OPTIMIZATION TECHNIQUES DRIVEN APPROACHES TO AUTOMATIC KEYPHRASE EXTRACTION

A Thesis Submitted
In Partial Fulfillment of the Requirements for the
Degree of

DOCTOR OF PHILOSOPHY

in

Mathematics

by

Reetika Singh
(Roll No. 2K22/PHDAM/501)

Under the Supervision of

Dr. Goonjan Jain



Department of Applied Mathematics
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

June 2026

© Delhi Technological University-2026

All rights reserved.

CANDIDATE'S DECLARATION

I, **Reetika Singh**, hereby certify that the work which is being presented in the thesis entitled **"Optimization Techniques Driven Approaches to Automatic Keyphrase Extraction"** in partial fulfillment of the requirements for the award of the Degree of **Doctor of Philosophy** in Mathematics, submitted in the Department of Applied Mathematics, Delhi Technological University is an authentic record of my own work carried out during the period from 19th January 2023 to 15th July 2026 under the supervision of Dr. Goonjan Jain.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Reetika Singh

Roll No: 2K22/PHDAM/501

CERTIFICATE

Certified that **Ms. Reetika Singh** (2K22/PHDAM/501) has carried out their research work presented in this thesis entitled **"Optimization Techniques Driven Approaches to Automatic Keyphrase Extraction"** for the award of **Doctor of Philosophy** from Department of Applied Mathematics, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student herself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Dr. Goonjan Jain
Supervisor
Department of Applied Mathematics
Delhi Technological University
Delhi.

Prof. R. Sriavastava
Head Of Department
Department of Applied Mathematics
Delhi Technological University
Delhi.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor, Dr. Goonjan Jain, for her constant motivation, invaluable guidance, and unwavering support throughout my research journey. It has been a privilege and a great pleasure to work under her supervision. Her encouragement, insightful suggestions, and continuous assistance have played a crucial role in the successful completion of this work.

I am deeply thankful to the DRC Chairperson and Head of the Department, Prof. R. Srivastava, and all the DRC members for their support and for providing the facilities and academic environment necessary for carrying out my research. I also extend my sincere thanks to the Dean (PG) and the office staff for their continued cooperation and assistance.

I would also like to express my gratitude to my fellow research scholars for their valuable discussions, encouragement, and support throughout this journey. Their willingness to share ideas and review my work has been immensely helpful.

I am sincerely thankful to Delhi Technological University for providing the fellowship support that made my Ph.D. research possible.

Finally, I am deeply indebted to my family for their unconditional love, endless patience, constant encouragement, and unwavering belief in me. Their support has been the foundation of my strength and success, and without them, this journey would not have been possible.

Above all, I bow with deep reverence and heartfelt gratitude to Radha Krishna, whose divine blessings, guidance, and grace have given me the strength, wisdom, and perseverance to overcome every challenge and successfully complete this journey.

Date: October 11, 2026 **REETIKA SINGH**

Place: Delhi, India.

This Thesis is dedicated to
The Almighty,
My Teachers,
&
My Family Members.

Contents

CANDIDATE'S DECLARATION i

CERTIFICATE BY THE SUPERVISOR iii

Acknowledgement v

Abstract xvii

List of Tables xix

List of Figures xxi

List of Abbreviations xxv

1 Introduction 1

1.1 Background and Motivation 2

1.2 Evolution of Textual Data Sources 4

1.3 Fundamentals of Keyword and Keyphrase Extraction 6

1.3.1 General Pipeline for Keyword Extraction 8

1.4 Role of Keyword Extraction in Natural Language Processing 9

1.5 Challenges in Keyword Extraction Across Heterogeneous Data 11

1.5.1 Challenges in Textual Documents 12

1.5.2 Challenges in Social Media Text 13

1.5.3 Challenges in Multimodal Data 14

1.5.4 Challenges in Application-Oriented Tasks 15

1.6 Motivation for Optimization-Based Keyword Extraction 16

1.7 Problem Statement 17

1.8 Organization of the Thesis 17

2 Literature Review 19

2.1 Keyword Extraction Methods for Structured Textual Documents 20

2.1.1 Statistical Approaches 20

2.1.2 Graph-Based Approaches . . . 21
2.1.3 Embedding-Based Approaches . . . 22
2.1.4 Machine Learning and Deep Learning Approaches . . . 24
2.1.5 Other Approaches . . . 24
2.1.6 Limitations of Existing Approaches . . . 25
2.2 Keyword Extraction Methods for Social Media Text . . . 26
2.2.1 Limitations of Existing Approaches . . . 28
2.3 Multimodal Keyword Extraction . . . 29
2.3.1 Limitations of Existing Approaches . . . 31
2.4 Downstream Applications of Keyword Extraction . . . 32
2.4.1 Limitations of Existing Approaches . . . 33
2.5 Research Gaps . . . 34
2.6 Research Objectives . . . 35

3 Keyword Extraction from Textual Documents 37
3.1 Text Preprocessing . . . 37
3.1.1 Standard Preprocessing and Linguistic Processing . . . 38
3.1.2 Proposed Extended Preprocessing Strategy . . . 40
3.2 FUBAKE Methodology . . . 42
3.2.1 Candidate Keyphrase Preparation and Co-occurrence Graph Construction . . . 44
3.2.2 Initialization of Pheromone Intensity, Desirability, and Bayesian Belief 45
3.2.2.1 Initialization of Pheromone Intensity . . . 46
3.2.2.2 Initialization of Desirability . . . 48
3.2.2.3 Initial Bayesian Belief Calculation . . . 48
3.2.3 Fuzzy Bayesian Ant Colony Optimization for Keyphrase Selection . . . 49
3.2.3.1 Fuzzy Modeling of Pheromone Intensity and Desirability . . . 49
3.2.3.2 Bayesian Belief Update . . . 52
3.2.3.3 Pheromone Intensity Update . . . 54
3.2.3.4 Bayesian Convergence and Termination Criteria . . . 55
3.2.4 Algorithmic Summary of FUBAKE . . . 56
3.3 TopS-Key Methodology . . . 57
3.3.1 Feature Scoring of Candidate Keyphrases . . . 59
3.3.2 Dimensionality Reduction using Principal Component Analysis . . . 62
3.3.3 Applying Fuzzy TOPSIS with Shannon Entropy Weights . . . 64
3.3.3.1 Shannon Entropy-Based Criterion Weighting . . . 65
3.3.3.2 Applying Fuzzy TOPSIS for Candidate Ranking . . . 66
3.3.4 Algorithmic Summary of TopS-Key . . . 69

3.4 ForestRank Methodology . . . 70
3.4.1 Document Preparation and Candidate Representation . . . 70
3.4.2 Feature Significance Analysis using Random Forest Classification . . . 72
3.4.3 MCDM Technique - WASPAS . . . 74
3.4.4 Algorithmic Summary of ForestRank . . . 75
3.5 Experimental Evaluation and Comparative Analysis . . . 76
3.5.1 Datasets . . . 76
3.5.2 Evaluation Metrics . . . 77
3.5.3 Similarity Threshold Analysis . . . 78
3.5.3.1 FuBAKE Similarity Threshold Analysis . . . 79
3.5.3.2 TopS-Key Similarity Threshold Analysis . . . 80
3.5.4 Parameter Sensitivity Analysis . . . 81
3.5.4.1 FuBAKE Weighting Parameter Analysis (α, β, and γ) . . . 81
3.5.4.2 FuBAKE Pheromone Evaporation Rate Analysis (ρ) . . . 82
3.5.4.3 FuBAKE Convergence Analysis . . . 83
3.5.4.4 TopS-Key PCA Variance Threshold Sensitivity Analysis . . . 84
3.5.5 Ablation Study Analysis . . . 86
3.5.5.1 FuBAKE Ablation Study Analysis . . . 86
3.5.5.2 TopS-Key Ablation Study Analysis . . . 87
3.5.6 Comparative Performance Analysis . . . 89
3.5.6.1 Performance Comparison at Top@3 . . . 90
3.5.6.2 Performance Comparison at Top@5 . . . 92
3.5.6.3 Performance Comparison at Top@10 . . . 92
3.5.7 Statistical Significance Analysis . . . 94
3.5.7.1 Statistical Validation of FuBAKE . . . 95
3.5.7.2 Statistical Validation of TopS-Key . . . 96
3.5.8 Evaluation of ForestRank . . . 98
3.5.8.1 Feature Importance Analysis . . . 98
3.5.8.2 Performance Impact of Feature Refinement . . . 99
3.5.8.3 Comparative Performance Analysis . . . 99
3.6 Conclusion . . . 100

4 Keyword Extraction from Social Media Text **103**
4.1 Social Media Text Preprocessing . . . 104
4.1.1 Standard Preprocessing and Linguistic Normalization for Social Media Text . . . 104
4.1.2 Extended Preprocessing Strategy for Social Media Text . . . 106

4.2 COM-Key: Commonsense-Augmented Fuzzy Graphs with Multi-Centrality for Keyword Discovery and Knowledge Mining in Social Media Data 107
4.2.1 Semantic Embedding and Clustering of Candidate Tokens 108
4.2.2 Commonsense-Enriched Fuzzy Graph Construction 110
4.2.2.1 Semantic Expansion via ConceptNet Neighborhoods 112
4.2.2.2 Edge Weighting via Structural Graph Connectivity 112
4.2.2.3 Fuzzy Membership Estimation via Contextual Embedding Similarity 113
4.2.3 Valuation of Node Importance via Fuzzy Centrality Modeling 113
4.2.3.1 Fuzzy Closeness Centrality 114
4.2.3.2 Fuzzy Eigenvector Centrality 114
4.2.3.3 Fuzzy Katz Centrality 115
4.2.3.4 Fuzzy Harmonic Centrality 115
4.2.3.5 Fuzzy PageRank Centrality 116
4.2.4 Multi-Criteria Keyword Ranking via WASPAS Aggregation of Fuzzy Centralities 116
4.2.5 Algorithmic Summary of COM-Key 118
4.3 Experimental Evaluation and Analysis of COM-Key 119
4.3.1 Datasets 119
4.3.2 Evaluation Protocol 120
4.3.2.1 Gold Standard Annotations 120
4.3.2.2 Inter-Annotator Agreement 120
4.3.3 Evaluation Metrics 121
4.3.4 Results and Discussion 122
4.3.4.1 Performance Analysis 123
4.3.4.2 Discussion 123
4.3.5 Comparative Analysis 124
4.3.6 Statistical Performance Analysis 125
4.3.6.1 Paired t-Test 125
4.3.6.2 ANOVA Analysis 126
4.3.6.3 Tukey HSD Test 126
4.4 QuToX: A Quantum-Inspired Game-Theoretic Computational Framework for Keyword Extraction and Interpretable Toxicity Detection in Short-Text Data . 127
4.4.1 Semantic Representation using Contextual Word Embeddings 128
4.4.2 Similarity and Interaction Modeling 129
4.4.3 Quantum Game-Theoretic Keyword Selection 131
4.4.4 Algorithmic Summary of QuToX 134
4.5 Experimental Evaluation 134

4.5.1 Dataset . . . 134
4.5.2 Keyword Extraction Analysis . . . 135
4.6 HumorKey: A Novel Keyword Extraction Framework Using Bit-Level Encoding and Fisher–Yates Shuffling for Humor Detection . . . 136
4.6.1 Bit-Level Structural Encoding . . . 137
4.6.2 Permutation via Fisher-Yates Shuffling . . . 137
4.6.3 Contextual Diffusion . . . 138
4.6.4 Semantic Adjustment and Keyword Selection . . . 138
4.6.5 Algorithmic Summary of HumorKey . . . 139
4.7 Conclusion . . . 139

5 Multimodal Keyword Extraction 141
5.1 Multimodal Content . . . 142
5.1.1 Types of Multimodal Content and Image-Text Data in Social Media . . . 143
5.1.2 Memes as Multimodal Communication Units . . . 143
5.1.3 Offensive and Contextual Memes . . . 145
5.1.4 Need for Multimodal Reasoning in Offensive Meme Interpretation . . . 146
5.1.5 Multimodal Keyword Extraction . . . 147
5.2 Knowledge-Enriched Semantic Keyword Extraction with NSGA-II for Offensive Meme Detection via Thermodynamic Modeling . . . 148
5.2.1 Multi-modal Semantic Acquisition . . . 149
5.2.1.1 Scene-Level Semantic Captioning using BLIP . . . 150
5.2.1.2 Embedded Text Extraction using OCR . . . 151
5.2.1.3 Visual Demographic and Emotive Analysis . . . 152
5.2.1.4 Unified Multimodal Semantic Representation . . . 154
5.2.2 Offensive Lexicon Expansion via ConceptNet . . . 155
5.2.3 Offensive-Aware Keyword Extraction via NSGA-II . . . 157
5.2.3.1 Candidate Keyword Generation and Semantic Representation 157
5.2.3.2 Multi-objective Keyword Optimization . . . 160
5.2.3.3 NSGA-II-Based Evolutionary Optimization and Final Keyword Selection . . . 162
5.2.4 Algorithmic Summary . . . 164
5.3 Conclusion . . . 164

6 Downstream Applications of Keyword Extraction 167
6.1 Toxicity Detection in Tweets . . . 168
6.1.1 Lexicon Construction Using N-gram Expansion . . . 170
6.1.2 Statistical Labeling of N-grams . . . 171
6.1.3 Threshold-Based Toxicity Clustering . . . 172

6.1.4 Rule-Based Toxicity Prediction . . . 173
6.1.5 Toxicity Detection Evaluation . . . 174
6.1.6 Toxicity Threshold (δ) Sensitivity Analysis . . . 176
6.1.7 Experimental Results . . . 178
6.1.7.1 Evaluation Metrics . . . 180
6.1.7.2 Results . . . 182
6.1.8 Comparative Analysis . . . 184
6.1.9 Ten-Fold Cross-Validation . . . 186
6.1.10 Statistical Performance Analysis . . . 187
6.1.10.1 Two-Tailed Z-Test Statistical Analysis . . . 187
6.1.10.2 Bootstrapping-Based Statistical Evaluation . . . 189
6.2 Humor Detection from Text . . . 190
6.2.1 Feature Representation . . . 191
6.2.1.1 Lexical Complexity Representation . . . 192
6.2.1.2 Sentiment-Based Representation . . . 192
6.2.1.3 Semantic Incongruity Representation . . . 193
6.2.1.4 Sarcasm-Oriented Pragmatic Representation . . . 194
6.2.2 Humor Classification . . . 194
6.2.3 Evaluation . . . 197
6.2.3.1 Datasets . . . 197
6.2.3.2 Comparative Results and Discussion . . . 197
6.2.3.3 Statistical Validation . . . 200
6.3 Offensive Meme Detection . . . 202
6.3.1 Extraction of Explicit and Implicit Offense-Indicative Features . . . 203
6.3.1.1 Explicit Offense-Oriented Features . . . 204
6.3.1.2 Implicit Offense-Oriented Features . . . 205
6.3.2 Thermodynamic Offense Modeling and Meme Classification . . . 206
6.3.3 Experimental Evaluation . . . 209
6.3.3.1 Datasets . . . 210
6.3.3.2 Evaluation Metrics . . . 211
6.3.3.3 Performance of the Proposed Method . . . 211
6.3.3.4 KDE-Based Threshold Analysis . . . 212
6.3.3.5 Comparative Analysis . . . 213
6.3.3.6 Statistical Significance Analysis . . . 215
6.4 Conclusion . . . 217

7 Conclusion, Future Scope, and Social Impact **219**
7.1 Summary of Findings . . . 220

7.2 Key Contributions 222
7.3 Limitations of the Study 223
7.4 Future Scope 225
7.5 Social Impact and Relevance to Sustainable Development Goals 227
7.6 Conclusion 231

Bibliography **233**

List of Publications **253**

Plagiarism Report **255**

Abstract

The rapid growth of digital communication platforms has resulted in an unprecedented increase in the volume and diversity of textual and multimodal data generated across online environments. Information is now disseminated through structured documents, social media posts, reviews, conversational text, and multimodal image-text content, each characterized by varying levels of contextual dependency, semantic ambiguity, and linguistic noise. These challenges significantly limit the effectiveness of conventional keyphrase extraction approaches, which primarily rely on frequency statistics, heuristic ranking strategies, or fixed graph-scoring mechanisms that often fail to preserve contextual semantics and domain adaptability.

This thesis presents optimization driven approaches to automatic keyphrase extraction by formulating keyword extraction as a semantic optimization problem rather than a purely statistical ranking task. The proposed research integrates graph-based representations, contextual semantic modeling, probabilistic reasoning, fuzzy decision mechanisms, quantum inspired interaction modeling, and multimodal semantic analysis to improve semantic relevance, diversity, redundancy reduction, and contextual coverage within extracted keyphrase subsets.

The research first investigates keyphrase extraction from structured textual documents through the development of frameworks such as FuBAKE, TopS-Key, and ForestRank, which combine graph-based ranking with semantic and structural optimization strategies. The work is subsequently extended to noisy social media environments through frameworks including COM-Key, QuToX, and HumorKey, designed to capture contextual semantics, emotional polarity, semantic incongruity, sarcasm, and toxicity oriented language behavior in short and highly dynamic text. To further address multimodal semantic understanding, the thesis proposes a knowledge-enriched semantic keyword extraction framework with NSGA-II optimization for offensive meme detection. The framework integrates visual semantics, OCR-based textual extraction, semantic embeddings, offensive lexicon expansion, and thermodynamic offense

modeling for interpretable multimodal offensiveness prediction.

The extracted semantic keyword representations are further utilized in downstream applications including toxic tweet detection, humor detection, and offensive meme classification, demonstrating that optimization driven keyword extraction can function as a transferable semantic representation layer for complex language understanding tasks. Extensive experimental evaluation across multiple benchmark datasets, supported by comparative analysis, cross-validation, sensitivity analysis, ablation studies, and statistical significance testing, demonstrates that the proposed methodologies consistently outperform existing baseline approaches in terms of semantic relevance, contextual coverage, interpretability, and downstream predictive performance.

Overall, this thesis establishes optimization-driven automatic keyphrase extraction as a unified and interpretable semantic representation framework for downstream natural language understanding and multimodal reasoning, while providing a scalable foundation for future research in social media analysis and responsible AI-driven language technologies.

List of Tables

3.1 Summary of benchmark datasets used for experimental evaluation 77

3.2 Formulas for precision, recall, and F1-score 78

3.3 Convergence behavior of FuBAKE across benchmark datasets in terms of Fraction of Paragraphs (FP), Average Iteration (AI), and Cumulative Fraction (CF). 84

3.4 Comparative performance of FuBAKE and TopS-Key against benchmark methods at Top@3 90

3.5 Comparative performance of FuBAKE and TopS-Key against benchmark methods at Top@5 92

3.6 Comparative performance of FuBAKE and TopS-Key against benchmark methods at Top@10 93

3.7 Feature importance scores in ForestRank 98

3.8 Performance comparison using all features and optimized features 99

4.1 Keywords identified by three independent annotators across the three datasets 121

4.2 Evaluation metrics used for assessing keyword extraction performance 121

4.3 Evaluation metrics at top-$n = 3, 5, 10$ across datasets and annotators 122

4.4 Comparison of keyword extraction methods on IPL 2019 dataset across three readers 124

4.5 Comparison of keyword extraction methods on Elections 2019 dataset across three readers 125

4.6 Statistical summary of tweet length and extracted keywords 135

6.1 Evaluation Metrics for Toxicity Classification Model 183

6.2 Evaluation Metrics for Toxicity Classification Model 185
6.3 Per-fold evaluation metrics obtained using ten-fold cross-validation 187
6.4 Comparative Performance Evaluation of Humor Detection Models 198
6.5 Performance Evaluation of the Proposed Offensive Meme Detection Framework Across Benchmark Datasets 212
6.6 Comparative Performance Analysis Across Benchmark Offensive Meme Datasets 214

7.1 Research-specific mapping of the thesis to selected United Nations Sustainable Development Goals. 229

List of Figures

1.1 Worldwide growth of structured and unstructured digital data from 2006 to 2020. 2

1.2 Growth in scientific and engineering research publications in India and globally from 2012 to 2022. 3

1.3 Evolution of textual data sources from structured documents to social media, multimodal, and application-oriented content. 5

1.4 General pipeline for automatic keyword extraction. 9

1.5 Major roles of keyword extraction in Natural Language Processing. 10

2.1 Modalities in keyword extraction illustrating the transition from unimodal to multimodal settings and the integration of textual, audio, and visual information. 30

3.1 Overall workflow of the proposed FUBAKE methodology 43

3.2 Workflow of the proposed TopS-Key framework integrating preprocessing, feature scoring, PCA-based dimensionality reduction, entropy weighting, and Fuzzy TOPSIS ranking. 58

3.3 Workflow of the proposed ForestRank framework integrating preprocessing, feature scoring, Random Forest-based feature selection, and WASPAS-based keyphrase ranking. 71

3.4 Precision, Recall, and F1-score variation across different similarity thresholds for Inspec, SemEval, and DUC datasets. 79

3.5 Precision trends across varying similarity thresholds for all datasets at Top@3 80

3.6 Recall trends across varying similarity thresholds for all datasets at Top@3 . 80

3.7 F1-score trends across varying similarity thresholds for all datasets at Top@3 81

3.8 F1-score variation across different values of α, β, and γ for Inspec, SemEval, and DUC datasets. 82

3.9 Radar charts illustrating F1-score variation across pheromone evaporation rates for Inspec, SemEval, and DUC datasets. 83

3.10 F1-score trends across varying PCA cumulative variance thresholds for all benchmark datasets 85

3.11 FuBAKE ablation analysis across Inspec, SemEval, and DUC-2001 datasets. The red dashed line denotes the complete FuBAKE baseline. 87

3.12 Feature-level ablation results showing the impact of removing individual TopSKey features across benchmark datasets 88

3.13 Percentage decrease in average F1-score resulting from exclusion of individual TopSKey components 89

3.14 Two-tailed Z-test distributions illustrating FuBAKE's aggregate statistical superiority over benchmark methods across Inspec, SemEval2017, and DUC2001. 95

3.15 Wilcoxon signed-rank significance analysis illustrating FuBAKE's paired document-level superiority across benchmark datasets. 96

3.16 Two-tailed Z-test distributions illustrating TopS-Key's aggregate statistical superiority over benchmark methods across benchmark datasets. 97

3.17 ANOVA and Tukey HSD significance analysis illustrating TopS-Key's statistically distinct performance tier across benchmark datasets. 97

3.18 Comparative performance of ForestRank against benchmark methods across Top@5, Top@10, and Top@15. 100

4.1 Sample raw tweet from the IPL 2019 dataset before preprocessing 105

4.2 Overall methodological framework of COM-Key for social media keyword extraction 108

4.3 Illustrative ConceptNet semantic neighborhood for the keyword *Cricket*, demonstrating semantically typed commonsense relations across associated concepts 111

4.4 Pairwise Sørensen-Dice coefficients between annotators across datasets . . . 122

4.5 Trend analysis of Precision, Recall, and F1-score across datasets with increasing number of extracted keywords 123

4.6 Paired t-test results on F1-scores for IPL and Elections datasets 126
4.7 Tukey HSD-based boxplot analysis of F1-scores for IPL and Elections datasets 127
4.8 Workflow of the QuToX keyword extraction framework 128
4.9 Flow of the quantum keyword extraction process from payoff-based state initialization to superposition, entanglement, parameterized evolution, and measurement-based keyword selection 131
4.10 Distribution of toxic and non-toxic tweets in the dataset 135
4.11 Workflow of the HumorKey keyword extraction framework based on bit-level encoding and structured permutation 137

5.1 Examples of humorous memes illustrating how meaning emerges from the interaction between visual context and textual content 144
5.2 Examples of offensive memes where meaning emerges from cross-modal interaction between image and text 145
5.3 Illustration of how meme interpretation evolves from text-only, image-only, and visual attribute-level analysis to integrated multimodal reasoning. 146
5.4 Overall workflow of the proposed knowledge-enriched multimodal keyword extraction framework for meme-based content 149

6.1 Downstream toxicity detection framework of QuToX using quantum-inspired semantically enriched keyword representations extracted from tweets. 169
6.2 Distribution of n-gram clusters based on computed toxicity scores and occurrence frequency within the tweet corpus. 175
6.3 Word cloud visualization of dominant keywords present within (a) toxic lexical clusters and (b) non-toxic lexical clusters. 176
6.4 Variation of evaluation metrics with respect to the toxicity threshold parameter δ, illustrating the optimal performance balance near $\delta = 0.55$. 177
6.5 Confusion matrix illustrating the toxicity classification performance of the proposed QuToX framework. 179
6.6 Receiver Operating Characteristic (ROC) curve of the proposed toxicity detection framework. 183

6.7 Precision–Recall (PR) curve of the proposed toxicity detection framework. . . 183
6.8 Comparative evaluation of the proposed QuToX framework against traditional machine learning classifiers. 186
6.9 Standard normal distribution illustrating the Two-Tailed Z-Test and rejection regions. 188
6.10 CDF of bootstrapped F1-Scores with estimated 95% confidence interval. . . . 189
6.11 Downstream humor detection framework using semantically enriched keyword representations extracted through the proposed HumorKey methodology. 191
6.12 F1-Score comparison between transformer-based baseline models and the proposed HumorKey framework across both datasets. 199
6.13 Statistical validation results showing mean F1-Scores and standard deviation across repeated experimental runs for both datasets. 201
6.14 Downstream offensive meme detection framework utilizing multimodal offensive-aware keywords for explicit and implicit feature extraction, thermodynamic offense modeling, and final meme classification. 203
6.15 ROC curves of the proposed framework across benchmark offensive meme datasets. 212
6.16 KDE distributions of thermodynamic offense scores for offensive and non-offensive meme instances across benchmark datasets. 213
6.17 Two-tailed Z-test results across benchmark datasets at a 99% confidence level. 215
6.18 Bootstrapped CDF distributions of F1-Scores with 95% confidence intervals across benchmark datasets. 216

7.1 Research contributions to selected United Nations Sustainable Development Goals. 228

List of Abbreviations

Abbreviation	Description
ACO	Ant Colony Optimization
AI	Artificial Intelligence
ANOVA	Analysis of Variance
ASEW	Adaptive Shannon Entropy Weights
AUC	Area Under Curve
BERT	Bidirectional Encoder Representations from Transformers
BLIP	Bootstrapping Language-Image Pre-training
CDF	Cumulative Distribution Function
CF	Cumulative Fraction
CNN	Convolutional Neural Network
CNOT	Controlled-NOT
CRAFT	Character Region Awareness for Text Detection
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
FN	False Negative
FNIS	Fuzzy Negative Ideal Solution
FP	False Positive
FPIS	Fuzzy Positive Ideal Solution
GOT	Game of Thrones
HITS	Hyperlink-Induced Topic Search
HSD	Honestly Significant Difference
HUX	Half Uniform Crossover
KEA	Keyphrase Extraction Algorithm

List of Abbreviations (continued)

Abbreviation	Description
KECNW	Keyword Extraction based on Complex Networks and Wikipedia
KDE	Kernel Density Estimation
KNN	k-Nearest Neighbors
LDA	Latent Dirichlet Allocation
MCC	Matthews Correlation Coefficient
MCDM	Multi-Criteria Decision Making
MMBT	Multimodal Bitransformers
MMDL	Multimodal Deep Learning
mBERT	Multilingual Bidirectional Encoder Representations from Transformers
NER	Named Entity Recognition
NLP	Natural Language Processing
NPD	Noun Phrase Density
NSGA-II	Non-dominated Sorting Genetic Algorithm II
OCR	Optical Character Recognition
PCA	Principal Component Analysis
POS	Part-of-Speech
PR	Precision–Recall
QGT	Quantum Game Theory
RBF	Radial Basis Function
RGB	Red Green Blue
ROC	Receiver Operating Characteristic
SDGs	Sustainable Development Goals
SIR	Standard Instance Recurrence
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
TKG	Twitter Keyword Graph
TP	True Positive
TSR	Token Set Ratio
URL	Uniform Resource Locator

List of Abbreviations (continued)

Abbreviation	Description
VADER	Valence Aware Dictionary and Sentiment Reasoner
ViT	Vision Transformer
WASPAS	Weighted Aggregated Sum Product Assessment
WPM	Weighted Product Model
WSM	Weighted Sum Model
YOLOv3	You Only Look Once Version 3

Chapter 1

Introduction

The widespread adoption of digital technologies and online communication platforms has led to a substantial increase in the amount of textual information generated across different domains. Large volumes of text are continuously produced through research publications, news articles, blogs, web pages, product reviews, social networking platforms, and multimedia content containing textual descriptions. As a result, textual data has become increasingly diverse in structure, length, style, and context. Although this growth has improved access to information, it has also created significant challenges in organizing, analyzing, and retrieving relevant content from large collections of data.

As illustrated in Figure 1.1, the volume of digital data has increased rapidly over the years, with unstructured data accounting for the largest proportion of worldwide data growth. The increasing dominance of unstructured content highlights the need for intelligent methods that can efficiently process, summarize, and retrieve relevant information from large-scale repositories. This rapid growth in data generation has made manual analysis impractical and has created a strong demand for automated techniques capable of extracting the most informative and representative content.

In this context, keyword and keyphrase extraction has gained considerable importance as a fundamental task in Natural Language Processing. The objective of keyword extraction is to identify a set of representative terms or phrases that effectively summarize the main content of

[0]Source of Figure 1.1: https://twitter.com/timothy_hughes/status/619075227021090817

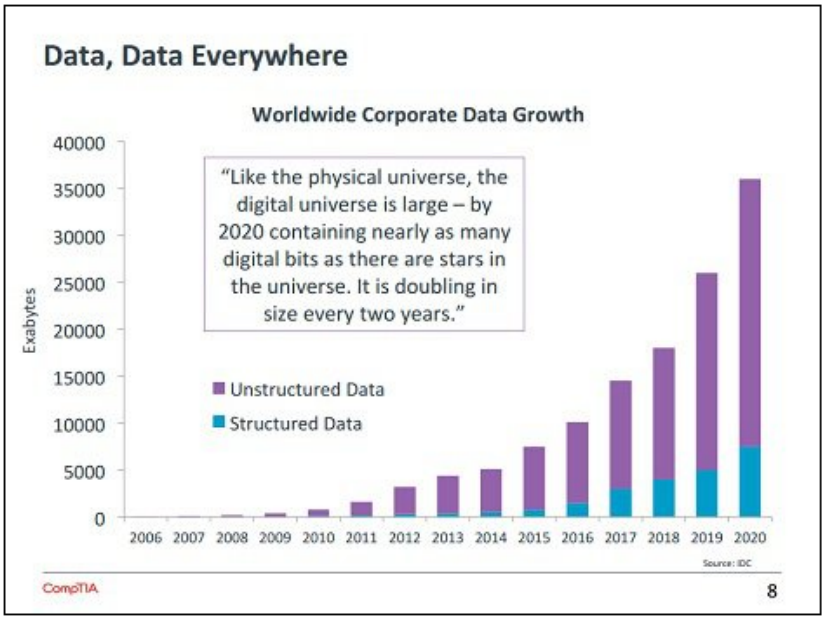

Figure 1.1: Worldwide growth of structured and unstructured digital data from 2006 to 2020.

a document. These keywords provide a compact semantic representation of textual information and support a variety of applications, including information retrieval, text summarization, document categorization, topic discovery, sentiment analysis, recommendation systems, and content filtering. However, the increasing heterogeneity of textual data, ranging from formal documents to short social media posts and multimodal content, requires keyword extraction methods that are capable of capturing semantic relevance, contextual meaning, and domain-specific characteristics in a reliable and scalable manner.

1.1 Background and Motivation

The increasing digitization of information has fundamentally changed the way knowledge is created, stored, and disseminated across different domains. Advances in internet technologies, online communication platforms, and digital repositories have resulted in an unprecedented growth in the volume of textual data generated worldwide [1, 2]. Such data is continuously produced through research publications, books, technical reports, news articles, websites, blogs, online reviews, discussion forums, and social media platforms. The availability of large-scale textual resources has improved access to information and facilitated knowledge sharing; however, it has also introduced significant challenges in managing, organizing, and retrieving relevant content from vast document collections.

The rapid growth of research publications and digital archives has further intensified these challenges. Academic databases and scientific repositories now contain millions of documents spanning diverse disciplines, making manual exploration increasingly difficult. Researchers, practitioners, and decision makers are often required to process large amounts of textual information within limited time, creating a need for automated methods that can identify the most relevant and informative content efficiently.

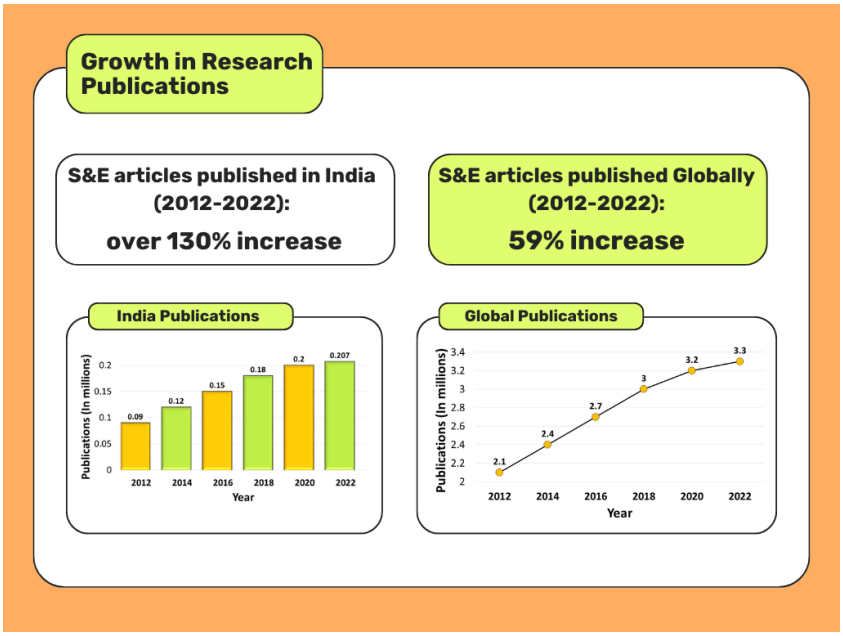

Figure 1.2: Growth in scientific and engineering research publications in India and globally from 2012 to 2022.

As illustrated in Figure 1.2, the number of scientific and engineering publications has increased considerably over the last decade at both national and global levels. India witnessed a growth of more than 130% in scientific publications between 2012 and 2022, while global scientific output increased by nearly 59% during the same period. This sustained increase in publication volume reflects the broader expansion of digital text repositories and highlights the growing difficulty of extracting useful information from large-scale document collections [3].

In this context, keyword and keyphrase extraction has emerged as an important research area in Natural Language Processing. The objective of keyword extraction is to identify a set of representative terms or phrases that capture the central concepts of a document. These keywords provide a concise semantic representation of textual content and support several downstream tasks, including information retrieval, text summarization, document indexing, topic discov-

[0]Source of Figure 1.2: Based on publication statistics reported for scientific and engineering articles from 2012 to 2022.

ery, recommendation systems, sentiment analysis, and document classification [4, 5, 6].

However, the effectiveness of keyword extraction is strongly influenced by the nature of the input data. Traditional approaches are generally designed for long and structured documents, where term frequency and positional information are often sufficient to identify important words. Such methods are less effective when applied to short, noisy, and context-dependent text such as tweets, comments, memes, and online discussions. Moreover, multimodal content introduces additional challenges because semantic meaning may arise from interactions between textual and visual information. These limitations indicate the need for more robust and adaptive keyword extraction frameworks capable of handling heterogeneous data sources while preserving contextual relevance and semantic meaning [7, 8, 9, 10].

The motivation for this thesis is based on the following observations:

- The rapid growth of textual and multimodal data requires automated techniques for efficient information analysis.
- Manual analysis of large-scale document collections is time-consuming and impractical.
- Traditional keyword extraction methods are often insufficient for short, noisy, and context-dependent text.
- Existing approaches do not adequately capture semantic relationships and contextual meaning.
- Multimodal content introduces additional challenges due to interactions between text and visual information.
- Optimization-driven frameworks can improve keyword ranking by considering multiple semantic and contextual factors simultaneously.
- Extracted keywords can serve as interpretable intermediate representations for downstream tasks such as toxicity detection, humor analysis, and offensive content detection.

1.2 Evolution of Textual Data Sources

The nature of textual data has evolved considerably over time due to advances in communication technologies, internet accessibility, and digital content generation. Early text analysis

methods were primarily developed for formal and structured documents such as books, research papers, technical reports, news articles, and domain-specific records. However, the rapid growth of social media platforms, multimedia systems, and application-specific digital environments has significantly diversified the nature of textual information. Textual data is no longer limited to well-structured documents but now includes short social media posts, noisy online discussions, memes, image-text content, and task-specific data sources. As a result, the requirements of keyword extraction methods have also evolved from simple frequency-based analysis to more context-aware, semantic, and multimodal approaches.

Figure 1.3 illustrates the major categories of textual data considered in this thesis, beginning with structured text documents and extending to social media text, multimodal content, and application-oriented data.

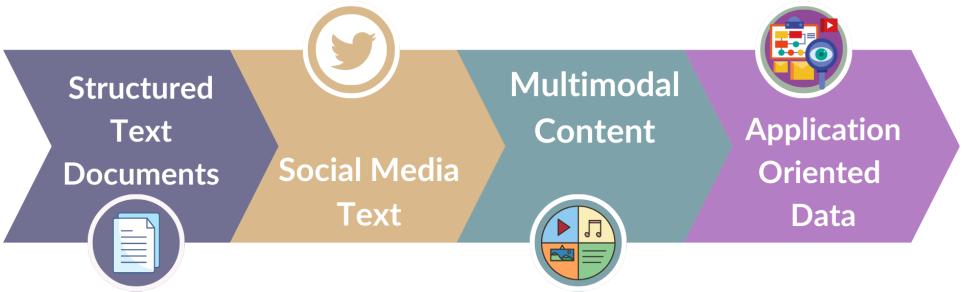

Figure 1.3: Evolution of textual data sources from structured documents to social media, multimodal, and application-oriented content.

Structured text documents represent the earliest and most widely studied form of textual data. These documents include books, research articles, technical reports, newspapers, legal records, policy documents, and other forms of formal written communication. Such data is generally characterized by well-defined grammatical structure, coherent context, longer document length, and relatively low levels of noise. Traditional keyword extraction methods were primarily designed for these forms of text, where important terms can often be identified through statistical features such as frequency, position, and co-occurrence patterns.

The emergence of social media platforms introduced a new category of textual data that differs significantly from traditional documents. Social media text includes tweets, comments, reviews, blog posts, online discussions, and chat messages. Unlike structured documents, this form of text is often short, informal, noisy, and highly dynamic. It frequently contains abbreviations, slang, emojis, hashtags, spelling variations, code-mixed expressions, and incomplete sentences. These characteristics make semantic understanding more difficult and require keyword extraction techniques that can capture context, user intent, and latent relationships

between words.

The increasing popularity of multimedia platforms has further expanded the scope of textual analysis to multimodal content. Multimodal data combines text with other modalities such as images, videos, audio, and metadata. Examples include memes, advertisements, infographics, multimedia news posts, video captions, and image-text social media content. In such data, semantic meaning often emerges from the interaction between textual and visual components rather than from text alone. Consequently, keyword extraction from multimodal content requires techniques capable of jointly modeling information from multiple modalities and achieving effective cross-modal semantic alignment.

In addition to these data types, there has been a growing emphasis on application-oriented data, where keyword extraction is performed to support a specific downstream task. Such tasks include toxicity detection, humor detection, offensive content identification, fake news analysis, sentiment prediction, and recommendation systems. In these scenarios, the relevance of a keyword is not determined solely by its statistical importance within the text but also by its usefulness for the target application. As a result, application-oriented keyword extraction requires domain-aware and task-specific approaches that can identify the most informative semantic cues and ensure task-aware keyword relevance for a particular problem setting.

1.3 Fundamentals of Keyword and Keyphrase Extraction

Keywords and keyphrases are widely used to represent the essential concepts present within a document. Since modern textual repositories contain large volumes of information, users often require a compact representation of document content rather than reading the entire text. Keywords and keyphrases address this need by identifying the most relevant terms that summarize the semantic meaning of a document.

Definition 1.1. A keyword is a single word that captures an important concept discussed in a document [11, 12].

Definition 1.2. A keyphrase is a sequence of one or more words that collectively represents an important concept, topic, or semantic aspect of a document [11, 13].

Although the terms keyword and keyphrase are frequently used interchangeably, they differ in the amount of information they convey. A keyword is generally a single term, such as “classification” or “optimization”, while a keyphrase may consist of multiple words, such as

“machine learning” or “social media analysis”. In many situations, keyphrases are more useful because they preserve contextual information and reduce ambiguity.

Definition 1.3. Keyword or keyphrase extraction is the process of identifying representative words or phrases that best describe the content of a document [14].

The main objective of keyword extraction is to produce a concise semantic representation of a document by selecting the most informative terms. These extracted terms can then be used in several applications such as information retrieval, document indexing, summarization, clustering, recommendation systems, and topic analysis. In large document repositories, keyword extraction helps users quickly understand the main ideas without examining the complete content.

Traditionally, keywords were manually assigned by authors, editors, or domain experts. This approach is still common in research articles, books, technical reports, and digital libraries. However, manual assignment is difficult to scale because it requires considerable effort and may vary across individuals. Two experts may assign different keywords to the same document depending on their interpretation, leading to inconsistency.

Definition 1.4. Automatic keyword extraction refers to the process of identifying important keywords or keyphrases from a document without manual intervention [13].

Automatic keyword extraction methods attempt to overcome the limitations of manual annotation by using computational approaches to determine which terms are most important. These methods usually rely on textual properties such as word frequency, term position, co-occurrence patterns, syntactic structure, semantic relationships, and contextual relevance. Based on these properties, the system ranks candidate terms and selects those that best represent the document.

Automatic keyword extraction methods can be grouped into supervised and unsupervised approaches.

Definition 1.5. Supervised keyword extraction refers to methods that use labeled training data to learn which candidate terms should be classified as keywords [12].

Definition 1.6. Unsupervised keyword extraction refers to methods that identify important terms without relying on annotated training data [14, 15].

Supervised methods often achieve good performance because they learn patterns from labeled examples. However, they require annotated datasets, which are not always available. Unsupervised methods are more flexible because they depend only on the information contained within

the document itself. As a result, unsupervised approaches are widely used in many real-world applications.

Another distinction in keyword extraction is between extractive and abstractive methods.

Definition 1.7. Extractive keyword extraction refers to methods that select keywords directly from the original document text [13, 14].

Definition 1.8. Abstractive keyword extraction refers to methods that generate keywords or phrases that may not explicitly appear in the original document but still reflect its meaning [5].

Most traditional methods are extractive because they rank and select existing terms from the document. Abstractive methods, on the other hand, attempt to generate more expressive and meaningful representations of the content. Although abstractive methods may produce more natural summaries, they are generally more complex and computationally expensive.

In general, keyword extraction involves several stages, including preprocessing, candidate term generation, feature computation, ranking, and final selection. During preprocessing, unnecessary elements such as stop words and punctuation are removed. Candidate terms are then generated using methods such as noun phrase extraction, n-gram generation, or part-of-speech tagging. Different statistical, syntactic, and semantic features are computed for each candidate term, and a ranking mechanism is applied to identify the most important keywords or keyphrases.

1.3.1 General Pipeline for Keyword Extraction

Automatic keyword extraction generally follows a sequence of stages that transform raw textual content into a final set of representative keywords or keyphrases. Although different methods may vary in terms of feature design, ranking mechanisms, and learning strategies, most keyword extraction systems follow a common workflow. This workflow typically includes preprocessing, candidate generation, feature extraction, ranking, and final keyword selection. Figure 1.4 illustrates the general pipeline of an automatic keyword extraction system.

The process begins with a raw document, which may be a research article, technical report, news article, online review, tweet, comment, or any other textual source. The quality and structure of the input text can significantly influence the effectiveness of the keyword extraction process. In the preprocessing stage, the raw text is cleaned and normalized in order to remove irrelevant information. Common preprocessing operations include tokenization,

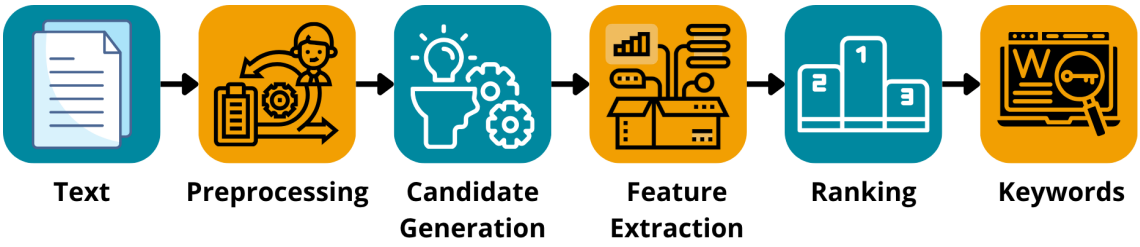

Figure 1.4: General pipeline for automatic keyword extraction.

stop word removal, punctuation removal, lowercasing, stemming, lemmatization, and part-of-speech tagging. Additional preprocessing steps may also be required for social media text, such as handling hashtags, emojis, URLs, abbreviations, and spelling variations.

After preprocessing, a set of candidate keywords or keyphrases is generated. Candidate terms are generally extracted using techniques such as noun phrase extraction, n-gram generation, syntactic rules, or part-of-speech patterns. The purpose of this stage is to identify all potentially relevant terms that may later be selected as final keywords. Once the candidate terms are identified, different features are computed for each term in order to estimate its importance. These features may include term frequency, inverse document frequency, positional importance, co-occurrence information, semantic similarity, graph centrality, contextual relevance, and embedding-based representations. The choice of features often depends on the nature of the data and the keyword extraction method being used.

In the ranking stage, each candidate term is assigned an importance score based on the extracted features. Different ranking methods may be employed, including statistical scoring functions, graph-based ranking, machine learning classifiers, fuzzy logic, optimization techniques, and multi-criteria decision-making frameworks. The highest ranked terms are considered to be the most representative of the document. In the final stage, the top-ranked candidate terms are selected as the final keywords or keyphrases.

1.4 Role of Keyword Extraction in Natural Language Processing

Keyword extraction plays an important role in Natural Language Processing because it provides a compact representation of textual content in terms of its most informative words or phrases. Instead of processing complete documents, many Natural Language Processing tasks rely on extracted keywords to identify the major themes, concepts, and semantic relationships

present in the text. By reducing a document to a smaller set of representative terms, keyword extraction improves computational efficiency, supports interpretability, and facilitates a variety of downstream analytical tasks.

Figure 1.5 illustrates some of the major roles of keyword extraction in Natural Language Processing, including text representation, knowledge discovery, document indexing, information retrieval, text clustering, text classification, text summarization, and topic modeling.

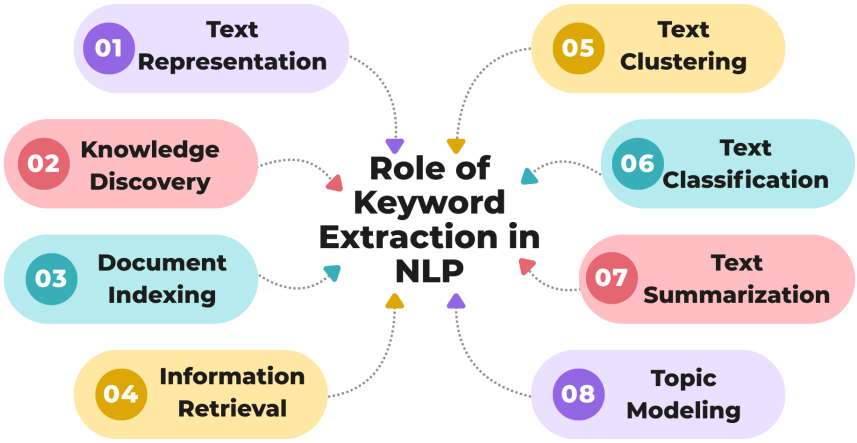

Figure 1.5: Major roles of keyword extraction in Natural Language Processing.

Text Representation: Keywords provide a concise representation of a document by capturing its most important concepts. Instead of processing the entire text, many Natural Language Processing systems use extracted keywords to describe the semantic content of a document in a compact form. Such representations are particularly useful for reducing dimensionality, improving computational efficiency, and preserving important contextual information in large textual repositories [16, 17, 7].

Knowledge Discovery: Keyword extraction supports knowledge discovery by revealing hidden relationships, emerging themes, and semantic patterns present in textual data. By analyzing frequently occurring keywords and their associations, it becomes possible to identify trends, discover latent topics, and uncover useful domain knowledge from large-scale document collections [18, 19, 7].

Document Indexing: Extracted keywords are widely used for document indexing in digital libraries, search engines, and document management systems. Documents indexed using representative keywords can be organized more efficiently and retrieved more accurately. Effective indexing improves the accessibility and navigability of large document collections [13, 20, 8].

Information Retrieval: One of the most important applications of keyword extraction is information retrieval. Search engines and retrieval systems use keywords to match user queries with relevant documents. Documents containing semantically relevant keywords are more likely to be retrieved in response to a search query, thereby improving retrieval effectiveness and relevance [21, 20, 22].

Text Clustering: Keyword extraction can be used to group similar documents into clusters based on shared themes or concepts. Documents that contain similar keywords are often semantically related and can therefore be grouped together. Such clustering is useful in applications such as document organization, news grouping, recommendation systems, and topic discovery [23, 24, 25].

Text Classification: Extracted keywords can serve as discriminative features for text classification tasks. In many classification problems, keywords help identify the presence of specific categories, sentiments, topics, or labels within a document. Applications include spam detection, sentiment analysis, toxicity detection, fake news classification, and document categorization [26, 27, 7].

Text Summarization: Keywords and keyphrases provide a concise overview of the content of a document and can therefore support automatic text summarization. Important keywords often indicate the central topics of a document and help summarization systems identify which sentences or concepts should be included in the final summary [15, 28, 29].

Topic Modeling: In topic modeling tasks, extracted keywords are often used to represent the dominant themes within a document collection. Topic modeling methods commonly rely on representative terms to identify latent semantic structures and group documents according to shared topics. Recent topic modeling frameworks such as BERTopic further use keyword-based topic representations to improve the interpretability and coherence of discovered topics [30, 31].

1.5 Challenges in Keyword Extraction Across Heterogeneous Data

Keyword extraction performance is highly dependent on the nature of the input data. Different forms of textual content exhibit different linguistic, structural, and semantic characteristics, which directly influence the effectiveness of extraction methods. Techniques that perform

well for long and structured documents are often less effective when applied to short social media posts, multimodal content, or task-specific textual data. This variation arises because the importance of a keyword is not determined solely by its frequency of occurrence but also by its contextual meaning, semantic relevance, and relationship with other concepts in the data.

The complexity of keyword extraction further increases when the input contains noisy language, informal expressions, abbreviations, code-mixing, visual information, or domain-specific terminology. In many situations, highly informative terms may appear only once within a document, while frequently occurring words may contribute little to the actual meaning of the text. Similarly, in multimodal data, semantic interpretation often depends on the interaction between textual and visual content rather than on either modality independently.

As a result, keyword extraction presents different challenges across structured documents, social media text, multimodal content, and application-oriented tasks. Understanding these challenges is essential for developing keyword extraction frameworks that are robust, adaptive, and capable of operating effectively across heterogeneous data environments.

1.5.1 Challenges in Textual Documents

Keyword extraction from structured textual documents presents several challenges despite the relatively organized nature of such data. Traditional methods are often developed for long documents such as research articles, technical reports, and news articles, where important terms are usually identified using features such as frequency, position, and co-occurrence. However, these features do not always reflect the true semantic importance of terms within a document.

A major limitation of frequency-based methods is that they often assign high importance to terms that occur repeatedly throughout the document. Although frequent terms may sometimes indicate important concepts, many highly repeated words can be generic or less informative. At the same time, semantically important terms may appear only once or twice and therefore receive a low importance score. As a result, traditional approaches may overemphasize common terms while missing rare but meaningful concepts.

Another challenge arises from redundancy among candidate terms. Different words or phrases may refer to similar concepts, resulting in extracted keyword sets that contain overlapping information. For example, terms such as “machine learning”, “learning algorithm”, and “artificial intelligence” may all represent related ideas within the same document. Without redun-

dancy control, the final keyword set may become repetitive rather than diverse and representative.

The importance of a term may also vary across different sections of a document. Keywords appearing in the title, abstract, conclusion, or section headings are often more informative than those appearing in the main body text. In addition, long documents may contain multiple themes and subtopics, making it difficult to identify a balanced set of keywords that adequately represents the entire document.

Structured documents also contain domain-specific terminology, particularly in scientific, legal, medical, and technical fields. Such terms may not be recognized effectively by general-purpose keyword extraction methods. Furthermore, most traditional approaches rely on a single scoring mechanism, such as frequency, graph centrality, or semantic similarity. However, no single feature is sufficient to capture all aspects of keyword relevance. Effective keyword extraction therefore requires the combined consideration of multiple factors, including statistical importance, contextual relevance, semantic meaning, and inter-term relationships.

1.5.2 Challenges in Social Media Text

Social media text introduces several challenges that differ significantly from those observed in structured documents. Unlike research articles, books, or news reports, social media posts are generally short, informal, and highly dynamic. Platforms such as Twitter, Reddit, Instagram, and YouTube comments often contain limited textual content, making it difficult to identify representative keywords using conventional methods.

One of the primary challenges is the short length of social media content. Since posts often contain only a few words or sentences, they provide limited contextual information and sparse textual representations. As a result, important terms may not appear frequently enough to be distinguished from less relevant words.

Another challenge arises from the informal nature of social media language. User-generated content frequently contains slang, abbreviations, emojis, hashtags, mentions, spelling variations, and grammatical inconsistencies. Many posts also include code-mixed or multilingual text, where multiple languages are used within the same sentence. Such linguistic variability introduces significant noise and makes it difficult for traditional keyword extraction methods to accurately identify meaningful terms and semantic relationships.

The semantic meaning of social media text is often implicit and highly context-dependent. A

short post may rely on sarcasm, humor, cultural references, previous conversations, or ongoing events to convey its meaning. Consequently, the relevance of a keyword cannot always be inferred from the text alone, particularly when the message depends on hidden context or implied intent.

Social media data is also characterized by high topical diversity and rapid topic shifts. Trending events, viral discussions, and breaking news can quickly alter the dominant themes present in online conversations. In addition, new slang terms, abbreviations, hashtags, and expressions emerge continuously, making it difficult for static keyword extraction methods to remain effective over time.

Furthermore, most traditional keyword extraction methods are designed for long and structured documents where important concepts are repeated across multiple sections. Such methods often rely on features such as term frequency, position, and document structure, which are less effective for short and noisy social media text. Therefore, keyword extraction from social media requires approaches that can handle sparse context, informal language, evolving vocabulary, and hidden semantic cues.

1.5.3 Challenges in Multimodal Data

Multimodal keyword extraction introduces additional challenges because the semantic meaning of the content is often distributed across multiple modalities rather than being contained within text alone. In many cases, textual information derives its meaning from associated visual content, making standalone text analysis insufficient. Examples such as memes, advertisements, social media posts, and multimedia news articles often require the combined interpretation of both text and images in order to understand the intended message.

One of the primary challenges in multimodal data is semantic alignment between textual and visual modalities. A word or phrase may have different meanings depending on the accompanying image, while the same image may convey different meanings when paired with different textual content. Establishing a meaningful relationship between these modalities is therefore a complex task.

Another challenge arises from the heterogeneous nature of multimodal features. Textual and visual information are represented using different feature spaces and exhibit different statistical properties. Text is sequential and symbolic in nature, whereas images contain spatial, visual, and object-level information. Combining these heterogeneous representations into a unified

framework is often difficult.

The semantic meaning of multimodal content is also frequently implicit. In memes and visual social media posts, meaning may emerge through sarcasm, symbolism, cultural references, or hidden associations between text and images. In such cases, neither the text nor the image is sufficient on its own to explain the actual meaning of the content.

Another difficulty is that the contribution of each modality may vary across different samples. In some cases, the image carries most of the semantic information, whereas in others, the text is more informative. Determining the relative importance of different modalities is therefore a significant challenge in multimodal keyword extraction.

Furthermore, most traditional keyword extraction methods rely on a single objective or a single type of feature. Such approaches are often unable to balance competing semantic contributions from text and images simultaneously. Joint modeling of multiple modalities also increases computational complexity because it requires the extraction, integration, and optimization of heterogeneous information from different sources.

1.5.4 Challenges in Application-Oriented Tasks

Application oriented keyword extraction presents challenges that differ from those encountered in general purpose keyword extraction. In many downstream tasks, the most important keywords are not necessarily the most frequent or semantically dominant terms in the text. Instead, the extracted keywords must be highly discriminative for the specific task being performed.

Different applications prioritize different semantic cues. For example, toxicity detection may rely on abusive words or hate expressions, whereas humor detection may depend on sarcasm, exaggeration, or irony. Similarly, offensive content detection may require keywords associated with stereotypes, coded language, or symbolic references. As a result, the same document may contain different sets of relevant keywords depending on the target application. Another challenge is that task labels strongly influence keyword importance. A term that is useful for one task may be irrelevant for another. Generic keyword extraction methods generally focus on identifying terms that are representative of the overall document content. However, such methods may fail to capture the specific semantic cues required for classification, prediction, or detection tasks.

Application oriented tasks also often involve subtle and implicit meanings. Toxicity, humor,

offensiveness, and sentiment may be expressed indirectly through sarcasm, wordplay, hidden intent, or contextual references. Such patterns are difficult to capture using traditional statistical methods alone.

Therefore, keyword extraction in application-oriented settings must be aligned with the objective of the downstream task. The extracted keywords should not only summarize the document but also improve interpretability, predictive performance, and decision making.

1.6 Motivation for Optimization-Based Keyword Extraction

Optimization based keyword extraction has gained importance because the task of selecting representative keywords can be naturally viewed as a decision-making problem. Given a large set of candidate terms, the objective is to identify a small subset of keywords that best represents the semantic content of the document. However, selecting such a subset is not straightforward because multiple criteria must often be satisfied simultaneously.

An effective keyword set should not only contain highly relevant terms but should also avoid redundancy, preserve diversity, capture contextual importance, and, in some cases, support a specific downstream application. These requirements are often conflicting in nature. For example, a highly frequent term may not be semantically important, while a semantically meaningful term may appear only once in the document. Similarly, two candidate keywords may both be highly relevant but may represent overlapping concepts. Therefore, keyword extraction cannot be reduced to a simple ranking problem based on a single score.

Traditional keyword extraction methods generally rely on one dominant feature, such as frequency, graph centrality, positional importance, or semantic similarity. Although such approaches are computationally simple, they are often unable to balance the trade-offs among multiple relevance criteria. As a result, the extracted keyword set may contain redundant terms, overlook semantically important concepts, or fail to adapt to different types of data.

Definition 1.9. Optimization refers to the process of identifying the best possible solution from a set of feasible alternatives according to one or more objective functions and constraints [32, 33].

Definition 1.10. Multi-objective optimization refers to the process of simultaneously optimizing two or more potentially conflicting objective functions in order to identify a balanced solution [34].

Optimization techniques provide a principled mathematical framework for solving complex selection problems under multiple objectives and constraints. Unlike heuristic ranking methods, optimization based approaches can explicitly model competing criteria and search for solutions that achieve a better balance among them.

Optimization methods are particularly useful in keyword extraction because they allow the simultaneous consideration of multiple features such as semantic relevance, contextual importance, positional information, redundancy reduction, and task-specific significance. Multi-objective optimization techniques further enable the modeling of conflicting criteria and support the selection of more representative keyword sets.

Another advantage of optimization based methods is their adaptability across different forms of data. Structured documents, social media posts, multimodal content, and application-oriented tasks all exhibit different characteristics and require different relevance criteria. Optimization based frameworks can incorporate these criteria within a unified model and adjust the keyword selection process according to the nature of the data. Previous studies have shown that optimization techniques such as ant colony optimization and particle swarm optimization can improve feature selection, term extraction, and keyword ranking in text mining applications [35, 36].

Therefore, optimization based keyword extraction provides a more flexible, robust, and mathematically grounded alternative to conventional single-score methods, making it particularly suitable for handling heterogeneous and complex textual data.

1.7 Problem Statement

Automatic keyphrase extraction is formulated as an optimization problem for unsupervised keyword ranking and selection.

1.8 Organization of the Thesis

This thesis is organized into seven chapters that collectively address the problem of semantic keyword extraction across heterogeneous data sources.

Chapter 1 introduces the research problem and presents the background, motivation, challenges, and problem statement associated with keyword extraction in diverse textual and multimodal environments. Chapter 2 reviews the existing literature on keyword and keyphrase ex-

traction methods across textual documents, social media text, multimodal data, and application-oriented tasks. The chapter also identifies the major research gaps and presents the research objectives of the thesis.

Chapter 3 presents the proposed optimization-driven frameworks for keyword extraction from structured textual documents. Chapter 4 focuses on keyword extraction from social media text. It presents methods designed to handle short, noisy, and context-dependent content. Chapter 5 extends keyword extraction to multimodal settings, where semantic meaning depends on the interaction between textual and visual information. Chapter 6 demonstrates the utility of extracted keywords in downstream applications, including toxicity detection, humor recognition, and offensive content identification.

Finally, Chapter 7 summarizes the major findings of the thesis, highlights the key contributions, and discusses potential directions for future research.

Chapter 2

Literature Review

A review of the existing literature is essential for understanding the development of keyword extraction methods and the limitations that remain unresolved. Over time, keyword extraction has progressed from simple statistical techniques based on frequency and position to more advanced methods incorporating semantic information, graph structures, contextual embeddings, machine learning, deep learning, and optimization strategies.

The evolution of these methods has been closely linked to the changing nature of textual data. While early approaches were primarily developed for long and structured documents, recent research increasingly addresses short social media text, multimodal content, and application-specific tasks. Consequently, keyword extraction methods have become more sophisticated in order to improve semantic relevance, contextual understanding, and robustness across diverse data types.

This chapter reviews the literature on keyword extraction across three major categories: structured textual documents, social media text, and multimodal and application-oriented data. For each category, the major methods, key contributions, and existing limitations are discussed, followed by the identification of research gaps that motivate the work presented in this thesis.

2.1 Keyword Extraction Methods for Structured Textual Documents

Structured textual documents constitute the earliest and most widely studied setting for keyword extraction research. These documents include research papers, books, technical reports, news articles, and other forms of formal written content. Since such documents are generally longer and more organized than social media text, they provide richer contextual information for identifying important keywords and keyphrases.

A wide range of methods has been developed for keyword extraction from structured documents. Early approaches mainly relied on statistical features such as frequency, position, and co-occurrence. Later studies introduced graph-based methods, embedding-based representations, machine learning models, deep learning techniques, and optimization-driven frameworks to better capture semantic importance and contextual relevance.

The following subsections review the major categories of keyword extraction methods for structured textual documents, along with their main contributions and limitations.

2.1.1 Statistical Approaches

Statistical approaches represent the earliest and most widely used category of keyword extraction methods. These methods are based on the assumption that important terms can be identified using simple statistical properties such as frequency, position, co-occurrence, and distribution within a document. Because of their simplicity and low computational cost, statistical methods have been extensively applied to structured textual documents.

One of the most influential statistical methods is Term Frequency-Inverse Document Frequency (TF-IDF), which measures the importance of a term according to its frequency within a document and its rarity across a collection of documents [37]. TF-IDF has been widely used for keyword extraction because it is simple, efficient, and easy to interpret. However, the method is highly dependent on term frequency and often fails to capture semantic meaning or contextual relevance.

Several extensions of frequency-based approaches have subsequently been proposed. YAKE [38] introduced a statistical framework based on multiple local features, including term frequency, word position, casing, sentence position, and term dispersion. RAKE [39] further

improved keyword extraction by ranking candidate phrases according to word co-occurrence patterns and degree-to-frequency ratios. These methods are unsupervised and domain independent, making them suitable for a wide range of textual documents.

Other statistical methods have attempted to improve keyword extraction by incorporating topic-level and feature-weighting information. Latent Dirichlet Allocation-based approaches combine topic modeling with multi-feature weighting to identify keywords that are both frequent and semantically related to latent topics within the document [40]. Similarly, [41] incorporated TF-IDF features into neural network-based representations, while [42] explored PCA-based feature optimization to improve keyword selection.

Recent studies have further extended statistical methods by incorporating entropy-based ranking, adaptive weighting, and hybrid scoring mechanisms. Entropy-based methods estimate the informativeness of candidate terms according to their distribution across the document, while AdaptiveUKE [43] combines statistical information with gated topic modeling to improve unsupervised keyword extraction. Recent hybrid ranking models also integrate positional, structural, and statistical features to improve keyword quality in long documents.

Although statistical approaches are computationally efficient and easy to implement, they have several limitations. Most statistical methods rely heavily on frequency-based indicators and therefore may overlook semantically important terms that occur infrequently. They also fail to capture contextual meaning, semantic similarity, and long-range relationships between terms. As a result, purely statistical approaches are often insufficient for handling complex and semantically rich textual documents.

2.1.2 Graph-Based Approaches

Graph-based approaches model candidate words or phrases as nodes in a graph, while edges capture co-occurrence, syntactic, semantic, or positional relationships between them. The importance of a candidate term is then estimated using graph centrality measures such as degree centrality, PageRank, or eigenvector centrality. Unlike purely statistical methods, graph-based approaches capture the structural relationships among terms and therefore provide richer representations of document content.

Early graph-based keyword extraction methods were influenced by ranking algorithms such as HITS [44] and PageRank [45]. One of the most influential approaches in this category is TextRank [14], which constructs a co-occurrence graph of terms and ranks them using graph

centrality. TextRank became a foundational method for many later graph-based techniques due to its simplicity and unsupervised nature.

Several extensions of TextRank were subsequently proposed to improve keyword ranking. PositionRank [46] incorporated positional information into graph-based ranking by assigning higher importance to words appearing earlier in the document. TopicRank [47] grouped similar candidate phrases into topics before applying graph ranking, thereby reducing redundancy among extracted keywords. DegExt [48] explored degree-based graph features for keyword ranking, while KECNW [49] exploited word co-occurrence networks for identifying important keywords.

RAKE [39] also employs graph-based relationships by representing words as nodes and their co-occurrence as edges. Although RAKE is often considered a statistical method, its graph representation enables it to model relationships between candidate terms more effectively than simple frequency-based methods.

Recent graph-based methods have increasingly incorporated semantic, fuzzy, and clustering-based mechanisms. FRAKE [50] integrates graph centrality with textual features to improve robustness across different domains. TripleRank [51] improves graph-based ranking by emphasizing semantic diversity and reducing redundancy among extracted keywords.

Graph-based methods have also been adapted for short and noisy text. More recent methods such as HCUKE [52] combine hierarchical clustering with graph centrality, while SDRank [53] refines graph ranking by modeling both local and global dependencies among candidate terms.

Although graph-based approaches are effective in capturing structural relationships among terms, they still face several limitations. Most graph-based methods rely heavily on local co-occurrence patterns and may fail to capture deeper semantic meaning, contextual variation, and long-range dependencies. In addition, their performance is often sensitive to graph construction strategies, edge weighting mechanisms, and parameter settings, making them less adaptable across different domains and document types.

2.1.3 Embedding-Based Approaches

Embedding-based approaches were introduced to overcome the limitations of frequency-based and graph-based methods by capturing semantic similarity between words and phrases. These methods represent words, sentences, or documents as dense vector embeddings in a continuous

space, where semantically similar terms are positioned closer together. As a result, embedding-based approaches are more effective in identifying meaningful keywords even when important terms occur infrequently.

One of the earliest embedding methods is Word2Vec [54], which uses shallow neural networks to learn distributed word representations from large corpora. GloVe [55] further improved word embeddings by combining local context information with global word co-occurrence statistics. These embedding techniques enabled keyword extraction methods to move beyond simple frequency counts and capture deeper semantic relationships between terms.

The introduction of contextual language models further advanced embedding-based keyword extraction. BERT [56] generates context-sensitive word representations by considering both left and right contextual information. RoBERTa [57] improved the BERT framework through enhanced training strategies and larger datasets. Such contextual embeddings are particularly useful because the same word may have different meanings depending on the surrounding context.

Several keyword extraction methods have been built upon these embedding models. KeyBERT [58] uses BERT embeddings and cosine similarity to identify keywords that are semantically similar to the document representation. SIFRank [59] further improved embedding-based keyword extraction by combining sentence embeddings with word importance weighting, thereby improving the ranking of semantically relevant keyphrases.

Some studies have also integrated embeddings with clustering and topic modeling techniques. Early clustering-based methods [60, 61] grouped semantically related terms before ranking them. EATS [62] combined topic modeling with clustering schemes driven by genetic algorithms to generate representative keywords. More recent methods such as [63] incorporated structural and positional information into embedding-based keyword ranking, while AdaptiveUKE [43] combined embeddings with gated topic modeling to improve unsupervised keyword extraction.

Although embedding-based approaches are more effective than traditional methods in capturing semantic similarity and contextual meaning, they also have several limitations. Static embeddings such as Word2Vec and GloVe cannot fully represent context-dependent meanings, while contextual models such as BERT and RoBERTa are computationally expensive and require substantial resources. In addition, embedding-based methods may still struggle with redundancy, domain-specific terminology, and long-document representations.

2.1.4 Machine Learning and Deep Learning Approaches

Machine learning and deep learning approaches formulate keyword extraction as a classification, sequence labeling, or ranking problem. Unlike statistical and graph-based methods, these approaches learn complex relationships between candidate terms and document content using labeled data and semantic representations.

Early supervised methods relied on manually designed lexical, positional, and linguistic features. KEA [64] is one of the earliest supervised keyword extraction systems and uses features such as TF-IDF and word position for keyword classification. MAUI [65] further extended supervised extraction by incorporating additional semantic and linguistic features. Support Vector Machine-based methods [66, 67] improved keyword extraction by learning from candidate term representations derived from statistical and linguistic information.

Several studies treated keyword extraction as a sequence labeling problem. [68] applied Conditional Random Fields to identify keywords directly from text sequences, while [69] explored ranking-based learning strategies for keyword extraction. More recent work by [70] introduced pseudo-labeling techniques to improve keyword extraction in low-resource settings.

Recent deep learning approaches have increasingly focused on contextual learning and semantic feature integration. Phraseformer [16] combines graph embeddings with transformer representations to jointly capture structural and semantic information. Similarly, [71] proposed a multi-perspective neural framework that integrates lexical, syntactic, and semantic information for improved keyword ranking. Recent advances have also explored semantic refinement and large language model integration. [72] introduced diffusion-based refinement to improve the semantic consistency of extracted keyphrases.

Although machine learning and deep learning methods generally provide better performance than traditional approaches, they also have several limitations. Supervised methods require large annotated datasets, which are often expensive to construct. Deep learning models are computationally expensive, less interpretable, and may not generalize effectively across domains with limited training data.

2.1.5 Other Approaches

In addition to statistical, graph-based, embedding-based, and learning-based methods, several alternative approaches have been proposed for keyword extraction. These methods often com-

bine multiple techniques in order to improve semantic relevance, reduce redundancy, and adapt to different document types.

Entropy-based methods estimate keyword importance according to the information content and distribution of candidate terms within the document [73].

Clustering-based approaches have also been explored for keyword extraction. These methods group semantically related words or phrases before selecting representative keywords from each cluster. More recent work has used clustering for sample selection and semantic coverage [74]. Other methods incorporate structural, positional, and domain-specific information to improve keyword extraction in technical documents [63].

Optimization-driven methods have become increasingly important in recent years. SDRank [53] incorporates both local and global dependencies into the ranking process, while SEKE [7] uses an ensemble of specialized experts for keyword extraction. ForestRank [75] combines ensemble learning with decision-theoretic ranking strategies, and FUBAKE [76] integrates fuzzy logic, uncertainty modeling, and optimization mechanisms for keyword extraction.

Recent studies have also explored decision-making frameworks such as TOPSIS and Fuzzy TOPSIS for text ranking and keyword selection. These methods are useful because they allow the simultaneous consideration of multiple criteria, including semantic importance, contextual relevance, redundancy, and diversity [77, 78]. A recent example is TopS-Key [79], which combines contextual PCA, adaptive Shannon entropy weighting, and fuzzy TOPSIS to improve semantic keyphrase extraction through multi-criteria ranking.

Although these alternative approaches often provide greater flexibility and robustness than conventional methods, many of them are computationally complex and difficult to generalize across domains. In addition, hybrid methods often require careful feature selection, parameter tuning, and optimization in order to achieve consistent performance.ig

2.1.6 Limitations of Existing Approaches

Despite the large number of keyword extraction methods proposed for structured textual documents, several limitations remain unresolved. Statistical approaches are computationally efficient but rely heavily on frequency-based indicators and often fail to capture semantic meaning and contextual relevance. As a result, important low-frequency terms may be ignored, while highly frequent but less informative terms may be overemphasized.

Graph-based methods improve keyword extraction by modeling relationships among terms, but they are strongly dependent on graph construction strategies and local co-occurrence patterns. Most graph-based approaches do not effectively capture deeper semantic dependencies, contextual variation, or long-range relationships between words.

Embedding-based and deep learning methods provide stronger semantic representations, but they are computationally expensive and often require large amounts of training data. Many of these methods also suffer from redundancy in the extracted keyword set and may not generalize effectively across domains with limited labeled data. In addition, most contextual embedding models are difficult to interpret and provide limited transparency regarding why specific keywords are selected.

Another important limitation is that most existing methods rely on a single ranking mechanism or a fixed set of features. Such methods are generally unable to balance multiple relevance criteria simultaneously, including semantic importance, contextual relevance, positional significance, diversity, and redundancy reduction. Existing approaches also provide limited adaptability to domain-specific terminology and varying document structures.

These limitations indicate the need for more robust and optimization-driven keyword extraction frameworks that can integrate multiple relevance criteria, improve semantic understanding, reduce redundancy, and adapt effectively across different forms of structured textual documents.

2.2 Keyword Extraction Methods for Social Media Text

Keyword extraction from social media text has become an important research area due to the rapid expansion of user-generated content on platforms such as Twitter, Reddit, and online discussion forums. In contrast to structured textual documents, social media content is typically short, informal, and highly dynamic, with limited contextual information. The widespread use of hashtags, abbreviations, emojis, spelling variations, and code-mixed language introduces significant linguistic noise, making the identification of meaningful keywords particularly challenging. Moreover, important terms in such text often appear infrequently, reducing the effectiveness of traditional frequency-based methods. These characteristics necessitate the development of specialized keyword extraction techniques that can operate effectively under sparse and noisy conditions.

Early approaches to keyword extraction in social media focused on graph-based representa-

tions of short text, where candidate terms are modeled as nodes and their relationships are captured through co-occurrence or dependency structures. Centrality-based ranking methods were then applied to identify important keywords, demonstrating that structural relationships can partially compensate for limited contextual information [80, 81]. These methods primarily rely on graph topology to rank candidate terms, highlighting the importance of connectivity patterns in determining keyword significance. However, such approaches remain limited in capturing deeper semantic meaning, as they primarily rely on local relationships between words.

To address these limitations, hybrid approaches have been proposed that integrate multiple sources of information. Methods combining named entity recognition with disambiguation techniques enable the extraction of semantically meaningful entities beyond simple word statistics [82]. Such approaches leverage external knowledge bases and contextual cues to improve the identification of relevant entities in short text. Task-oriented keyword extraction frameworks further extend this idea by selecting keyphrases based on their relevance to specific objectives rather than purely statistical importance [83]. This methodology introduces a task-aware ranking mechanism, allowing keyword selection to align with downstream application requirements. This shift reflects a broader transition toward context-aware and application-driven keyword extraction in social media environments.

Subsequent research has focused on enhancing graph-based methods by incorporating semantic, fuzzy, and semi-supervised mechanisms. Semantic graph-based ranking models improve keyword quality by integrating contextual relationships and ranking strategies [84], where semantic similarity and graph structure are jointly considered to refine keyword importance. Semi-supervised approaches based on lexical chains and centrality measures capture semantic continuity across terms and improve keyword identification in sparse text [85], enabling better propagation of contextual relevance across related terms. In addition, fuzzy graph-based methods introduce uncertainty modeling into term relationships, enabling more robust keyword extraction in noisy and ambiguous environments [86, 87]. These methods assign degrees of association rather than binary relationships, which is particularly effective for handling imprecise language. Unsupervised statistical ranking methods such as SwiftRank have also been adapted for short text, demonstrating effectiveness in identifying salient terms with minimal supervision [88]. Such methods rely on statistical scoring and ranking heuristics that are computationally efficient and adaptable to real-time data.

Recent advancements have increasingly focused on learning-based and neural approaches to

better capture semantic and contextual information. Deep learning-based models treat keyword extraction as a supervised task and learn representations that improve semantic relevance [89]. These models utilize neural architectures to capture complex feature interactions and contextual dependencies within short text. Sequence labeling formulations further enhance performance by modeling dependencies between neighboring tokens [90], allowing the model to identify multi-word keyphrases more effectively. Multi-feature ranking frameworks have also been proposed to integrate semantic, structural, and positional information, thereby improving keyword extraction in constrained text environments [91]. Such approaches combine multiple scoring criteria to improve robustness and adaptability.

Embedding-based approaches have further advanced keyword extraction in social media text. Contextual word embeddings provide dense semantic representations that capture variations in meaning depending on context, enabling more accurate identification of relevant keywords [92]. These representations allow the model to distinguish between different semantic usages of the same term. More recently, hybrid approaches combining graph-based ranking with large language models have been proposed to jointly leverage structural relationships and contextual semantics, resulting in improved robustness in noisy and context-limited environments [21]. This integration enables the model to benefit from both global semantic understanding and local structural information.

Despite these advancements, keyword extraction from social media text remains a challenging problem. Sparse context, multilingual and code-mixed content, sarcasm, and rapidly evolving vocabulary continue to limit the effectiveness of existing methods. Furthermore, many approaches are tailored to specific datasets or platforms, reducing their generalizability. In addition, most existing methods rely on a limited set of features and do not effectively balance multiple relevance criteria such as semantic importance, contextual relevance, diversity, and redundancy. These limitations highlight the need for more adaptive, context-aware, and multi-objective keyword extraction frameworks capable of operating effectively across diverse and dynamic social media environments.

2.2.1 Limitations of Existing Approaches

Despite notable progress, keyword extraction from social media text continues to face several challenges. The inherent sparsity of short text limits contextual understanding, reducing the effectiveness of statistical and structural methods. Additionally, linguistic noise such as

abbreviations, emojis, and code-mixed language hampers reliable semantic interpretation and generalization.

Graph-based and learning-based approaches improve performance but remain limited by local context, data dependency, and computational cost. Moreover, most methods optimize a single criterion and fail to balance multiple factors such as relevance, diversity, and redundancy.

Finally, many models are domain-specific and struggle to adapt to evolving topics and diverse platforms. These limitations highlight the need for adaptive, context-aware, and multi-objective keyword extraction frameworks.

2.3 Multimodal Keyword Extraction

The increasing availability of multimodal data has extended keyword extraction beyond traditional text-based analysis to settings where information is distributed across multiple modalities such as audio, images, and video. In many real-world scenarios, textual content alone is insufficient to capture the complete semantic meaning, as visual and acoustic signals provide complementary contextual cues. This has led to the development of multimodal keyword extraction methods that integrate heterogeneous data sources to improve the relevance, accuracy, and robustness of extracted keywords.

Unlike unimodal approaches, multimodal keyword extraction requires effective alignment and fusion of representations across different modalities. These methods aim to capture cross-modal dependencies, reduce ambiguity, and enhance semantic consistency by leveraging complementary information as given in Figure 2.1. As a result, keyword extraction is increasingly formulated as a cross-modal representation and ranking problem, where the objective is to identify meaningful keywords that are consistent across multiple sources of information.

A significant direction in multimodal keyword extraction focuses on audio-text alignment, particularly in open-vocabulary keyword spotting. Instead of treating keyword detection as a closed-set classification problem, recent approaches learn a shared embedding space where spoken audio and textual keyword representations are directly aligned. This enables flexible detection of user-defined keywords through cross-modal similarity [93]. Subsequent work addressed modality mismatch by introducing homogeneous embedding strategies that better align acoustic and textual representations [94]. Adversarial metric learning further improved alignment by suppressing modality-specific noise and enhancing generalization across unseen keywords [95]. More recent approaches incorporate attention-based mechanisms to model

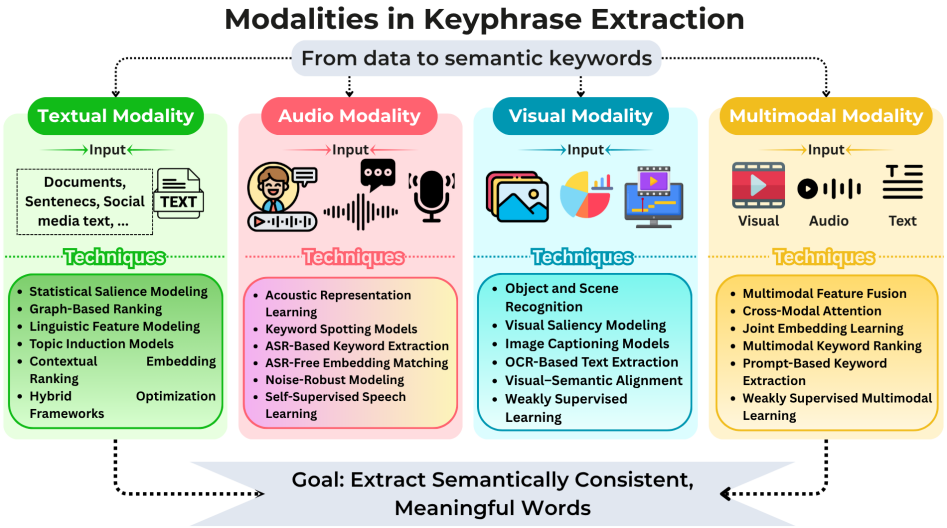

Figure 2.1: Modalities in keyword extraction illustrating the transition from unimodal to multimodal settings and the integration of textual, audio, and visual information.

fine-grained temporal correspondence between speech segments and textual representations, improving localization accuracy in continuous speech [96]. Additionally, contrastive learning frameworks have been proposed to distinguish between keyword and non-keyword segments, reducing false detections in streaming environments [97].

In the visual domain, early approaches primarily focused on image captioning, where semantic concepts are embedded implicitly within generated descriptions. However, such methods do not explicitly extract or rank keywords. More recent work has shifted toward direct visual keyword extraction, where models are trained to predict ranked keyword sets from images. Visual semantic attention mechanisms enable the identification of salient objects and concepts, allowing explicit keyword prediction instead of full sentence generation [98]. In parallel, document image analysis has emerged as an important area for keyword extraction. Deep neural models have demonstrated the ability to extract keywords from handwritten documents and complex layouts under noisy conditions [99]. Systems such as ChartOCR extend this capability by extracting semantically meaningful components such as titles, axis labels, and legends from charts, which naturally function as keywords for indexing [100]. Weakly supervised and adversarial learning approaches have further improved robustness while reducing dependence on large annotated datasets [101].

Recent advances emphasize stronger integration between visual and textual modalities. Multimodal keyphrase extraction frameworks incorporate visual entity enhancement to refine keyword salience, demonstrating that visual context can directly influence keyword ranking rather

than acting as auxiliary information [102]. Furthermore, self-supervised transformer-based models have been proposed for keyword extraction in historical and low-resource document images, leveraging masked learning objectives to improve robustness under varying conditions and limited supervision [103].

Video-based keyword extraction represents a more complex multimodal scenario, where audio, visual, and textual signals must be jointly modeled. Recent approaches explicitly treat keyword extraction as the primary objective by jointly modeling multimodal inputs. Frameworks designed for long-form instructional videos analyze transcripts, slide content, and temporal alignment to extract salient keyphrases representing core concepts [104]. Similarly, multimodal keyword and hashtag prediction methods integrate visual frames, audio signals, and textual metadata to generate ranked keyword sets grounded in multimodal evidence [105, 106].

2.3.1 Limitations of Existing Approaches

Despite significant progress, multimodal keyword extraction remains a challenging problem. A major limitation lies in the effective alignment of heterogeneous modalities, as differences in representation space, scale, and noise characteristics often lead to inconsistencies in learned embeddings. Many existing approaches rely on simple fusion strategies that fail to capture complex cross-modal interactions, limiting their ability to fully exploit complementary information.

Another critical challenge is the imbalance in modality contribution. In many methods, one modality often text dominates the extraction process, while other modalities contribute only marginally. This results in suboptimal keyword selection, particularly in scenarios where visual or acoustic cues carry essential semantic information.

Data-related challenges also persist. Multimodal keyword extraction often requires large, well-aligned datasets, which are difficult to obtain. Although weakly supervised and self-supervised approaches have been proposed, they still face limitations in generalization and robustness across diverse domains.

Furthermore, most existing methods focus on a limited set of objectives, such as similarity or ranking, without explicitly balancing multiple criteria including semantic relevance, contextual consistency, diversity, and redundancy. The absence of multi-objective optimization frameworks restricts the ability of current models to produce high-quality and interpretable keyword sets.

Finally, computational complexity remains a significant concern, particularly for large-scale or real-time applications involving multiple modalities. These limitations highlight the need for more adaptive, efficient, and optimization driven approaches that can effectively integrate multimodal information while maintaining semantic consistency and robustness.

2.4 Downstream Applications of Keyword Extraction

Keyword extraction has been increasingly utilized in downstream tasks where lexical units such as words and phrases serve as interpretable signals for decision-making. In these settings, keywords are not extracted solely for representation but are implicitly or explicitly aligned with specific task objectives, including content moderation, emotion understanding, and multimodal analysis.

In multimodal offensive content detection, several studies have demonstrated the importance of lexical cues in identifying harmful intent. The Hateful Memes framework [107] combines textual and visual information to detect implicit hate, where keywords such as slurs or context-dependent phrases act as critical indicators. Similarly, MultiOFF [108] employs multimodal fusion to classify offensive content, while MOMENTA [109] introduces attention-based mechanisms to capture interactions between textual and visual features. These approaches primarily rely on end-to-end representations, where keyword-level signals are embedded within learned features rather than explicitly extracted or optimized.

Subsequent work has focused on improving robustness and contextual reasoning. Adaptive learning strategies [110] and ensemble-based approaches [111] have been proposed to enhance performance across diverse and noisy inputs. Fairness-aware modeling [112] addresses bias in multimodal systems, while commonsense reasoning frameworks [113] incorporate external knowledge to better interpret implicit and context-dependent meanings. Although these methods improve prediction accuracy, they still lack explicit modeling of keyword-level representations, limiting interpretability.

In emotion and sentiment analysis, keyword-level cues play a central role in capturing affective signals. The Memotion benchmark [114, 115] introduces multi-label classification tasks involving humor, sarcasm, and offense, where keywords represent underlying emotional indicators. Several approaches utilize multimodal embeddings derived from models such as BERT [116], GloVe [55], and visual encoders to capture semantic and affective features. Multi-task learning frameworks [117] and joint fusion models [118] further improve performance by

modeling interactions across modalities. However, these methods treat keywords implicitly within representations rather than as structured outputs.

Knowledge-enhanced and representation-driven approaches have also been explored to improve semantic grounding. Models incorporating external knowledge and multimodal representations, such as CLIP [119] and knowledge-aware frameworks [120], improve alignment between textual and visual semantics. These methods enhance contextual understanding but still rely on latent feature representations without explicitly extracting or optimizing keyword sets.

In video and social media applications, keyword-level outputs are often generated through tagging and recommendation systems. Multimodal hashtag prediction models [105, 106] integrate visual, textual, and audio signals to produce ranked keyword sets. While these approaches effectively generate keywords, they are typically optimized for recommendation objectives rather than semantic relevance or interpretability.

Across these applications, a common limitation is that keyword extraction is rarely treated as a primary objective. Instead, keywords are embedded within learned representations or used as auxiliary signals, resulting in limited interpretability and reduced control over keyword selection. This highlights the need for task-aware keyword extraction frameworks that explicitly model keywords as primary outputs and align them with downstream objectives, enabling more interpretable, robust, and application-specific solutions.

2.4.1 Limitations of Existing Approaches

Despite extensive use of keyword-level signals in downstream applications, several limitations remain. A key issue is that most existing methods do not treat keyword extraction as a primary objective. Instead, keywords are embedded within latent representations or used as auxiliary features, which limits interpretability and reduces control over keyword selection.

Another limitation is the lack of task-aware keyword modeling. Current approaches often rely on generic feature learning without explicitly aligning keyword extraction with application-specific objectives, resulting in suboptimal performance in tasks requiring fine-grained semantic understanding.

Furthermore, most methods depend on implicit multimodal correlations rather than structured keyword-level reasoning. This restricts their ability to capture meaningful lexical cues, particularly in complex scenarios involving ambiguity, sarcasm, or contextual dependencies.

Scalability and generalization also remain challenges. Many models are tailored to specific datasets or domains and struggle to adapt to diverse and evolving content. In addition, the absence of multi-objective optimization limits the ability to balance competing factors such as relevance, interpretability, and robustness.

These limitations highlight the need for explicit, task-aware, and optimization-driven keyword extraction frameworks that can better support downstream applications while maintaining interpretability and semantic consistency.

2.5 Research Gaps

A cross-domain analysis of keyword extraction methods reveals several overarching limitations that persist irrespective of data type, modeling strategy, or application setting.

A central issue is the absence of a unified framework for estimating keyword relevance. Existing methods rely on disparate scoring mechanisms that are often loosely integrated, leading to fragmented decision-making. The lack of joint modeling of multiple relevance factors results in keyword sets that may capture isolated aspects of importance but fail to provide a holistic representation of content.

Another key limitation lies in the inability of current approaches to robustly handle contextual variability and uncertainty. While advancements in representation learning have improved semantic modeling, most methods remain sensitive to variations in input structure, noise levels, and domain characteristics. This limits their effectiveness in dynamic environments where meaning is highly context-dependent.

A further gap is the weak coupling between keyword extraction and end-task objectives. In many cases, keyword extraction is treated as a standalone process, without explicitly considering how the extracted keywords will be utilized. This disconnect reduces the effectiveness of keywords in supporting downstream tasks and limits their interpretability in practical applications.

In multimodal scenarios, the challenge extends to the integration of heterogeneous information sources. Existing approaches often lack mechanisms for consistent semantic alignment across modalities, resulting in incomplete or conflicting keyword representations. This highlights the need for more coherent cross-modal modeling strategies.

Finally, scalability and adaptability remain open challenges. Many existing models are de-

signed under constrained settings and do not generalize effectively across domains, data distributions, or evolving content streams. This restricts their applicability in real-world systems where flexibility and robustness are essential.

These observations collectively indicate the need for more integrated, adaptive, and task-aware keyword extraction frameworks that can jointly model multiple relevance criteria, accommodate uncertainty, and operate effectively across diverse and multimodal environments.

2.6 Research Objectives

The objectives of this thesis are as follows:

1. To conduct a comprehensive literature survey.
2. To develop a model by the fusion of NLP and MCDM techniques that enhances the process of automatic keyphrase extraction.
3. To advance the keyphrase extraction process with a hybrid model that combines various optimization techniques.
4. To foster the swarm optimization techniques in extracting keyphrases automatically.
5. To integrate keyphrase extraction with optimization techniques for downstream NLP applications.

Chapter 3

Keyword Extraction from Textual Documents

Keyword extraction from textual documents is a fundamental task in natural language processing, aimed at identifying representative terms that capture the core semantics of a document. Despite significant advancements, existing methods often rely on isolated relevance criteria and lack optimization-driven mechanisms, leading to suboptimal keyword selection. This chapter focuses on developing robust methodologies for keyword extraction by integrating linguistic processing with optimization-based strategies.

The chapter begins with text preprocessing, which forms the foundation for candidate term generation. Standard preprocessing techniques are first discussed, followed by a proposed extended preprocessing strategy designed to enhance semantic representation. Subsequently, three proposed approaches, *FuBAKE*, *TopS-Key*, and *ForestRank*, are introduced to improve keyword extraction through hybrid optimization and multi-criteria decision-making.

3.1 Text Preprocessing

Text preprocessing is a fundamental stage in keyword extraction, as it transforms raw textual data into a structured and analyzable form. Real-world text often contains noise, inconsis-

tencies, and redundant elements that can adversely affect the identification of meaningful keywords. Therefore, preprocessing aims to standardize the text, eliminate irrelevant components, and retain linguistically significant information. An effective preprocessing pipeline enhances the quality of candidate terms and provides a reliable foundation for subsequent keyword extraction methods.

This section presents both conventional preprocessing techniques and an extended preprocessing strategy. The standard methods focus on linguistic normalization and candidate term generation, while the proposed approach introduces enhancements to improve semantic representation and selection quality.

3.1.1 Standard Preprocessing and Linguistic Processing

Standard preprocessing constitutes a foundational stage in keyword extraction, as it transforms raw textual data into a structured and linguistically enriched representation. Textual documents typically contain noise, inconsistencies, and redundant elements that can obscure meaningful patterns and hinder the identification of representative keywords. Therefore, preprocessing aims to normalize the text, eliminate irrelevant components, and extract linguistically significant units that can serve as reliable candidate keyphrases.

Let a document be represented as a sequence of tokens ($T = \{t_1, t_2, \ldots, t_n\}$), where each t_i corresponds to a lexical unit derived from the original text. The objective of preprocessing is to construct a refined candidate set $C \subseteq T$, containing only informative and semantically relevant terms that can be further analyzed for keyword extraction.

The preprocessing pipeline begins with tokenization, which segments the raw document D into smaller units called tokens. These tokens typically correspond to words or meaningful symbols and form the basic units for subsequent analysis [121]. Formally, tokenization defines a mapping $\mathscr{T} : D \to T$. This step is essential as it enables the system to operate at the word level, allowing individual lexical units to be analyzed, filtered, and combined.

Following tokenization, text normalization is performed to ensure consistency in representation. This includes operations such as converting text to lowercase, removing special characters, and handling formatting variations. By reducing inconsistencies caused by different writing styles, normalization improves the robustness and comparability of tokens across the document.

Subsequently, stopword removal is applied to eliminate frequently occurring words that carry

minimal semantic value, such as articles, prepositions, and conjunctions. These terms often do not contribute meaningfully to the document content and may introduce noise in keyword extraction. Let S denote the stopword set; the filtered token set is defined as Eq. 3.1.

$$T' = \{t_i \in \tilde{T} \mid t_i \notin S\} \tag{3.1}$$

By removing such non-informative tokens, this step enhances the discriminative quality of the remaining terms and focuses the analysis on content-bearing words.

Lemmatization is then employed to perform morphological normalization by reducing words to their canonical or base forms. Unlike stemming, which applies heuristic truncation, lemmatization utilizes linguistic knowledge to produce valid base forms. For instance, words such as "running," "runs," and "ran" are mapped to "run." This process reduces redundancy and ensures that semantically equivalent variations are treated uniformly, thereby improving the consistency of the representation [122].

To incorporate syntactic information, Part-of-Speech (POS) tagging is performed, assigning grammatical categories such as nouns, verbs, and adjectives to each token based on its role in the sentence. This information is particularly valuable for keyword extraction, as important terms are often associated with nouns and their modifiers [123]. Formally, a tagging function $\mathrm{POS}(t_i)$ assigns a label to each token, and a linguistically constrained candidate set is defined as Eq. 3.2.

$$C_1 = \{t_i \in T'' \mid \mathrm{POS}(t_i) \in \{\mathrm{NOUN}, \mathrm{ADJ}, \mathrm{NUM}\}\} \tag{3.2}$$

This filtering ensures that only structurally relevant tokens are retained for further processing.

While individual tokens provide useful information, meaningful keywords often appear as multi-word expressions. Therefore, noun phrase extraction is performed to identify syntactically coherent phrases. Noun phrases are constructed by grouping consecutive tokens whose POS tags belong to the set $\{\mathrm{NOUN}, \mathrm{ADJ}, \mathrm{NUM}\}$ until a token outside this set or a punctuation boundary is encountered. Formally, a noun phrase p_j is defined as $p_j = (t_i, t_{i+1}, \ldots, t_{i+k})$ such that $\mathrm{POS}(t_{i+m}) \in \{\mathrm{NOUN}, \mathrm{ADJ}, \mathrm{NUM}\}$ for all $m = 0, \ldots, k$. The set of extracted phrases is denoted as $C_2 = \{p_1, p_2, \ldots, p_r\}$. This step enhances semantic representation by capturing complete concepts, such as "deep neural network model," rather than isolated words.

In addition to syntactic structures, domain-specific entities also play a crucial role in iden-

tifying important information. Named Entity Recognition (NER) is therefore employed to detect and classify entities such as persons, organizations, locations, and dates within the text [124]. These entities often represent key concepts and contribute significantly to the semantic understanding of the document. The entity set is defined as $E = \{t_i \mid \text{NER}(t_i) \neq \emptyset\}$.

Finally, the candidate keyphrase set is constructed by combining all extracted components $C = C_1 \cup C_2 \cup E$. This unified set integrates linguistically valid tokens, semantically rich phrases, and domain-specific entities, providing a comprehensive pool of candidate terms for keyword extraction.

The overall preprocessing pipeline can thus be summarized as Eq. 3.3.

$$D \to T \to \tilde{T} \to T' \to T'' \to C \tag{3.3}$$

Although these standard preprocessing techniques significantly improve the quality of candidate terms by reducing noise and enforcing linguistic constraints, they remain predominantly rule-based and do not incorporate semantic weighting or optimization. Consequently, they may not fully capture deeper contextual relationships or effectively eliminate redundancy, which motivates the extended preprocessing strategy presented in the subsequent subsection.

3.1.2 Proposed Extended Preprocessing Strategy

Although standard preprocessing techniques effectively normalize textual data and extract linguistically relevant candidate terms, they often result in the generation of redundant or highly similar phrases. Such redundancy typically arises due to variations in word ordering, minor lexical differences, and morphological inconsistencies. Consequently, multiple candidate phrases may correspond to the same underlying semantic concept, leading to duplication and reduced efficiency in subsequent keyword extraction stages. To overcome this limitation, an extended preprocessing strategy is proposed, which incorporates similarity-based filtering to refine and consolidate the candidate set.

The proposed approach employs fuzzy string matching to identify and eliminate duplicate or near-duplicate phrases. Unlike exact matching, fuzzy matching enables the comparison of strings that are not identical but exhibit a high degree of similarity. This property is particularly useful in natural language processing, where semantically equivalent phrases often appear in slightly different forms [125]. Among various similarity measures, the token set ratio (TSR) is

adopted in this work due to its ability to effectively capture token-level overlap while remaining robust to variations in word order.

Let p_i and p_j denote two candidate phrases, and let S_i and S_j represent their corresponding sets of unique tokens. The similarity between these phrases is computed using the token set ratio defined as Eq. 3.4.

$$TSR(p_i, p_j) = 100 \times \max\left(\text{Sim}(S_c, S_c \cup S_{i'}), \text{Sim}(S_c, S_c \cup S_{j'}), \text{Sim}(p_i, p_j)\right) \tag{3.4}$$

where $S_c = S_i \cap S_j$ denotes the set of common tokens, and $S_{i'} = S_i - S_c$, $S_{j'} = S_j - S_c$ represent the non-overlapping tokens. This formulation ensures that both shared and distinct components of the phrases are considered during similarity computation.

The similarity function $\text{Sim}(X, Y)$ is derived from normalized edit distance, defined as Eq. 3.5.

$$\text{Sim}(X, Y) = 1 - \frac{d(X, Y)}{\max(\text{len}(X), \text{len}(Y))} \tag{3.5}$$

where $d(X, Y)$ represents the Levenshtein distance [126]. The Levenshtein distance is a widely used string similarity metric that quantifies the minimum number of single character operations: insertions, deletions, or substitutions required to transform one string into another.

For instance, converting the word “kitten” to “sitting” requires three such operations. By normalizing this distance, the similarity score is constrained within the range $[0, 1]$, enabling consistent comparison across phrases of varying lengths. To ensure effective removal of redundancy, a similarity threshold θ is introduced. If $\text{TSR}(p_i, p_j) \geq \theta$, the two phrases are considered semantically equivalent, and only one representative phrase is retained. Otherwise, both phrases are preserved as distinct candidates. A relatively high threshold is employed to ensure strict filtering, thereby eliminating only strongly similar phrases while preserving meaningful diversity within the candidate set.

After applying the TSR-based filtering mechanism, the refined set of candidate keyphrases is denoted as $\mathscr{K} = \{k_1, k_2, \ldots, k_m\}$, where each k_i represents a unique and representative phrase derived from the initial candidate pool. This refined set forms a more compact and semantically consistent basis for subsequent keyword extraction.

Overall, the proposed extended preprocessing strategy enhances the quality of candidate terms by systematically reducing redundancy and consolidating semantically equivalent phrases. In contrast to standard preprocessing, which primarily relies on rule-based filtering, this approach

introduces a similarity-aware mechanism that captures subtle linguistic variations. As a result, the preprocessing stage becomes more robust and better aligned with the objective of accurate and efficient keyword extraction.

3.2 FUBAKE Methodology

This section presents the proposed FUBAKE (Bayesian Gamified Fuzzy Ant Colony Optimization for Semantic-Driven Unsupervised Keyphrase Extraction) methodology for extracting representative keyphrases from textual documents. The framework integrates fuzzy logic, Bayesian modeling, and ant colony optimization to address key challenges in unsupervised keyword extraction, including redundancy, uncertainty, and semantic relevance.

The overall workflow of the proposed methodology is illustrated in Figure 3.1.

The process begins with preprocessing and candidate keyphrase generation, followed by the construction of a co-occurrence graph $G = (V, E)$, where nodes represent candidate phrases and edges capture their co-occurrence relationships within the document.

In the next stage, multiple features are initialized for each candidate, including pheromone intensity, desirability, and Bayesian belief. These features incorporate statistical, semantic, and structural information, enabling a comprehensive representation of candidate importance. To handle uncertainty in these features, fuzzy logic is applied, where values are mapped to linguistic variables using membership functions and subsequently converted into crisp scores through defuzzification.

The optimization process is driven by an ant colony-based mechanism, in which artificial agents probabilistically explore the graph to identify high-quality paths. A roulette wheel selection strategy is employed to guide path selection, while a payoff function evaluates the quality of selected paths. Based on this evaluation, pheromone intensity and Bayesian belief are iteratively updated, reinforcing promising candidates and suppressing less relevant ones until convergence is achieved.

Finally, the selected paths are analyzed to extract keyphrases, which are then ranked based on their importance. The output is a ranked list of predicted keyphrases that effectively represent the semantic content of the document.

The detailed formulation of each component of the FUBAKE methodology is presented in the following subsections.

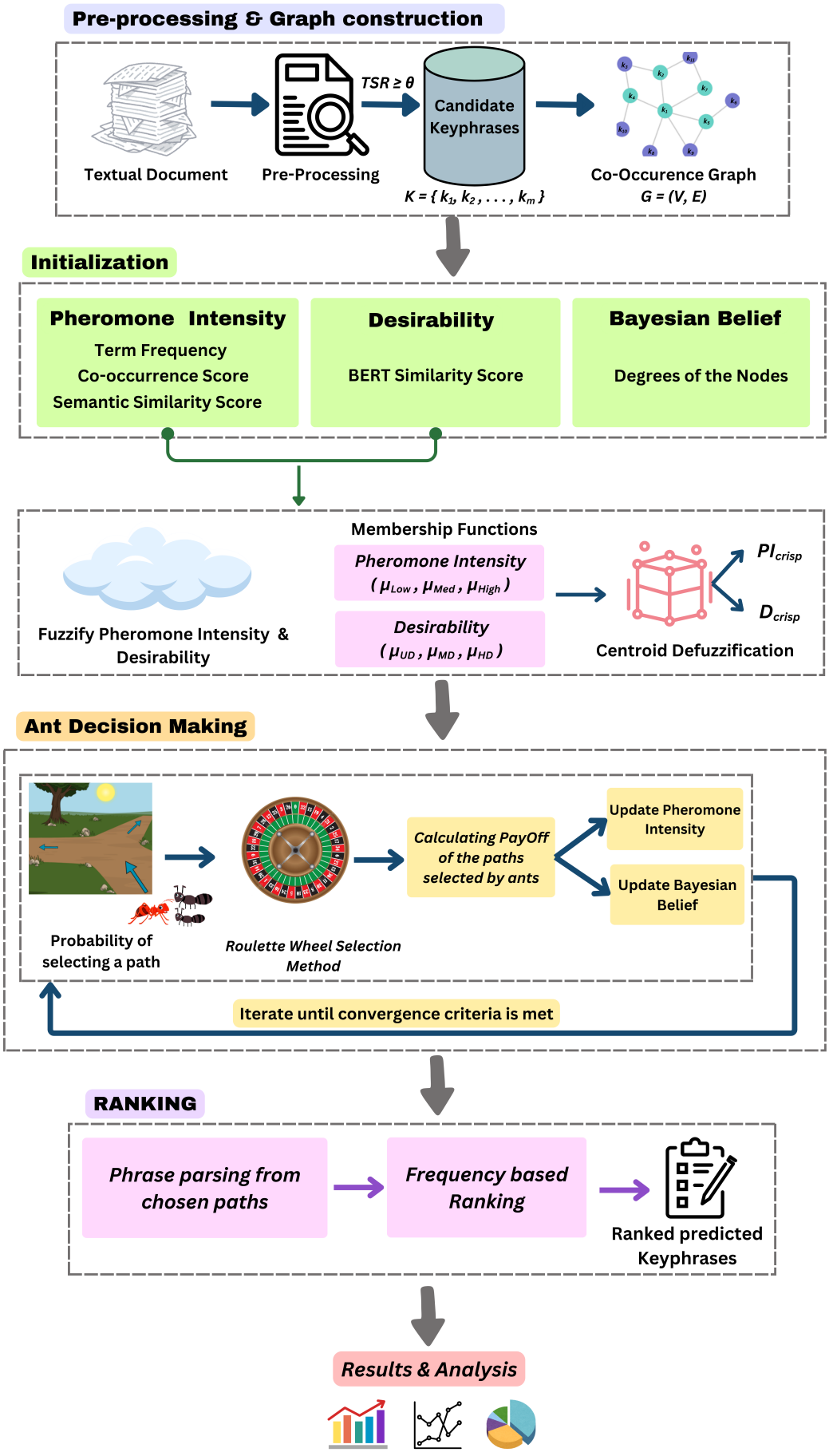

Figure 3.1: Overall workflow of the proposed FUBAKE methodology

3.2.1 Candidate Keyphrase Preparation and Co-occurrence Graph Construction

After applying the preprocessing and extended candidate refinement procedures described in Section 3.1, the document is represented as a refined collection of unique candidate keyphrases. At this stage, the focus shifts from linguistic filtering to structural modeling, where the extracted candidate phrases are organized into a semantic graph for further analysis. Let the final set of refined candidate keyphrases be represented as $\mathscr{K} = \{k_1, k_2, \ldots, k_m\}$, where each k_i denotes a distinct candidate phrase retained after linguistic preprocessing and redundancy reduction.

To evaluate the contextual importance of these candidate phrases, it is necessary to model the relationships that exist among them throughout the document. For this purpose, a co-occurrence graph is constructed, enabling candidate phrases to be represented not as isolated units but as interconnected semantic components. This graph-based formulation captures how frequently and in what contexts candidate phrases appear together, thereby providing a structural foundation for subsequent optimization.

Consider a document as an ordered sequence of sentences $D = \{s_1, s_2, \ldots, s_k\}$, where each sentence s_i contributes contextual information to the overall document structure. Based on the refined candidate set $\mathscr{K}$, a graph is defined as $G = (V, E)$, where V represents the set of nodes corresponding to candidate keyphrases, and E represents the set of edges that encode contextual associations between them.

A direct connection is established between two candidate phrases when they appear within the same sentence, as sentence-level co-occurrence often indicates strong contextual or thematic association. This relationship is formally defined as Eq. 3.6.

$$E(k_i, k_j) = 1 \quad \text{if } k_i \text{ and } k_j \text{ occur in the same sentence } s_q \tag{3.6}$$

While sentence-level relationships capture immediate semantic interactions, relevant concepts in textual documents frequently extend across neighboring sentences. Restricting associations to individual sentences may therefore overlook broader contextual dependencies. To incorporate wider semantic continuity, an extended context window is introduced around each sentence. For a given sentence s_i, the contextual neighborhood is defined as Eq. 3.7.

$$\mathscr{C}(s_i) = \{s_j \mid j \in [i-3, i+3],\ j \neq i\} \tag{3.7}$$

This contextual range includes up to three preceding and three succeeding sentences, enabling the graph to capture medium-range semantic relationships while maintaining contextual relevance. If two candidate phrases appear within this expanded sentence neighborhood, they are also considered contextually related and connected within the graph.

To preserve the independent significance of each candidate phrase, self-connections are additionally incorporated as $E(k_i, k_i) = 1$.

These self-loops ensure that each node retains intrinsic relevance, even when external co-occurrence connections are sparse.

The resulting co-occurrence graph provides a structured representation of semantic interactions among candidate keyphrases by integrating both local sentence-level associations and broader contextual dependencies. Through this representation, the document is transformed into an interconnected semantic network that serves as the basis for subsequent stages of the FUBAKE framework, including feature initialization, fuzzy modeling, Bayesian reasoning, and optimization.

3.2.2 Initialization of Pheromone Intensity, Desirability, and Bayesian Belief

Once the co-occurrence graph has been constructed, the next stage of the FUBAKE framework focuses on initializing the fundamental parameters that guide the optimization process. In Ant Colony Optimization (ACO), the quality of search depends heavily on how effectively the search space is initialized. Since the proposed framework aims to identify semantically meaningful and contextually important keyphrases, initialization must incorporate multiple dimensions of relevance rather than relying on a single scoring criterion.

Accordingly, three complementary components are established at the outset: pheromone intensity, desirability, and Bayesian belief. Together, these components provide a balanced representation of candidate keyphrase significance by integrating statistical prominence, lexical cohesion, semantic relatedness, and graph-structural importance. Pheromone intensity serves as the initial attractiveness of candidate paths, desirability captures contextual semantic similarity, and Bayesian belief models structural uncertainty within the graph. This multi-dimensional

initialization forms the basis for subsequent optimization and path selection. The methodology described in this subsection is adapted from the proposed FUBAKE framework presented in the associated research work, but is elaborated here in a more comprehensive thesis-oriented form.

3.2.2.1 Initialization of Pheromone Intensity

Pheromone intensity represents the initial strength assigned to relationships between candidate keyphrases and acts as a guiding signal during the ACO search process. In biological ant systems, pheromones represent chemical trails that influence path selection; analogously, in FUBAKE, pheromone intensity quantifies the attractiveness of traversing between keyphrase pairs in the co-occurrence graph.

To ensure that this initialization captures multiple aspects of relevance, pheromone intensity is derived from three complementary measures: term frequency, lexical co-occurrence, and semantic similarity. These metrics collectively ensure that candidate keyphrases are evaluated not only by how frequently they appear, but also by how cohesively their constituent terms interact and how strongly they align with the document's semantic content.

The first component, Term Frequency (TF), measures the statistical prominence of a candidate keyphrase within the document. Frequently occurring terms are generally more likely to represent central concepts. For a candidate keyphrase k, TF is defined as Eq. 3.8.

$$\mathrm{TF}(k) = \frac{\sum_{w \in k} \mathrm{count}(w, D)}{\text{Total Words in } D} \tag{3.8}$$

where $\mathrm{count}(w, D)$ denotes the number of occurrences of word w in document D. This normalization ensures comparability across documents of varying lengths.

While TF captures statistical prominence, frequency alone does not guarantee phrase integrity. Therefore, lexical cohesion is incorporated through a co-occurrence score, which evaluates how consistently the constituent words of a candidate phrase appear together within sentence-level contexts. For a keyphrase $k = \{w_1, w_2, \ldots, w_n\}$, the co-occurrence score is defined as Eq. 3.9.

$$c(k) = \frac{1}{\binom{n}{2}} \sum_{i<j} \sum_{s \in D} \mathbf{1}(w_i \in s \wedge w_j \in s) \tag{3.9}$$

where the indicator function equals 1 when the word pair co-occurs within sentence s, and

0 otherwise. This formulation measures internal phrase cohesion and ensures that candidate phrases are not merely frequent but structurally meaningful.

Beyond statistical and lexical dimensions, semantic relevance is also essential. A phrase may occur frequently and cohesively yet still be peripheral to the document's primary theme. To address this, semantic similarity is computed by embedding both the candidate phrase and the full document into a vector space and measuring their cosine similarity as Eq. 3.10.

$$\text{Semantic Similarity}(k, D) = \frac{\mathbf{v}_k \cdot \mathbf{v}_D}{\|\mathbf{v}_k\| \, \|\mathbf{v}_D\|} \tag{3.10}$$

where $\mathbf{v}_k$ and $\mathbf{v}_D$ denote the vector representations of the keyphrase and document, respectively. Higher similarity values indicate stronger thematic alignment.

Since ACO operates on graph edges rather than isolated nodes, these metrics are extended to keyphrase pairs. For two candidate phrases k_I and k_{II}, pairwise relevance is estimated by averaging their individual scores by Eqs. 3.11, 3.12, 3.13.

$$\text{Avg TF}(k_I, k_{II}) = \frac{\text{TF}(k_I) + \text{TF}(k_{II})}{2} \tag{3.11}$$

$$\text{Avg Co-occur}(k_I, k_{II}) = \frac{c(k_I) + c(k_{II})}{2} \tag{3.12}$$

$$\text{Avg Semantic Sim}(k_I, k_{II}) = \frac{\text{SemanticSim}(k_I) + \text{SemanticSim}(k_{II})}{2} \tag{3.13}$$

The initial pheromone intensity is then defined as the arithmetic mean of these three dimensions evaluated using Eq. 3.14.

$$\text{Initial PI} = \frac{\text{Avg TF} + \text{Avg Co-occur} + \text{Avg Semantic Sim}}{3} \tag{3.14}$$

This balanced formulation ensures that statistical frequency, lexical cohesion, and semantic relevance contribute equally to initialization, thereby preventing overdependence on any single factor.

To maintain numerical stability and comparability across all graph edges, Min-Max normalization is applied using Eq. 3.15.

$$\text{PI}_{\text{normalized}} = \frac{\text{PI} - \min(\text{PI})}{\max(\text{PI}) - \min(\text{PI})} \tag{3.15}$$

This scaling projects pheromone values into the range $[0,1]$, ensuring a stable starting point for probabilistic exploration.

3.2.2.2 Initialization of Desirability

While pheromone intensity captures statistical and structural relevance, desirability measures contextual semantic compatibility between connected keyphrases. In FUBAKE, desirability reflects how semantically related two candidate phrases are within contextual embedding space.

To compute this measure, Bidirectional Encoder Representations from Transformers (BERT) embeddings are employed due to their ability to capture contextualized semantic meaning. For each keyphrase, token embeddings are aggregated to form a phrase-level representation, and the desirability between two keyphrases is computed as cosine similarity using Eq. 3.16.

$$\text{Des}(k_I, k_{II}) = \frac{\mathbf{v}_{k_I} \cdot \mathbf{v}_{k_{II}}}{\|\mathbf{v}_{k_I}\| \, \|\mathbf{v}_{k_{II}}\|} \tag{3.16}$$

where, $\mathbf{v}_{k_I}$ and $\mathbf{v}_{k_{II}}$ are the BERT-based embeddings of the corresponding keyphrases. Higher desirability values indicate stronger semantic compatibility, thereby encouraging ants to traverse semantically coherent paths.

3.2.2.3 Initial Bayesian Belief Calculation

In addition to pheromone intensity and desirability, FUBAKE incorporates Bayesian belief to explicitly model uncertainty and structural prior knowledge. Bayesian Game Theory provides a framework for rational decision-making under incomplete information, making it suitable for graph exploration where ants must evaluate uncertain path quality.

Bayesian belief is initialized on graph edges rather than nodes, as edges represent relationships between candidate phrases. For an edge connecting k_I and k_{II}, the initial belief is defined as Eq. 3.17.

$$\text{Belief}(k_I, k_{II}) = \frac{1}{\deg(k_I) + \deg(k_{II})} \tag{3.17}$$

where $\deg(k_I)$ and $\deg(k_{II})$ denote the graph degrees of the connected nodes.

This inverse-degree formulation assigns relatively lower prior belief to highly connected hub nodes, which may be broadly connected but less semantically distinctive, while emphasizing edges associated with more context-specific candidates. As a result, Bayesian belief acts as a structural regularization mechanism that balances graph connectivity with semantic specificity.

Collectively, pheromone intensity, desirability, and Bayesian belief provide a robust initialization framework for the ACO process. Pheromone intensity captures statistical and semantic prominence, desirability ensures contextual alignment, and Bayesian belief introduces structural uncertainty modeling. Together, these components establish a stable and semantically informed foundation for iterative optimization in the FUBAKE methodology.

3.2.3 Fuzzy Bayesian Ant Colony Optimization for Keyphrase Selection

Following the initialization of pheromone intensity, desirability, and Bayesian belief, the final stage of the FUBAKE framework performs iterative optimization to identify the most semantically relevant keyphrases from the candidate graph. This stage integrates Ant Colony Optimization (ACO), fuzzy logic, and Bayesian reasoning into a unified decision-making framework. While ACO provides an adaptive mechanism for graph traversal and solution discovery, fuzzy logic addresses uncertainty in dynamically evolving path attributes, and Bayesian belief introduces probabilistic prior knowledge for more informed path evaluation. Together, these components enable the framework to balance semantic relevance, contextual desirability, structural importance, and uncertainty during keyphrase selection.

3.2.3.1 Fuzzy Modeling of Pheromone Intensity and Desirability

In classical Ant Colony Optimization (ACO), pheromone intensity evolves continuously through pheromone deposition and evaporation as ants traverse different paths. However, in practical applications such as keyphrase extraction, the exact pheromone level associated with a graph edge is inherently uncertain. This uncertainty arises because ant movement is probabilistic, pheromone deposition depends on path quality, and evaporation continuously modifies previously accumulated intensities over time. Consequently, pheromone values cannot be regarded as perfectly deterministic quantities; instead, they behave as dynamically fluctuating indicators of path importance.

A similar limitation applies to desirability, which reflects the semantic usefulness of a traversal path. Although desirability is initially estimated using semantic and contextual measures,

its interpretation during optimization remains uncertain because graph traversal continuously changes the search dynamics. Therefore, representing desirability using only crisp numerical values may oversimplify the evolving semantic relationships among candidate phrases.

To address this limitation, fuzzy logic is incorporated to represent both pheromone intensity and desirability using linguistic categories rather than rigid numerical thresholds. Specifically, each edge is described using approximate qualitative states such as low, medium, and high. This representation allows uncertainty to be modeled more naturally while preserving interpretability.

Let p_{ij}^0 denote the initial pheromone intensity associated with edge (i, j). Triangular fuzzy membership functions are employed to classify pheromone intensity into three linguistic states: Low, Medium, and High. These membership functions are defined as Eq. 3.18.

$$\mu_{\text{Low}}\left(p_{ij}\right) = \begin{cases} 1 & \text{if } p_{ij} = 0 \\ \frac{p_{\text{mid}} - p_{ij}}{p_{\text{mid}} - p_{\text{min}}} & \text{if } p_{\text{min}} \leq p_{ij} \leq p_{\text{mid}} \\ 0 & \text{if } p_{ij} > p_{\text{mid}} \end{cases}$$

$$\mu_{\text{Med}}\left(p_{ij}\right) = \begin{cases} 0 & \text{if } p_{ij} < p_{\text{min}} \text{ or } p_{ij} > p_{\text{max}} \\ \frac{p_{ij} - p_{\text{min}}}{p_{\text{mid}} - p_{\text{min}}} & \text{if } p_{\text{min}} \leq p_{ij} \leq p_{\text{mid}} \\ \frac{p_{\text{max}} - p_{ij}}{p_{\text{max}} - p_{\text{mid}}} & \text{if } p_{\text{mid}} < p_{ij} \leq p_{\text{max}} \end{cases} \tag{3.18}$$

$$\mu_{\text{High}}\left(p_{ij}\right) = \begin{cases} 0 & \text{if } p_{ij} \leq p_{\text{mid}} \\ \frac{p_{ij} - p_{\text{mid}}}{p_{\text{max}} - p_{\text{mid}}} & \text{if } p_{\text{mid}} < p_{ij} \leq p_{\text{max}} \\ 1 & \text{if } p_{ij} > p_{\text{max}} \end{cases}$$

where p_{min}, p_{mid}, and p_{max} denote the lower, intermediate, and upper pheromone boundaries, respectively.

Similarly, desirability $\bar{u}_{ij}$ for edge (i, j) is represented using fuzzy categories: Undesirable (UD), Moderately Desirable (MD), and Highly Desirable (HD). These are defined through analogous triangular membership functions defined by Eq. 3.19.

$$\mu_{\text{UD}}\left(\bar{u}_{ij}\right) = \begin{cases} 1 & \text{if } \bar{u}_{ij} = \bar{u}_{\min} \\ \frac{\bar{u}_{\text{mid}} - \bar{u}_{ij}}{\bar{u}_{\text{mid}} - \bar{u}_{\min}} & \text{if } \bar{u}_{\min} \leq \bar{u}_{ij} \leq \bar{u}_{\text{mid}} \\ 0 & \text{if } \bar{u}_{ij} > \bar{u}_{\text{mid}} \end{cases}$$

$$\mu_{\text{MD}}\left(\bar{u}_{ij}\right) = \begin{cases} 0 & \text{if } \bar{u}_{ij} \leq \bar{u}_{\min} \text{ or } \bar{u}_{ij} > \bar{u}_{\max} \\ \frac{\bar{u}_{ij} - \bar{u}_{\min}}{\bar{u}_{\text{mid}} - \bar{u}_{\min}} & \text{if } \bar{u}_{\min} \leq \bar{u}_{ij} \leq \bar{u}_{\text{mid}} \\ \frac{\bar{u}_{\max} - \bar{u}_{ij}}{\bar{u}_{\max} - \bar{u}_{\text{mid}}} & \text{if } \bar{u}_{\text{mid}} < \bar{u}_{ij} \leq \bar{u}_{\max} \end{cases} \tag{3.19}$$

$$\mu_{\text{HD}}\left(\bar{u}_{ij}\right) = \begin{cases} 0 & \text{if } \bar{u}_{ij} \leq \bar{u}_{\text{mid}} \\ \frac{\bar{u}_{ij} - \bar{u}_{\text{mid}}}{\bar{u}_{\max} - \bar{u}_{\text{mid}}} & \text{if } \bar{u}_{\text{mid}} < \bar{u}_{ij} \leq \bar{u}_{\max} \\ 1 & \text{if } \bar{u}_{ij} > \bar{u}_{\max} \end{cases}$$

where $\bar{u}_{\min}$, $\bar{u}_{\text{mid}}$, and $\bar{u}_{\max}$ denote the corresponding desirability boundaries.

Importantly, these fuzzy boundaries are not fixed constants. Instead, they are dynamically recalibrated at each iteration according to the evolving distribution of pheromone and desirability values across graph edges. This adaptive parameterization allows the fuzzy framework to remain responsive to stochastic changes in path quality while preserving meaningful distinctions among linguistic categories.

After fuzzification, the fuzzy values must be transformed back into crisp numerical quantities to enable computational decision-making within the ACO process. This is achieved through centroid-based defuzzification, which computes the center of gravity of the aggregated fuzzy membership distribution.

The defuzzified pheromone intensity is given by Eq. 3.20.

$$\text{PI}_{\text{crisp}} = \frac{\int_{p_{\min}}^{p_{\max}} p_{ij} \cdot \mu(p_{ij})\, dp}{\int_{p_{\min}}^{p_{\max}} \mu(p_{ij})\, dp} \tag{3.20}$$

where: $\mu(p_{ij}) = \max\left(\mu_{\text{Low}}(p_{ij}), \mu_{\text{Med}}(p_{ij}), \mu_{\text{High}}(p_{ij})\right)$.

Similarly, the defuzzified desirability is computed as Eq. 3.21.

$$\mathscr{D}_{\text{crisp}} = \frac{\int_{\bar{u}_{\min}}^{\bar{u}_{\max}} \bar{u}_{ij} \cdot \mu(\bar{u}_{ij})\, d\bar{u}}{\int_{\bar{u}_{\min}}^{\bar{u}_{\max}} \mu(\bar{u}_{ij})\, d\bar{u}} \tag{3.21}$$

Through this process, uncertain linguistic assessments are transformed into operational nu-

merical values suitable for computational optimization. In this manner, fuzzification enables realistic representation of uncertainty, whereas defuzzification ensures practical applicability within the optimization framework. By incorporating fuzzy logic into the modeling of pheromone intensity and desirability, the FuBAKE framework effectively accommodates stochastic variability without compromising interpretability or algorithmic precision. Consequently, the proposed fuzzy representation provides a more robust and adaptive foundation for subsequent path selection and iterative optimization processes.

To improve conceptual clarity and maintain thesis-oriented continuity, the original methodological formulation has been elaborated in this subsection through detailed theoretical exposition.

3.2.3.2 Bayesian Belief Update

In addition to pheromone intensity and desirability, the proposed framework incorporates Bayesian belief as a probabilistic mechanism for continuously refining confidence in candidate paths during optimization. While pheromone intensity reflects accumulated traversal experience and desirability captures semantic relevance, Bayesian belief provides an adaptive measure of confidence by integrating prior knowledge with newly observed evidence from ant behavior. This enables the system to progressively strengthen paths that consistently demonstrate relevance while reducing confidence in less informative alternatives.

For each candidate path (k_I, k_{II}), a belief value $B^t_{k_I,k_{II}}$ is maintained at iteration t. This belief represents the probability that the path forms part of an optimal traversal within the candidate keyphrase graph. Initially, each path is assigned a prior belief based on its structural properties, and this belief is subsequently revised as the optimization process unfolds.

At every iteration, ants traverse the graph according to the probabilistic decision mechanism governed by pheromone intensity and desirability. Thus, the probability that an ant selects a particular path depends on the relative strength of these two components. Formally, the path selection probability is defined as Eq. 3.22.

$$P\left(\text{choose}(k_I, k_{II}) \mid PI, \mathscr{D}\right) = \frac{\left(PI^t_{k_I,k_{II}}\right)^{\alpha} \cdot \left(\mathscr{D}^t_{k_I,k_{II}}\right)^{\beta}}{\sum_{k \in \mathscr{N}(k_I)} \left(PI^t_{k_I,k}\right)^{\alpha} \cdot \left(\mathscr{D}^t_{k_I,k}\right)^{\beta}} \tag{3.22}$$

where $\mathscr{N}(k_I)$ denotes the set of neighboring nodes connected to k_I, and the parameters α and β regulate the influence of pheromone intensity and desirability, respectively.

Once all ants complete their traversal decisions during iteration t, the observed frequency with which a path is selected is used as empirical evidence for Bayesian updating. Let $n^t_{k_I,k_{II}}$ denote the number of ants selecting path (k_I, k_{II}), and let N represent the total number of ants. The likelihood of the path is then estimated as Eq. 3.23.

$$L^t_{k_I,k_{II}} = \frac{n^t_{k_I,k_{II}}}{N} \tag{3.23}$$

This likelihood reflects the degree of collective support a path receives from the ant population during the current iteration. Paths selected more frequently are interpreted as having stronger empirical evidence, while infrequently selected paths receive weaker support.

The posterior belief is then obtained by combining the prior belief with the newly observed likelihood using Bayes' theorem as in Eq. 3.24.

$$B^t_{k_I,k_{II}} = \frac{L^t_{k_I,k_{II}} \cdot B^{t-1}_{k_I,k_{II}}}{\sum_{k \in \mathcal{N}(k_I)} L^t_{k_I,k} \cdot B^{t-1}_{k_I,k}} \tag{3.24}$$

This update mechanism preserves normalization across all competing paths connected to a given node, ensuring that belief values remain probabilistically valid. Candidate paths that repeatedly attract ant traversal gain stronger posterior confidence, whereas paths with limited support gradually lose prominence.

The Bayesian update process serves several important purposes within the optimization framework. First, it introduces a memory mechanism by preserving historical confidence through prior belief values. Second, it incorporates real-time behavioral evidence through likelihood estimation. Third, it balances prior structural assumptions with observed traversal outcomes, allowing path confidence to evolve adaptively over time.

Through this probabilistic refinement, Bayesian belief functions as a complementary decision layer within FUBAKE. Rather than relying solely on immediate pheromone or desirability values, the framework benefits from a continuously updated confidence model that reflects both historical and current evidence. This enhances stability, improves path discrimination, and strengthens the framework's ability to converge toward semantically meaningful keyphrase structures while minimizing the influence of noisy or inconsistent traversal patterns.

3.2.3.3 Pheromone Intensity Update

Once ants complete path traversal and payoff values are computed, the pheromone intensity associated with each candidate path must be updated to reflect the quality of the discovered solutions. In Ant Colony Optimization, pheromone updating serves as the primary reinforcement mechanism through which successful paths become increasingly attractive over time. Paths that consistently lead to semantically meaningful and structurally relevant keyphrase combinations accumulate stronger pheromone trails, while less useful paths gradually lose influence through evaporation.

Within the FUBAKE framework, pheromone updating fulfills two complementary objectives. First, it reinforces high-quality candidate paths identified through the payoff mechanism. Second, it prevents premature convergence by gradually weakening paths that are no longer competitive. This balance between reinforcement and evaporation enables the framework to simultaneously exploit strong solutions and preserve exploratory flexibility.

The pheromone intensity for path (k_I, k_{II}) at iteration $t+1$ is updated as Eq. 3.25.

$$\mathrm{PI}^{t+1}_{k_I,k_{II}} = (1-\rho)\cdot \mathrm{PI}^{t}_{k_I,k_{II}} + \Delta\mathrm{PI}_{k_I,k_{II}} \tag{3.25}$$

where ρ denotes the pheromone evaporation rate, subject to $0 < \rho < 1$. The evaporation term $(1-\rho)$ reduces the influence of previously accumulated pheromone, ensuring that outdated or suboptimal paths do not dominate indefinitely. This mechanism is essential for maintaining adaptability and avoiding search stagnation.

The reinforcement component $\Delta\mathrm{PI}_{k_I,k_{II}}$ represents the additional pheromone deposited on a path based on its relative payoff. It is defined as eq. 3.26.

$$\Delta\mathrm{PI}_{k_I,k_{II}} = \frac{U(k_I,k_{II})}{\sum_{k_m,k_n} U(k_m,k_n)} \tag{3.26}$$

where $U(k_I,k_{II})$ denotes the payoff of path (k_I,k_{II}), and the denominator represents the cumulative payoff across all candidate paths. This normalization ensures that pheromone reinforcement is proportional to relative path quality rather than absolute payoff magnitude.

By linking pheromone deposition directly to payoff, the framework ensures that reinforcement reflects a balanced combination of pheromone intensity, semantic desirability, and Bayesian belief. Consequently, paths that demonstrate strong semantic relevance and structural signifi-

cance receive progressively stronger reinforcement, while weaker alternatives naturally diminish.

This update mechanism transforms pheromone intensity into an adaptive memory structure. Rather than functioning as a static score, pheromone becomes a continuously evolving indicator of collective optimization success, shaped by both historical reinforcement and present path quality. Over successive iterations, this process progressively biases the search toward increasingly relevant keyphrase structures.

3.2.3.4 Bayesian Convergence and Termination Criteria

Since the FUBAKE framework operates iteratively, it is necessary to establish clear convergence conditions that determine when optimization has stabilized sufficiently to terminate. Without well-defined stopping criteria, the search process may either terminate prematurely before identifying optimal keyphrases or continue unnecessarily after meaningful convergence has already occurred.

To ensure both computational efficiency and solution reliability, the framework employs multiple termination criteria inspired by Bayesian confidence, optimization stability, and path dominance. These criteria collectively assess convergence from complementary perspectives, ensuring that the algorithm halts only when the search process has reached a stable and meaningful state.

Maximum Iteration Criterion The first convergence condition is a predefined maximum iteration threshold $T_{\max}$, which guarantees termination even if other criteria are not met. This serves as a computational safeguard and ensures bounded execution $t \geq T_{\max}$.

In the current implementation, $T_{\max} = 100$, providing sufficient opportunity for exploration while avoiding excessive computational overhead.

Ant Consensus Criterion The second criterion evaluates convergence through collective ant agreement. If all ants converge toward a single dominant path, the posterior probability distribution effectively collapses into deterministic selection. This occurs when $\max_{k_I,k_{II}} P_{k_I,k_{II}} = 1$.

Under this condition, one path completely dominates the selection process, indicating maximal confidence and practical convergence.

Pheromone Stability Criterion The third criterion measures stabilization in pheromone intensity across iterations. If pheromone values no longer change significantly, further iterations

are unlikely to meaningfully alter the solution. Formally, convergence is declared when Eq. 3.27.

$$\Delta PI_{ij}(t) = |PI_{ij}(t) - PI_{ij}(t-1)| < \varepsilon \tag{3.27}$$

for all paths, where ε is a small threshold (set to 10^{-3}). This criterion captures equilibrium in the optimization landscape.

Payoff Dominance Criterion The fourth criterion assesses whether one candidate path has become significantly superior to all others in terms of payoff. Let $p^\star$ denote the highest-payoff path. Convergence is declared when Eq. 3.28.

$$\frac{U(p^\star)}{\sum_{p \neq p^\star} U(p)} \geq \theta \tag{3.28}$$

where θ is the dominance threshold. In this framework, $\theta = 2.0$, ensuring that the leading path demonstrates clear superiority before termination.

Together, these criteria provide a robust convergence strategy by combining computational limits, probabilistic certainty, optimization stability, and payoff dominance. In practice, the optimization process terminates as soon as any one of these criteria is satisfied. This multi-criterion design improves both efficiency and reliability, ensuring that the extracted keyphrases are derived from stable, high-confidence, and semantically optimized paths.

3.2.4 Algorithmic Summary of FUBAKE

For clarity, Algorithm 1 summarizes the complete workflow of the proposed FUBAKE methodology, beginning with candidate preparation and graph construction, followed by initialization, fuzzy Bayesian ant colony optimization, and final keyphrase extraction.

Algorithm 1 FuBAKE Keyphrase Extraction

Require: Document D, Ants m, Iterations T, Threshold θ
Ensure: Optimized Keyphrase Set $\mathscr{K}^*$

Step 1: Preprocessing
Tokenize D, perform POS tagging, extract noun phrases, and filter redundant phrases using θ.
Step 2: Co-occurrence Graph Construction
Construct graph G with keyphrases as nodes.
for each sentence in D **do**
    Connect co-occurring keyphrases.
end for
Step 3: Initialization
for each keyphrase pair (k_i, k_j) **do**
    Compute term frequency, co-occurrence, and semantic similarity.
    Initialize pheromones and Bayesian belief values.
end for
Step 4: Bayesian Gamified Fuzzy ACO Optimization
for $t = 1$ to T **do**
    for each keyphrase pair **do**
        Compute fuzzy desirability scores.
    end for
    for each ant **do**
        Select keyphrases based on transition probabilities.
    end for
    Update Bayesian beliefs and pheromones.
end for
Step 5: Keyphrase Extraction
Rank and extract the best keyphrases $\mathscr{K}^*$.

3.3 TopS-Key Methodology

This section presents TopS-Key, a multi-criteria keyword extraction framework that integrates natural language processing, feature-based scoring, dimensionality reduction, and decision-theoretic ranking within a unified structure. Unlike graph-driven optimization approaches, TopS-Key formulates keyword extraction as a structured ranking problem in which candidate keyphrases are evaluated across multiple statistical, semantic, and positional dimensions before being prioritized through an optimized decision-making process.

The overall workflow of the proposed TopS-Key framework is illustrated in Figure 3.2. The methodology consists of four primary stages: data preprocessing, feature scoring, dimensionality reduction using Principal Component Analysis (PCA), and final keyword ranking through Fuzzy TOPSIS with Shannon entropy-based weighting.

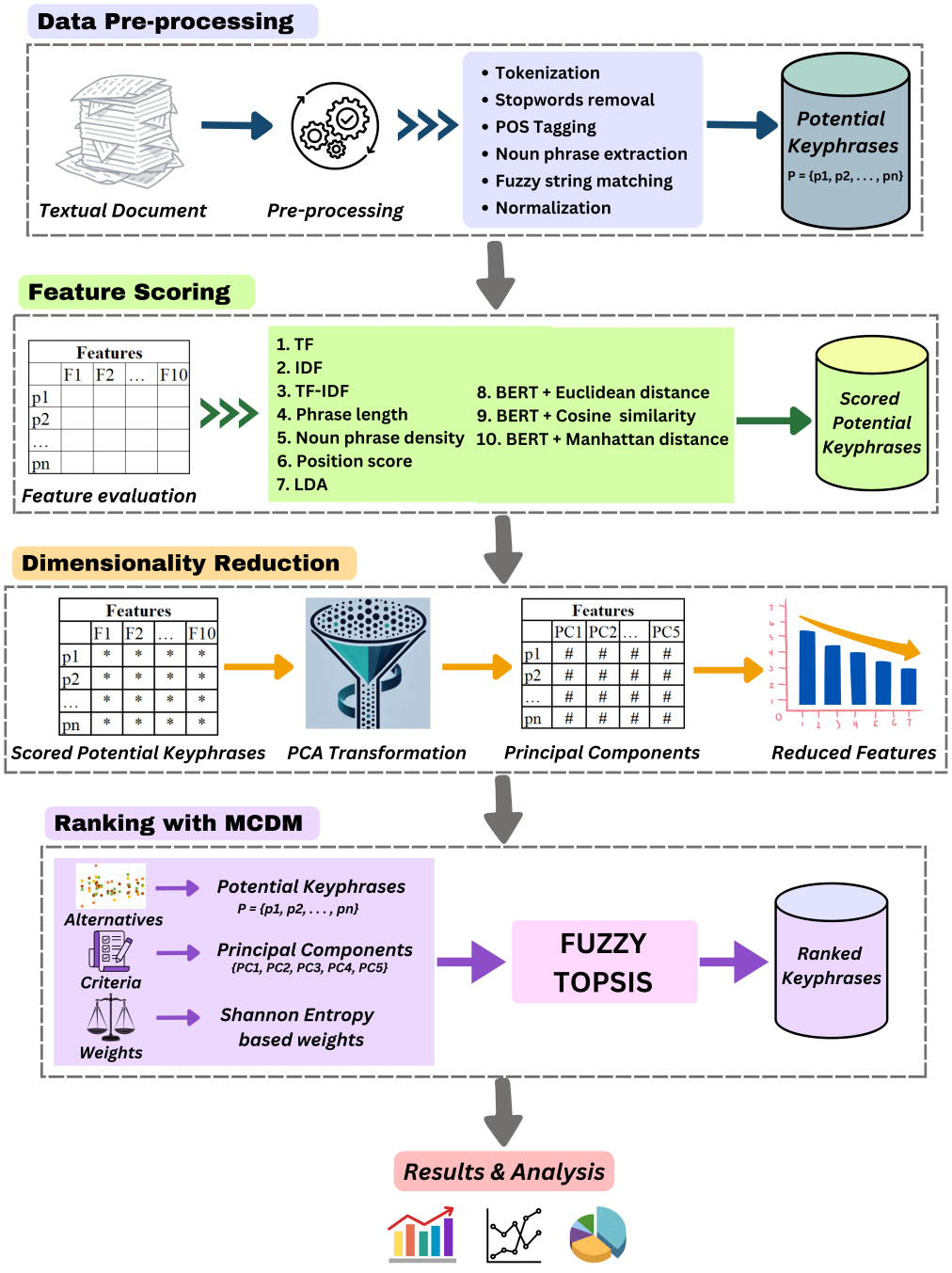

Figure 3.2: Workflow of the proposed TopS-Key framework integrating preprocessing, feature scoring, PCA-based dimensionality reduction, entropy weighting, and Fuzzy TOPSIS ranking.

The process begins with the extended preprocessing framework described earlier in this chapter, where linguistically valid and semantically refined candidate keyphrases are generated from the textual document. These candidate phrases are then evaluated using multiple feature-scoring mechanisms designed to capture diverse aspects of candidate relevance.

Since high-dimensional feature spaces may contain redundancy and correlated information, PCA is employed to reduce dimensional complexity while preserving the most informative variance within the feature set. This step improves computational efficiency and strengthens ranking stability by transforming the original feature space into a compact and representative structure.

Following dimensionality reduction, Shannon entropy is used to determine objective feature weights based on information contribution, thereby minimizing subjective bias in feature prioritization. Finally, Fuzzy TOPSIS ranks candidate keyphrases by measuring their relative closeness to the ideal solution and distance from the negative ideal solution.

Through this sequential integration of linguistic preprocessing, statistical feature engineering, dimensionality optimization, and fuzzy multi-criteria decision-making, TopS-Key provides a robust and interpretable framework for unsupervised keyword extraction. The detailed methodology of each stage is presented in the following subsections.

3.3.1 Feature Scoring of Candidate Keyphrases

Following the preprocessing and candidate refinement framework described earlier in this chapter, the TopS-Key methodology utilizes the resulting set of distinct candidate keyphrases for feature-based evaluation. At this stage, the objective is to quantify the relevance of each candidate phrase through multiple complementary indicators rather than relying on a single statistical or semantic criterion. Since keyword importance is inherently multi-dimensional, TopS-Key constructs a comprehensive feature space that captures statistical prominence, structural characteristics, positional significance, and semantic relevance.

Let the refined candidate set be denoted as $\mathscr{K} = \{k_1, k_2, \ldots, k_m\}$, where each candidate phrase k_i is evaluated using ten distinct feature scores. These features were selected to represent diverse yet complementary perspectives of keyphrase importance, ensuring balanced assessment prior to dimensionality reduction and ranking.

1. **Term Frequency (TF):** TF measures the frequency with which a candidate phrase ap-

pears within the document. Frequently occurring phrases are often indicative of central themes and therefore receive higher importance.

2. **Inverse Document Frequency (IDF):** IDF evaluates the rarity of a candidate phrase across the document corpus. Phrases that occur in fewer documents are considered more discriminative and informative evaluated using Eq. 3.29.

$$idf(i,d) = \log\left(\frac{|d|}{1+|d,i|}\right) \tag{3.29}$$

where $|d|$ represents the total number of documents and $|d,i|$ denotes the number of documents containing phrase i.

3. **TF-IDF:** TF-IDF combines local frequency and corpus-level rarity, emphasizing phrases that are prominent within the document while remaining distinctive across the corpus.

4. **Phrase Length:** Phrase length measures the number of constituent words in a candidate phrase. This feature captures the descriptive specificity of multi-word phrases.

5. **Noun Phrase Density (NPD):** NPD measures the proportional contribution of a candidate phrase relative to the overall document length evaluated using Eq. 3.30.

$$NPD(i) = \frac{N(w,i)}{N(w,d)} \tag{3.30}$$

where $N(w,i)$ is the number of words in phrase i, and $N(w,d)$ is the total number of words in the document.

6. **Position Score:** Position score captures the relative location of a candidate phrase within the document. In many formal texts, important concepts often appear earlier, making positional prominence a useful indicator.

7. **LDA-Based Semantic Score:** To capture thematic consistency, Latent Dirichlet Allocation (LDA) is employed to model latent topics and measure the semantic alignment between candidate phrases and document topics. Hellinger distance is used to compare topic distributions evaluated using Eq. 3.31.

$$H(P,P') = \frac{1}{\sqrt{2}}\sqrt{\sum_{j=1}^{k}\left(\sqrt{x_j}-\sqrt{x'_j}\right)^2} \tag{3.31}$$

The semantic score is then defined as Eq. 6.48. Higher values indicate stronger semantic relevance.

$$Semantic_score(i) = 1 - H(P, P') \tag{3.32}$$

8. **BERT-Based Cosine Similarity:** Contextualized BERT embeddings are used to measure semantic similarity between candidate phrases and document context through cosine similarity as Eq. 3.33.

$$C_{\text{BERT}} = \frac{\vec{v_i} \cdot \vec{v_d}}{\|\vec{v_i}\| \cdot \|\vec{v_d}\|} \tag{3.33}$$

where $\vec{v_i}$ and $\vec{v_d}$ denote phrase and document embeddings.

9. **BERT-Based Euclidean Distance:** Euclidean distance measures geometric separation between phrase and document embeddings as Eq. 3.34.

$$E_{\text{BERT}} = \sqrt{\sum_{j=1}^{n} (i_j - d_j)^2} \tag{3.34}$$

10. **BERT-Based Manhattan Distance:** Manhattan distance measures semantic separation through absolute component-wise differences evaluated as Eq. 3.35.

$$M_{\text{BERT}} = \sum_{j=1}^{n} |i_j - d_j| \tag{3.35}$$

Together, these ten features establish a comprehensive representation of candidate keyphrase relevance. Statistical measures capture prominence and rarity, structural features reflect linguistic composition, positional features capture document organization, and semantic features model contextual and thematic alignment. By integrating these heterogeneous dimensions, TopS-Key constructs a rich feature matrix that supports more informed ranking decisions.

However, because high-dimensional feature spaces may contain correlation and redundancy, direct ranking using all raw features may reduce efficiency and stability. Therefore, the next stage of the TopS-Key framework applies PCA to transform this feature space into a reduced and more informative representation while preserving the most significant variance.

3.3.2 Dimensionality Reduction using Principal Component Analysis

After computing multiple feature scores for each candidate keyphrase, the resulting feature space becomes highly descriptive but may also contain overlapping information, inter-feature dependency, and unnecessary complexity. Since several statistical, structural, and semantic indicators may partially capture similar patterns, directly using all features for ranking can introduce redundancy and reduce computational efficiency. To overcome this limitation, the proposed TopS-Key framework incorporates PCA as a dimensionality reduction mechanism that transforms the original feature space into a smaller set of orthogonal components while preserving the majority of informative variance.

PCA is a statistical transformation technique that converts correlated variables into a new set of linearly independent variables known as principal components. Each principal component represents a weighted combination of the original features, arranged in descending order of explained variance. Consequently, the first few components retain the most significant patterns present in the data, enabling dimensionality reduction without substantial information loss.

Let X denote the feature matrix, where each row corresponds to a candidate keyphrase and each column corresponds to one of the extracted feature scores. Because these features may exist on different numerical scales, standardization is first required to ensure uniform contribution across all dimensions. For this purpose, z-score normalization is applied as Eq. 3.36.

$$z = \frac{x - \mu}{\sigma} \tag{3.36}$$

where x represents the original feature value, μ is the mean of the feature, and σ is its standard deviation. This transformation normalizes each feature to zero mean and unit variance, thereby preventing scale-sensitive dominance.

Following standardization, the covariance matrix is computed to capture the relationships among features Eq. 3.37.

$$[C] = \frac{1}{n-1} X^T X \tag{3.37}$$

where X denotes the standardized feature matrix and $[C]$ represents the covariance matrix. This matrix provides a quantitative representation of how feature dimensions vary jointly and forms the mathematical basis for identifying redundant relationships.

The covariance matrix is then decomposed into eigenvalues and eigenvectors. Eigenvectors define the directions of the new transformed axes, while eigenvalues measure the amount of variance captured along those directions. Principal components associated with larger eigenvalues preserve more information and are therefore considered more significant.

To quantify the contribution of each component, the explained variance ratio is calculated as Eq. 3.38.

$$EVR_j = \frac{\lambda_j}{\sum_{m=1}^{k} \lambda_m} \tag{3.38}$$

where λ_j denotes the eigenvalue of the jth principal component, and the denominator represents total variance across all components.

Since the primary objective of PCA in TopS-Key is to preserve informative structure while reducing dimensionality, cumulative explained variance is used to determine the number of components to retain Eq. 3.39.

$$Cum_{EVR} = \sum_{m=1}^{k} EVR_m \tag{3.39}$$

In this framework, principal components are retained until cumulative explained variance reaches 90%. This threshold provides a practical balance between information preservation and dimensional simplification, ensuring that most of the meaningful variance is maintained while less informative components are excluded.

Once the optimal number of components is identified, the standardized feature matrix is projected into the reduced component space $X_{PC} = XV$.

where V represents the matrix of selected eigenvectors, and X_{PC} denotes the transformed principal component representation.

The integration of PCA within TopS-Key offers several methodological advantages. By converting correlated features into orthogonal components, PCA minimizes redundancy and reduces overlapping signals among feature dimensions. This not only improves computational efficiency but also enhances ranking stability by ensuring that subsequent decision-making is based on compact, information-rich representations rather than raw correlated inputs.

Although PCA is fundamentally a linear transformation, its application in TopS-Key remains highly effective because it is preceded by a carefully engineered feature space that already

captures multiple semantic, structural, and statistical perspectives. Moreover, the subsequent stages of Shannon entropy weighting and Fuzzy TOPSIS ranking introduce adaptive and non-linear decision mechanisms that further strengthen robustness. Thus, within the proposed architecture, PCA serves as a critical intermediate step that improves efficiency, preserves essential feature information, and establishes a stable foundation for optimized multi-criteria ranking.

3.3.3 Applying Fuzzy TOPSIS with Shannon Entropy Weights

Following dimensionality reduction through PCA, the retained principal components are treated as the final decision criteria for ranking candidate keyphrases. At this stage, keyword extraction is formulated as a Multi-Criteria Decision-Making (MCDM) problem, where each candidate keyphrase is considered an alternative and evaluated across multiple optimized criteria.

To perform this ranking, the proposed TopS-Key framework employs Fuzzy TOPSIS, an extension of the conventional TOPSIS method that integrates fuzzy logic to handle uncertainty and imprecision in feature evaluation. This approach ranks candidate keyphrases according to their closeness to the ideal solution and distance from the negative ideal solution, thereby enabling more reliable prioritization under heterogeneous feature conditions.

In the proposed framework, each candidate keyphrase is represented as an alternative, while the PCA-derived principal components serve as ranking criteria. Since criterion importance may vary, objective weighting is required before ranking. Instead of manually assigning weights, Shannon entropy is used to determine criterion importance based on the informational variability of each component. This ensures a data-driven and unbiased weighting mechanism.

The integration of PCA, Shannon entropy, and Fuzzy TOPSIS strengthens the framework in three ways: PCA reduces redundancy, entropy weighting determines criterion significance objectively, and fuzzy TOPSIS manages uncertainty during final ranking. Together, these stages provide a robust and optimized decision-making structure for identifying the most relevant keyphrases.

This stage is therefore divided into two components: (i) Shannon entropy-based criterion weighting, and (ii) Fuzzy TOPSIS-based candidate ranking.

3.3.3.1 Shannon Entropy-Based Criterion Weighting

Once the principal components are obtained through PCA, the next step is to determine their relative importance within the ranking framework. Since not all components contribute equally to candidate discrimination, an objective weighting mechanism is required to quantify the informational significance of each criterion. For this purpose, TopS-Key employs Shannon entropy, a well-established concept from information theory that measures uncertainty or disorder within a distribution.

The primary advantage of Shannon entropy lies in its ability to evaluate how much useful discriminative information a criterion provides across all candidate keyphrases. A criterion whose values vary significantly across alternatives contains greater distinguishing power and is therefore considered more informative. In contrast, a criterion with highly uniform values contributes less to decision-making because it offers limited differentiation. By using entropy, the weighting process becomes entirely data-driven, eliminating subjective bias and ensuring that criterion importance reflects actual statistical variability.

Let x_{ip} denote the value of the pth principal component for candidate keyphrase i, where m represents the total number of candidate keyphrases. The first step is normalization, which converts each criterion into a proportional distribution defined as Eq. 3.40.

$$N_{ip} = \frac{x_{ip}}{\sum_{i=1}^{m} x_{ip}} \tag{3.40}$$

This normalization ensures comparability by scaling the values of each principal component across all alternatives.

Using the normalized distribution, Shannon entropy for each principal component p is computed as Eq. 3.41.

$$H_p = -\frac{1}{\log(m)} \sum_{i=1}^{m} N_{ip} \log(N_{ip} + \varepsilon) \tag{3.41}$$

where ε is a small constant introduced to maintain numerical stability and avoid undefined logarithmic values. The normalization factor $\frac{1}{\log(m)}$ ensures that entropy values remain within the range $[0, 1]$.

A higher entropy value indicates that the criterion values are more evenly distributed across candidate phrases, suggesting lower discriminative capability. Conversely, lower entropy

implies greater variability and stronger informative potential. To quantify this informative strength, the degree of diversification is calculated as $D_p = 1 - H_p$.

Here, D_p represents the effective information contribution of the pth criterion. Criteria with larger diversification values are considered more useful because they provide stronger differentiation among alternatives.

Finally, the normalized weight assigned to each principal component is determined by Eq. 3.42.

$$w_p = \frac{D_p}{\sum_{p=1}^{n} D_p} \tag{3.42}$$

where n denotes the total number of retained principal components.

Through this process, each criterion receives a weight proportional to its informational contribution. Principal components with greater variability and stronger discriminative capacity are assigned higher weights, while less informative components receive lower influence. This entropy-based weighting strategy ensures that TopS-Key prioritizes criteria according to objective statistical significance rather than manual assumptions, thereby improving ranking reliability, adaptability, and overall decision quality.

3.3.3.2 Applying Fuzzy TOPSIS for Candidate Ranking

After computing Shannon entropy-based weights for the retained principal components, the final stage of TopS-Key involves ranking candidate keyphrases through the Fuzzy TOPSIS framework. At this stage, each candidate keyphrase is treated as an alternative, while each weighted principal component serves as a decision criterion. The objective is to identify candidate phrases that are simultaneously closest to the ideal solution and farthest from the least desirable solution. The following methodology is adapted and reformulated from the original TopS-Key framework into thesis oriented form.

The process begins with the construction of a decision matrix $D = [d_{ip}]$, where d_{ip} represents the value of candidate keyphrase i with respect to criterion p. Each criterion value is then multiplied by its corresponding Shannon entropy weight w_p to construct the weighted decision matrix as Eq. 3.43.

$$\tilde{d}_{ip} = w_p \cdot d_{ip} \tag{3.43}$$

Unlike conventional TOPSIS, which operates on crisp numerical values, TopS-Key introduces fuzzy logic to better model uncertainty and ambiguity. Each weighted criterion value is therefore transformed into a triangular fuzzy number as Eq. 3.44.

$$\tilde{d}_{ip} = (l_{ip}, m_{ip}, u_{ip}) \tag{3.44}$$

where l_{ip}, m_{ip}, and u_{ip} represent the lower, modal, and upper bounds, respectively. These values are computed as Eq. 3.45.

$$l_{ip} = \mu_{ip} - k \cdot sd_p, \qquad m_{ip} = \mu_{ip}, \qquad u_{ip} = \mu_{ip} + k \cdot sd_p \tag{3.45}$$

where μ_{ip} is the criterion mean, sd_p is the standard deviation of criterion p, and k controls the degree of fuzziness. This triangular representation enables the framework to capture uncertainty in feature interpretation more effectively than rigid scalar values.

Once the fuzzy decision matrix is constructed, the Fuzzy Positive Ideal Solution (FPIS) and Fuzzy Negative Ideal Solution (FNIS) are identified. FPIS represents the most desirable criterion values, while FNIS represents the least desirable as in Eq. 3.46.

$$F_p^+ = \left(\max_i u_{ip}, \max_i m_{ip}, \max_i l_{ip} \right), \qquad F_p^- = \left(\min_i l_{ip}, \min_i m_{ip}, \min_i u_{ip} \right) \tag{3.46}$$

These ideal solutions serve as benchmarks against which all candidate keyphrases are evaluated.

To ensure comparability across criteria with different scales, fuzzy normalization is performed as Eq. 3.47.

$$\tilde{r}_{ip} = \left(\frac{l_{ip}}{u_p^{\max}}, \frac{m_{ip}}{m_p^{\max}}, \frac{u_{ip}}{l_p^{\max}} \right) \tag{3.47}$$

where $u_p^{\max}$, $m_p^{\max}$, and $l_p^{\max}$ are derived from the FPIS of criterion p.

The next step involves computing the distance of each alternative from both FPIS and FNIS using Euclidean distance for triangular fuzzy numbers as in Eq. 3.48.

$$d(\tilde{d}, \tilde{d}') = \sqrt{\frac{1}{3}\left[(l - l')^2 + (m - m')^2 + (u - u')^2\right]} \tag{3.48}$$

Using this formulation, the total distances of each candidate keyphrase from the positive and negative ideal solutions are calculated as Eq. 3.49.

$$D_i^+ = \sum_{p=1}^{n} d(\tilde{r}_{ip}, \tilde{F}_p^+), \qquad D_i^- = \sum_{p=1}^{n} d(\tilde{r}_{ip}, \tilde{F}_p^-) \tag{3.49}$$

where D_i^+ denotes the distance from the ideal solution and D_i^- denotes the distance from the negative ideal solution.

Finally, the closeness coefficient for each candidate keyphrase is computed as Eq. 3.50.

$$C_i = \frac{D_i^-}{D_i^+ + D_i^-} \tag{3.50}$$

where $C_i \in [0,1]$. A larger value of C_i indicates that the candidate phrase is closer to the ideal solution and therefore more suitable as a final keyphrase.

Candidate keyphrases are subsequently ranked in descending order of their closeness coefficients, and the highest-ranked phrases are selected as the predicted keyphrases.

The integration of Fuzzy TOPSIS within TopS-Key offers several methodological advantages. First, fuzzy representation enables uncertainty aware decision-making by accommodating ambiguity in heterogeneous feature evaluation. Second, the simultaneous consideration of both ideal and negative solutions ensures balanced ranking rather than isolated score maximization. Third, entropy-based weighting guarantees objective criterion prioritization, eliminating subjective bias.

From a computational perspective, although the framework integrates BERT-based semantic modeling, PCA, entropy weighting, and fuzzy ranking, its downstream ranking components remain computationally efficient. After contextual embedding, the principal TopS-Key operations scale linearly with respect to the number of candidate keyphrases, ensuring practical scalability for large textual collections.

Thus, through the integration of PCA-based dimensionality reduction, Shannon entropy weighting, and Fuzzy TOPSIS ranking, TopS-Key provides a comprehensive multi-criteria decision framework capable of producing robust, interpretable, and semantically informed keyphrase rankings.

3.3.4 Algorithmic Summary of TopS-Key

For clarity, Algorithm 2 summarizes the complete workflow of the proposed TopS-Key methodology, beginning with candidate keyphrase generation and feature scoring, followed by PCA-based dimensionality reduction, Shannon entropy weighting, Fuzzy TOPSIS ranking, and final keyphrase extraction.

Algorithm 2 TopS-Key Keyphrase Extraction

Require: Document D, Feature Set F, PCA Variance Threshold τ, Fuzziness Parameter k
Ensure: Ranked Keyphrase Set $\mathscr{K}^*$

Step 1: Candidate Keyphrase Extraction
Apply preprocessing and extended preprocessing on D.
Extract refined candidate keyphrase set $\mathscr{K}$.
Step 2: Feature Score Computation
for each candidate keyphrase $k_i \in \mathscr{K}$ **do**
    Compute statistical, structural, positional, and semantic feature scores.
end for
Construct feature matrix **F**.
Step 3: Principal Component Analysis
Standardize feature matrix **F**.
Compute covariance matrix, eigenvalues, and eigenvectors.
Retain principal components satisfying cumulative explained variance threshold τ.
Transform **F** into reduced principal component matrix **PC**.
Step 4: Shannon Entropy-Based Weighting
for each principal component p **do**
    Normalize criterion values.
    Compute entropy score H_p.
    Compute diversification degree D_p.
    Derive criterion weight w_p.
end for
Step 5: Fuzzy TOPSIS Ranking
Construct decision matrix using weighted principal components.
Convert weighted values into triangular fuzzy numbers.
Identify FPIS and FNIS.
for each candidate keyphrase **do**
    Compute distance from FPIS and FNIS.
    Calculate closeness coefficient C_i.
end for
Step 6: Final Keyphrase Extraction
Rank candidate keyphrases in descending order of C_i.
Select top-ranked keyphrases as final output $\mathscr{K}^*$.

3.4 ForestRank Methodology

This section presents ForestRank, a structured keyphrase extraction framework that integrates feature engineering, ensemble-based feature selection, and multi-criteria decision-making as an unified architecture. The methodology is organized into four sequential stages. The process begins with document preprocessing, where textual data is transformed into refined candidate noun phrases using the preprocessing framework established earlier in this chapter. The second stage involves computing multiple feature scores for each candidate noun phrase to capture statistical, structural, and semantic relevance. The third stage applies the Random Forest classifier to identify the most significant and discriminative features from the constructed feature space. Finally, the fourth stage employs the Weighted Aggregated Sum Product Assessment (WASPAS) method to rank candidate noun phrases and extract the final keyphrases using the selected optimal features.

The overall workflow of the proposed ForestRank framework is illustrated in Figure 3.3. By combining preprocessing, feature scoring, feature significance learning, and optimized ranking, ForestRank provides a systematic approach for robust and interpretable keyphrase extraction.

Each stage of the methodology is described in detail in the following subsections.

3.4.1 Document Preparation and Candidate Representation

The ForestRank framework begins with the preprocessing and candidate refinement strategy established earlier in this chapter. Since the foundational preprocessing pipeline including tokenization, normalization, linguistic filtering, noun phrase extraction, and fuzzy similarity-based deduplication has already been discussed in detail, it is adopted here as the initial document preparation stage without methodological repetition. Through this process, each document is transformed into a refined set of representative noun phrases that serve as candidate keyphrases for subsequent analysis.

A distinguishing characteristic of ForestRank is the incorporation of supervised candidate labeling during this stage. In addition to candidate generation, extracted noun phrases are compared with author assigned or reference keyphrases using similarity-based matching to assign binary labels indicating whether a candidate phrase corresponds to an actual keyphrase. This labeling process establishes a structured learning environment in which candidate phrases

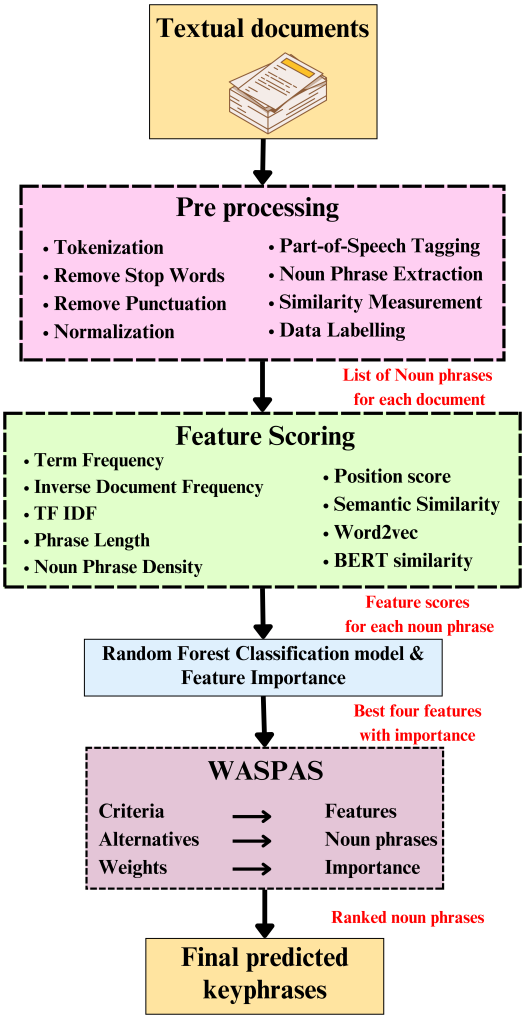

Figure 3.3: Workflow of the proposed ForestRank framework integrating preprocessing, feature scoring, Random Forest-based feature selection, and WASPAS-based keyphrase ranking.

are categorized as relevant or non-relevant, thereby enabling downstream feature significance analysis through Random Forest classification. The methodological structure is adapted from the original ForestRank framework and reformulated here in thesis style for clarity.

Formally, for a given document D, the preprocessing stage produces a candidate set $\mathcal{N} = \{n_1, n_2, \ldots, n_k\}$, where each n_i represents a refined noun phrase extracted from the document. Each candidate is then associated with a binary label $y_i \in \{0, 1\}$.

where $y_i = 1$ indicates correspondence with a valid keyphrase and $y_i = 0$ indicates a non-keyphrase candidate.

This structured representation serves two important purposes. First, it preserves the linguistic and semantic integrity of candidate phrases through standardized preprocessing. Second, it provides the labeled foundation necessary for supervised feature importance learning in sub-

sequent stages.

Thus, rather than functioning solely as a text-cleaning procedure, preprocessing in ForestRank acts as a bridge between linguistic candidate extraction and machine learning-driven feature optimization, ensuring that the candidate pool is both semantically refined and analytically structured for downstream ranking.

3.4.2 Feature Significance Analysis using Random Forest Classification

Following candidate preparation, each noun phrase is represented through a multi-dimensional feature space designed to capture diverse aspects of phrase relevance. Although multiple statistical, semantic, and structural features collectively improve descriptive richness, not all features contribute equally to keyphrase discrimination. Some may provide highly informative signals, while others may introduce redundancy or minimal predictive value. Therefore, before final ranking, ForestRank incorporates a feature significance analysis stage to identify the most discriminative features and reduce the dimensional space to its most informative subset.

To achieve this, the framework employs Random Forest classification, an ensemble-based supervised learning method that combines multiple decision trees to improve predictive robustness, reduce overfitting, and estimate feature importance systematically. Random Forest is particularly suitable in this context because it can handle heterogeneous feature types, remains resilient to noise, and provides an intrinsic mechanism for measuring the contribution of each feature toward classification accuracy. This methodology is adapted from the original ForestRank framework and reformulated here in thesis style for conceptual clarity.

Initially, each candidate noun phrase is associated with a feature matrix $\mathbf{F} \in \mathbb{R}^{n \times m}$, where n denotes the number of candidate noun phrases and m represents the number of extracted features. Since low-variance features often contribute little to discrimination, an initial variance threshold is applied to eliminate features with minimal informational spread evaluated using Eq. 3.51.

$$Var_X = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 \tag{3.51}$$

where $\bar{x}$ is the mean feature value. Features with negligible variance are excluded because they provide limited differentiation among candidate phrases.

Following variance filtering, feature standardization is performed to normalize all retained

features onto a common scale evaluated using Eq. 3.52.

$$z = \frac{x - \mu}{\sigma} \tag{3.52}$$

where x is the feature value, μ is the feature mean, and σ is the standard deviation. This step ensures that features with larger numerical scales do not disproportionately influence classification.

The standardized feature matrix is then paired with the binary label vector generated during preprocessing, where each noun phrase is classified as either a valid keyphrase or non-keyphrase. Using this labeled dataset, Random Forest is trained through multiple decision trees constructed from randomized subsets of samples and features. This ensemble strategy improves generalization by aggregating multiple independent learning paths.

Within each decision tree, node splitting is guided by impurity reduction, commonly measured using Gini impurity as in Eq. 3.53.

$$Gini = 1 - \sum_{i=1}^{C} p_i^2 \tag{3.53}$$

where p_i represents the probability of class membership and C is the total number of classes.

For a split based on feature f, the reduction in impurity is defined as Eq. 3.54.

$$\Delta Gini = G_{parent} - \left(\frac{N_{left}}{N_{total}} G_{left} + \frac{N_{right}}{N_{total}} G_{right} \right) \tag{3.54}$$

This quantity measures how effectively a feature improves classification purity at each split.

The overall importance of feature f is then computed by averaging its impurity reductions across all trees evaluated as Eq. 3.55.

$$F_{importance}(f) = \frac{1}{T} \sum_{t=1}^{T} \sum_{s \in S_t} \Delta Gini(s, f) \tag{3.55}$$

where T is the total number of trees, and S_t denotes the set of nodes in tree t where feature f is used.

Features with higher Gini importance values contribute more substantially to classification performance and are therefore considered more significant for keyphrase extraction. In the proposed framework, only features exceeding the predefined importance threshold are retained

for subsequent ranking.

This Random Forest-based feature significance stage serves multiple purposes. First, it identifies the most predictive features through supervised learning. Second, it reduces dimensional complexity by removing weak or redundant features. Third, it improves the reliability of downstream ranking by ensuring that only the strongest discriminative signals are forwarded to the final decision-making stage.

Thus, within ForestRank, Random Forest functions not merely as a classifier, but as an intelligent feature optimization mechanism that bridges feature engineering and final keyphrase ranking.

3.4.3 MCDM Technique - WASPAS

MCDM is a set of methods or procedures used to evaluate and prioritize multiple competing criteria when making decisions. In the context of keyphrase extraction, MCDM can be particularly useful as it involves multiple features and it provides a systematic approach to combine these diverse criteria and determine the overall importance of each candidate keyphrase. Thus, in the final step, we rank our noun phrases using the MCDM technique, WASPAS (Weighted Aggregated Sum Product Assessment). Our alternatives that need to be ranked are the noun phrases, the criteria are the best features we got in step 3, and the weights for each feature are the feature importance in step 3.

We start ranking using WASPAS by normalization to ensure comparability between different features calculated. Then we calculate the weighted sum and product using Eq. 3.56 and 3.57 respectively.

$$\text{Weighted Sum } = \sum_{i=1}^{n} w_i \cdot x^i_{\text{normalized}} \tag{3.56}$$

$$\text{Weighted Product } = \prod_{i=1}^{n} \left(x^i_{\text{normalized}}\right)^{w_i} \tag{3.57}$$

Where w_i is the weight of the ith feature. Then the WASPAS score is calculated by adding Weighted Sum and Weighted Product scores. By combining both the additive and multiplicative effects of the features, the WASPAS score provides a comprehensive ranking that considers the relative importance and interaction of each feature. This method helps identify the

most significant keyphrases for each document.

Thus, we get the ranked noun phrases, and the top-ranked are the predicted keyphrases for the document.

3.4.4 Algorithmic Summary of ForestRank

For clarity, Algorithm 3 summarizes the complete workflow of the proposed ForestRank methodology, beginning with candidate generation and supervised labeling, followed by feature computation, Random Forest-based feature selection, WASPAS-based ranking, and final keyphrase extraction.

Algorithm 3 ForestRank Keyphrase Extraction

Require: Document D, Reference Keyphrases K, Feature Set F, Variance Threshold θ, Importance Threshold δ
Ensure: Ranked Keyphrase Set $\mathscr{K}^*$

Step 1: Candidate Keyphrase Extraction and Labeling
Apply preprocessing on D.
Extract refined noun phrase candidate set $\mathscr{N}$.
for each candidate $n_i \in \mathscr{N}$ **do**
    Assign label $y_i \leftarrow 1$ if n_i matches K, else 0.
end for
Step 2: Feature Score Computation
for each candidate $n_i \in \mathscr{N}$ **do**
    Compute statistical, structural, and semantic feature scores.
end for
Construct feature matrix **F**.
Step 3: Feature Filtering and Normalization
Remove features with variance below threshold θ.
Standardize remaining features.
Step 4: Random Forest-Based Feature Selection
Train Random Forest classifier using feature matrix and labels.
Compute feature importance scores.
Select optimal feature subset F_{opt} based on importance threshold δ.
Step 5: WASPAS-Based Ranking
Normalize selected feature values.
for each candidate n_i **do**
    Compute weighted sum score.
    Compute weighted product score.
    Combine scores to obtain final ranking score.
end for
Step 6: Final Keyphrase Extraction
Rank candidates in descending order of final scores.
Select top-ranked candidates as final keyphrases $\mathscr{K}^*$.

3.5 Experimental Evaluation and Comparative Analysis

This section evaluates the performance of the three proposed keyphrase extraction frameworks presented in this chapter: FuBAKE, TopS-Key, and ForestRank. While the previous sections detailed their methodological foundations, the present section focuses on their empirical validation under a unified experimental framework.

The evaluation begins with a description of the benchmark datasets and performance metrics used for assessment. This is followed by individual threshold and parameter analyses for each model to identify their optimal configurations. Once each framework is evaluated under its best settings, a consolidated comparative analysis is conducted to examine their relative effectiveness across datasets.

To further understand model behavior, ablation analysis is performed to assess the contribution of key architectural components, followed by sensitivity analysis to evaluate robustness under varying parameter conditions.

Together, these experiments provide a comprehensive assessment of the effectiveness, stability, and comparative strengths of the proposed frameworks.

3.5.1 Datasets

To evaluate the effectiveness and generalizability of the proposed keyphrase extraction frameworks, experiments were conducted on three publicly available benchmark datasets: Inspec [12], SemEval2017 Task 10 [127], and DUC2001 [128]. These datasets were selected because they collectively represent diverse textual environments with substantial variation in domain, document length, and structural complexity, thereby enabling rigorous assessment across heterogeneous document types.

A practical limitation of benchmark-based keyphrase extraction is that some gold-standard keyphrases may not explicitly appear in the source document, while others may not align with extractable linguistic structures. Consequently, complete keyphrase recovery is theoretically unattainable, making these datasets suitable for realistic performance evaluation rather than absolute extraction.

The selected datasets span multiple application contexts, ranging from short technical abstracts to medium-length scientific paragraphs and long-form news articles. This diversity ensures

that the proposed frameworks are evaluated not only for extraction effectiveness but also for scalability, adaptability, and robustness under varying textual conditions.

Table 3.1 summarizes the primary characteristics of the benchmark datasets used in this study.

Table 3.1: Summary of benchmark datasets used for experimental evaluation

Attribute	**Inspec**	**SemEval2017 Task 10**	**DUC2001**
No. of Documents	2000	493	308
Primary Domain	Computer Science, Control, and Information Technology	Computer Science, Material Science, and Physics	News / Multi-topic
Document Length	Short abstracts	Medium-length scientific paragraphs	Long articles (approximately 740 words/article)
Publication Source	Journal papers published between 1998-2002	ScienceDirect open-access publications	TREC-9 news corpus
Primary Evaluation Focus	Short technical document keyphrase extraction	Scientific paragraph keyphrase extraction	Long-form document keyphrase extraction

Collectively, these datasets establish a balanced experimental foundation for evaluating the proposed models across short, medium, and long textual settings.

3.5.2 Evaluation Metrics

The proposed keyphrase extraction frameworks are evaluated using precision, recall, and F1-score. Unlike conventional approaches that rely on exact lexical matching, this study adopts semantic matching to provide a more realistic assessment of extraction quality. Exact matching may underestimate performance when predicted keyphrases differ slightly from gold-standard annotations in wording or token arrangement despite conveying equivalent meaning.

To address this limitation, predicted and actual keyphrases are compared using fuzzy string matching through the Token Set Ratio (TSR), which measures semantic similarity while accommodating lexical variation. For a predicted keyphrase k_p and an actual keyphrase k_g, it is evaluated using Eq. 3.58.

$$Similarity(k_p, k_g) = TSR(k_p, k_g) \tag{3.58}$$

A predicted keyphrase is considered a valid match if its similarity exceeds a predefined threshold θ as in Eq. 3.59.

$$Match(k_p, k_g) = \begin{cases} 1, & \text{if } TSR(k_p, k_g) \geq \theta \\ 0, & \text{otherwise} \end{cases} \tag{3.59}$$

The threshold θ controls matching strictness and plays an important role in evaluation sensitivity. Higher values impose stricter semantic equivalence, whereas lower values allow more flexible matching. Since threshold selection directly affects performance, it is analyzed separately in the threshold sensitivity analysis of each framework.

Matching is performed using a one-to-one strategy to avoid duplicate assignments. Based on this process, True Positives (TP), False Positives (FP), and False Negatives (FN) are identified and used to compute precision, recall, and F1-score. The mathematical formulations of these evaluation metrics are presented in Table 3.2.

Table 3.2: Formulas for precision, recall, and F1-score

Metric	**Precision**	**Recall**	**F1-score**
Formula	$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

Performance is further evaluated at multiple extraction levels (Top-3, Top-5, and Top-10) to account for variation in prediction depth. This semantic matching-based framework therefore provides a more robust and practically meaningful evaluation of keyphrase extraction effectiveness than strict exact-match approaches.

3.5.3 Similarity Threshold Analysis

Since evaluation in this study is based on semantic matching rather than exact lexical correspondence, the similarity threshold θ is a critical parameter that determines whether a predicted keyphrase sufficiently aligns with a gold-standard keyphrase. Threshold selection directly influences evaluation strictness and therefore affects precision, recall, and F1-score.

Lower threshold values allow more flexible semantic matching, thereby increasing retrieval breadth but potentially introducing weaker matches. In contrast, higher threshold values impose stricter semantic correspondence, improving precision-oriented evaluation while reducing retrieval flexibility. Consequently, threshold calibration is necessary to ensure balanced

and consistent assessment of model performance.

To systematically examine this effect, similarity threshold analysis is performed individually for each proposed framework.

3.5.3.1 FuBAKE Similarity Threshold Analysis

Within FuBAKE, the similarity threshold θ governs the strictness of fuzzy phrase matching during candidate refinement. Because this threshold controls how aggressively redundant or near-duplicate phrases are filtered, it directly affects the balance between semantic diversity and noise reduction.

To evaluate this behavior, threshold values ranging from 50 to 100 were systematically analyzed across all benchmark datasets using precision, recall, and F1-score, as illustrated in Figure 3.4.

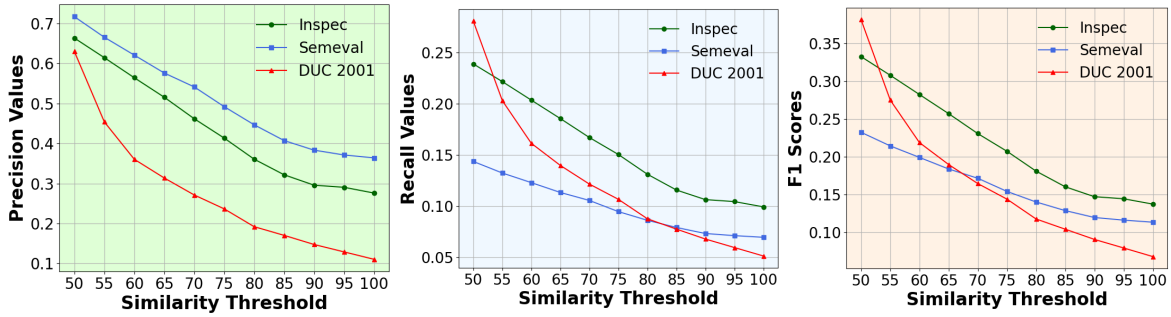

Figure 3.4: Precision, Recall, and F1-score variation across different similarity thresholds for Inspec, SemEval, and DUC datasets.

Across all datasets, lower threshold values permitted broader semantic flexibility, generally improving recall while increasing redundancy. As threshold values increased, phrase matching became progressively stricter, reducing duplicate inclusion but also limiting semantic coverage. This trade-off was especially evident in structurally complex datasets, where excessive strictness reduced contextual retention.

Among the evaluated settings, $\theta = 90$ provided the most effective balance between redundancy reduction and semantic preservation. This threshold improved candidate refinement quality while maintaining sufficient phrase diversity for downstream graph construction and Bayesian fuzzy ant colony optimization.

Accordingly, $\theta = 90$ was selected as the optimal similarity threshold for FuBAKE and was retained throughout subsequent experiments.

3.5.3.2 TopS-Key Similarity Threshold Analysis

For TopS-Key, similarity threshold analysis was conducted to examine how semantic matching strictness influences evaluation performance. Since TopS-Key also relies on semantic correspondence rather than exact lexical overlap, the threshold parameter θ determines the minimum similarity required for predicted keyphrases to be considered valid during evaluation.

Threshold values ranging from 0.70 to 0.90 were systematically analyzed at Top@3 using precision, recall, and F1-score across all benchmark datasets, as illustrated in Figures 3.5, 3.6, and 3.7.

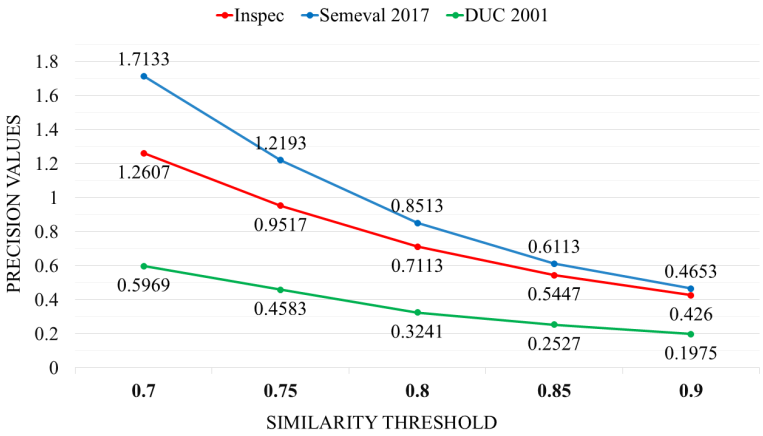

Figure 3.5: Precision trends across varying similarity thresholds for all datasets at Top@3

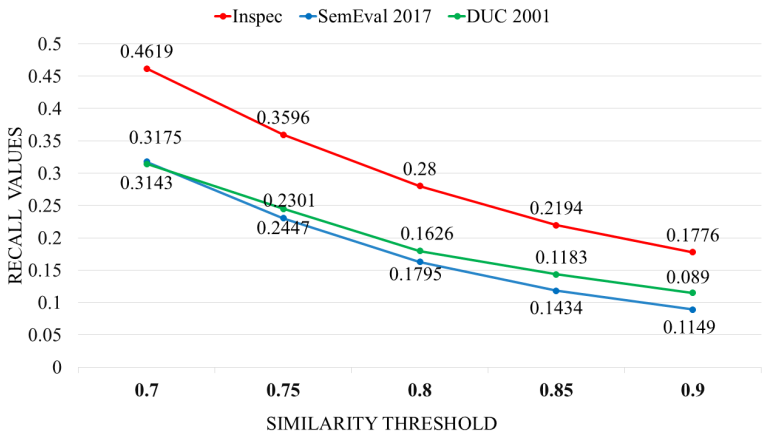

Figure 3.6: Recall trends across varying similarity thresholds for all datasets at Top@3

A consistent decline in precision, recall, and F1-score was observed as the threshold increased. Stricter thresholds reduced semantic flexibility by requiring stronger correspondence between predicted and gold-standard phrases, thereby lowering retrieval breadth. This effect became more pronounced for datasets containing longer and lexically diverse documents, where broader semantic variation was more common.

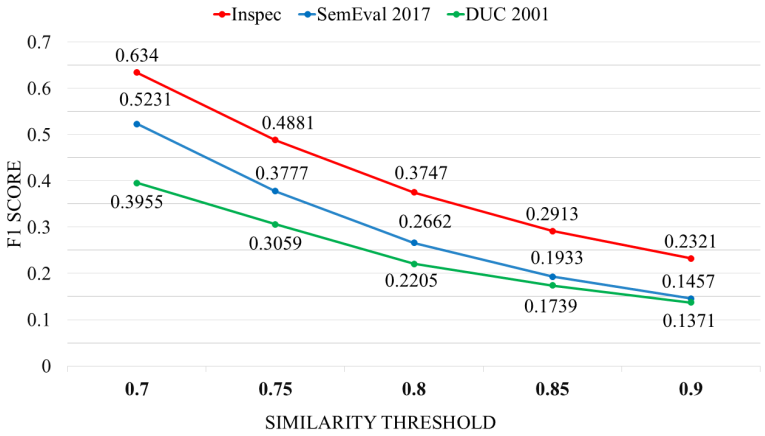

Figure 3.7: F1-score trends across varying similarity thresholds for all datasets at Top@3

Moderate threshold values generally provided stronger balance between semantic flexibility and retrieval effectiveness. However, despite favorable F1 performance at comparatively lower thresholds, the final TopS-Key configuration adopted $\theta = 0.90$. This decision prioritized stricter semantic precision, reduced false positives, and ensured stronger semantic validity of extracted keyphrases.

Although this stricter configuration introduced moderate reductions in recall and overall F1-score, TopS-Key demonstrated stable performance across the tested threshold range. Therefore, $\theta = 0.90$ was retained as the final similarity threshold for all subsequent TopS-Key experiments.

3.5.4 Parameter Sensitivity Analysis

Beyond similarity threshold calibration, the effectiveness of FuBAKE is also influenced by its internal optimization parameters. Since the framework integrates pheromone reinforcement, semantic desirability, Bayesian belief, and iterative path exploration, parameter sensitivity analysis is necessary to determine how these components collectively affect retrieval stability and performance. This analysis identifies empirically robust parameter settings while also providing deeper insight into the optimization dynamics of the framework.

3.5.4.1 FuBAKE Weighting Parameter Analysis (α, β, and γ)

In FuBAKE, the parameters α, β, and γ regulate the relative importance of pheromone intensity, desirability, and Bayesian belief during probabilistic path traversal. These parameters shape the decision-making process by controlling the balance between historical reinforce-

ment, local semantic evidence, and prior structural confidence.

To evaluate their influence, each parameter was varied independently over the interval $[0,3]$, while the remaining parameters were held constant. The resulting F1-score trends across all benchmark datasets are shown in Figure 3.8.

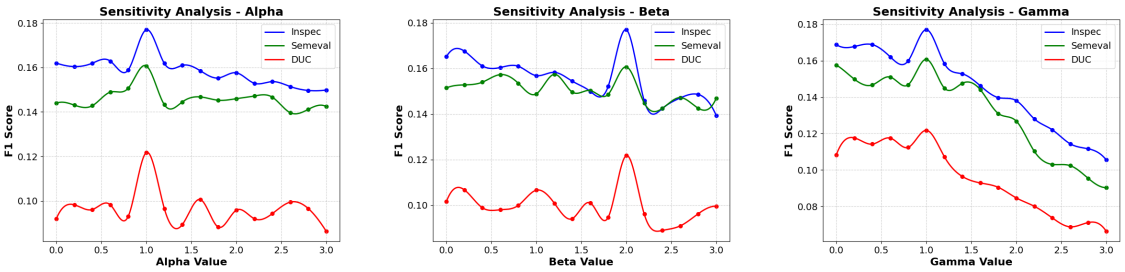

Figure 3.8: F1-score variation across different values of α, β, and γ for Inspec, SemEval, and DUC datasets.

The analysis indicates that moderate pheromone influence ($\alpha = 1$) provides the most stable performance, suggesting that historical path memory is valuable when used as reinforcement rather than as a dominant control signal. Excessive pheromone emphasis risks premature convergence by over-prioritizing previously traversed paths.

Among the three parameters, desirability demonstrated the strongest influence on performance. The optimal setting of $\beta = 2$ consistently produced superior F1-scores, indicating that semantic and contextual relevance play a central role in guiding effective path selection. This highlights the importance of emphasizing local phrase quality when identifying meaningful keyphrase structures.

Bayesian belief achieved its strongest contribution at $\gamma = 1$, where prior structural information effectively supported decision-making without constraining adaptive exploration. Larger values gradually reduced performance, suggesting that over-reliance on prior belief may suppress the framework's ability to respond dynamically to evolving optimization signals.

Overall, the results demonstrate that FuBAKE performs most effectively when semantic desirability is given slightly greater emphasis than pheromone reinforcement and Bayesian priors. Based on this analysis, $\alpha = 1$, $\beta = 2$, and $\gamma = 1$ were selected for all subsequent experiments.

3.5.4.2 FuBAKE Pheromone Evaporation Rate Analysis (ρ)

The pheromone evaporation rate ρ governs the rate at which accumulated pheromone information decays over successive iterations. This parameter is particularly important because it controls the balance between retaining useful historical learning and allowing continued ex-

ploration of alternative candidate paths.

To assess this behavior, ρ was varied from 0.1 to 1.0 in increments of 0.1. The resulting F1-score variations across all datasets are presented in Figure 3.9.

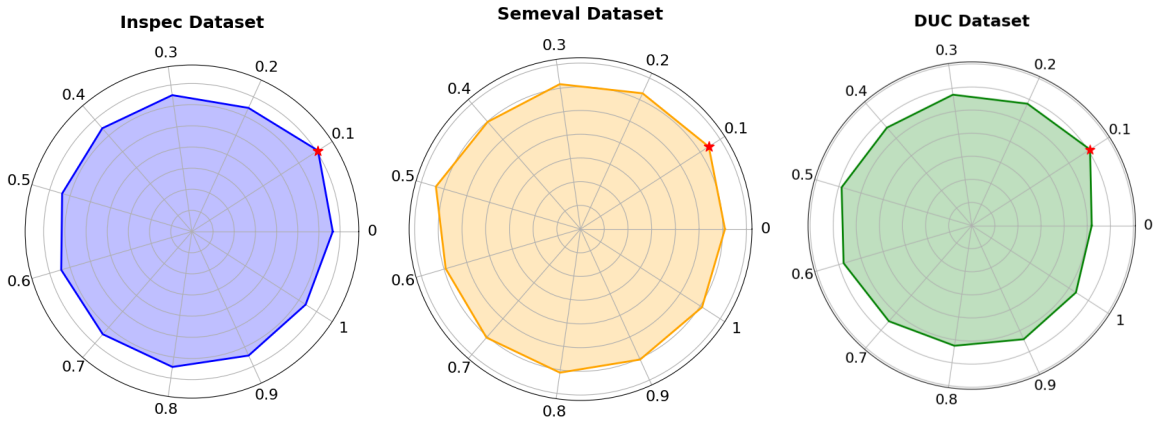

Figure 3.9: Radar charts illustrating F1-score variation across pheromone evaporation rates for Inspec, SemEval, and DUC datasets.

FuBAKE exhibited relatively stable performance across the tested evaporation range, indicating that the framework is not excessively sensitive to moderate changes in pheromone decay. Nevertheless, the strongest and most consistent performance was observed at $\rho = 0.1$.

This finding suggests that slower evaporation preserves beneficial historical information for a sufficient duration, enabling meaningful semantic paths to accumulate reinforcement while still allowing less informative paths to weaken gradually. In contrast, higher evaporation rates reduce historical continuity too aggressively, slightly diminishing optimization consistency.

Accordingly, $\rho = 0.1$ was adopted as the final pheromone evaporation setting due to its superior balance between exploration and reinforcement stability.

3.5.4.3 FuBAKE Convergence Analysis

Since FuBAKE is an iterative optimization framework, convergence behavior is a critical indicator of both computational efficiency and solution stability. A robust optimization process should converge reliably toward semantically meaningful solutions while avoiding unnecessary iterations or premature collapse into suboptimal paths.

To examine this property, convergence analysis was conducted across all benchmark datasets using four termination criteria: Ant Consensus, Pheromone Stability, Payoff Dominance, and Maximum Iteration. These criteria collectively represent different forms of optimization stabilization, ranging from strict collective agreement to mathematically stable saturation and

computational stopping conditions.

For each dataset, convergence behavior was analyzed using Fraction of Paragraphs (FP), Average Iteration (AI), and Cumulative Fraction (CF), as summarized in Table 3.3.

Table 3.3: Convergence behavior of FuBAKE across benchmark datasets in terms of Fraction of Paragraphs (FP), Average Iteration (AI), and Cumulative Fraction (CF).

Convergence Criterion	Inspec			SemEval 2017			DUC 2001		
	FP	AI	CF	FP	AI	CF	FP	AI	CF
Ant Consensus	71	30.4	71	72	46.8	72	63.5	53.8	63.5
Pheromone Stability	4	45.7	75	1	48.5	73	3	58.4	66.5
Payoff Dominance	5	60.3	80	2	56.2	75	2.5	61.7	69
Maximum Iteration	20	100	100	25	100	100	31	100	100

Across all datasets, Ant Consensus emerged as the dominant convergence mechanism, indicating that FuBAKE frequently stabilizes through strong collective path agreement. This suggests that the combined effects of pheromone reinforcement, semantic desirability, and Bayesian guidance often concentrate optimization toward high-confidence semantic paths relatively early.

Pheromone Stability and Payoff Dominance served as secondary convergence mechanisms, enabling mathematically stable termination when complete consensus was not achieved. Meanwhile, Maximum Iteration primarily occurred in structurally complex or semantically ambiguous documents, where prolonged exploration remained beneficial.

Overall, the convergence profile demonstrates that FuBAKE maintains an effective balance between exploration and exploitation. The framework consistently converges toward reliable semantic structures while preserving adaptability for more challenging textual environments, thereby confirming both its computational robustness and optimization stability.

3.5.4.4 TopS-Key PCA Variance Threshold Sensitivity Analysis

Within TopS-Key, Principal Component Analysis (PCA) is employed to reduce feature dimensionality while preserving the most informative variance. Since the retained variance threshold directly determines how much original feature information is preserved after dimensionality reduction, this parameter plays an important role in balancing semantic retention, redundancy reduction, and ranking efficiency.

To examine its effect, cumulative variance thresholds ranging from 0.80 to 1.00 were systematically evaluated across all benchmark datasets. Performance was assessed using F1-score at Top@3, Top@5, and Top@10, as illustrated in Figure 3.10.

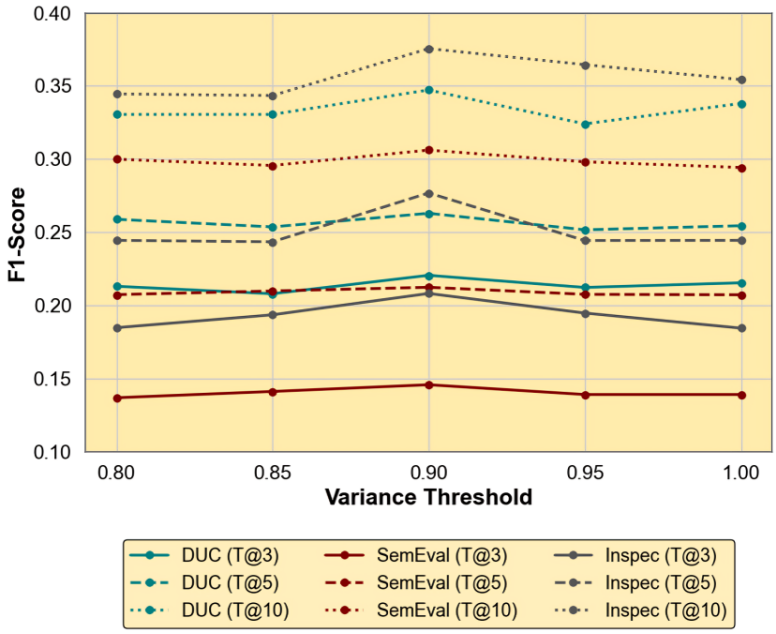

Figure 3.10: F1-score trends across varying PCA cumulative variance thresholds for all benchmark datasets

Across most datasets, performance improved as the retained variance threshold increased from 0.80 to 0.90, indicating that preserving a larger proportion of variance supports stronger semantic representation while still effectively reducing feature redundancy. This suggests that moderate dimensionality reduction may remove informative components prematurely, whereas retaining approximately 90% of variance preserves sufficient discriminative information for effective keyphrase ranking.

Beyond the 0.90 threshold, performance generally stabilized or exhibited slight decline. This pattern indicates that retaining excessive variance may reintroduce lower-importance components, noise, or redundant feature correlations that reduce the efficiency of PCA-driven feature compression. In particular, full variance retention diminished the practical benefits of dimensionality reduction without consistently improving ranking effectiveness.

Among the evaluated thresholds, 0.90 provided the most balanced and consistently effective performance across datasets and Top@N settings. This threshold preserved essential semantic structure while minimizing unnecessary redundancy, thereby supporting more stable entropy weighting and Fuzzy TOPSIS ranking.

Accordingly, a PCA cumulative variance threshold of 0.90 was selected as the final dimensionality reduction setting for TopS-Key and retained throughout subsequent experiments.

3.5.5 Ablation Study Analysis

To evaluate the functional contribution of individual features and architectural components within the proposed frameworks, ablation analysis is performed through systematic component removal. By selectively excluding specific features or methodological modules while preserving all other settings, this analysis quantifies the relative importance of each element in overall model performance.

Such evaluation is essential for determining whether performance gains arise from isolated components or from the synergistic integration of multiple mechanisms. It also provides deeper interpretability by identifying the most influential semantic, statistical, structural, and optimization factors within each framework.

3.5.5.1 FuBAKE Ablation Study Analysis

To examine the contribution of FuBAKE's major architectural components, an ablation study was conducted across the Inspec, SemEval, and DUC-2001 datasets by selectively removing or simplifying key modules while preserving the remainder of the framework.

Four ablated variants were analyzed: (i) removal of fuzzy inference, where raw pheromone and desirability values were used directly; (ii) removal of Bayesian belief updating, where beliefs remained static throughout optimization; (iii) removal of roulette-wheel selection, where stochastic traversal was replaced by deterministic greedy choice; and (iv) replacement of contextualized BERT embeddings with static Word2Vec embeddings. Figure 3.11 presents the resulting performance variations.

Among all components, replacing BERT embeddings produced the largest degradation, demonstrating that contextual semantic representation is central to FuBAKE's discriminative capability. The removal of roulette-wheel selection also caused substantial performance loss, particularly in recall, confirming that probabilistic exploration is critical for avoiding premature convergence and maintaining search diversity.

Fuzzy inference contributed moderate but consistent improvements, particularly by refining noisy decision boundaries and improving robustness under lexical variation. Bayesian belief updating produced the smallest immediate decline, yet remained valuable as a stabilizing mechanism that supports convergence consistency.

Overall, the results indicate that FuBAKE derives its strength from the integration of contextual

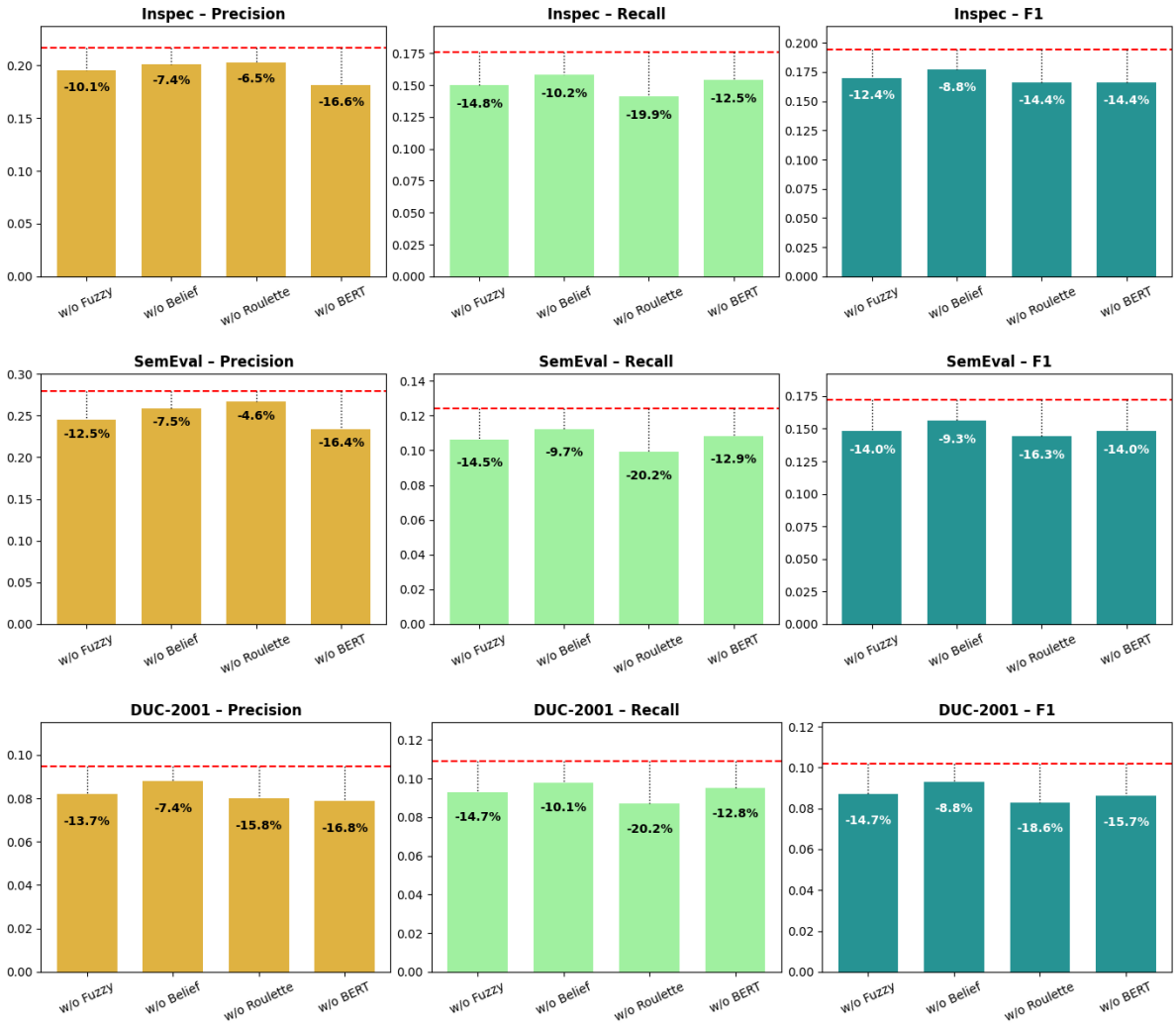

Figure 3.11: FuBAKE ablation analysis across Inspec, SemEval, and DUC-2001 datasets. The red dashed line denotes the complete FuBAKE baseline.

semantic representation and adaptive probabilistic exploration, supported by fuzzy refinement and Bayesian stabilization.

3.5.5.2 TopS-Key Ablation Study Analysis

For TopS-Key, ablation analysis was conducted at both feature and architectural levels to assess the relative importance of individual features as well as broader methodological components.

Feature-Level Ablation Analysis Each feature was removed independently and performance was re-evaluated across SemEval2017 Task 10, Inspec, and DUC2001 using precision, recall, and F1-score at Top@3, Top@5, and Top@10. Figure 3.12 summarizes the resulting performance variations.

Across all datasets, contextual semantic similarity features demonstrated the strongest influ-

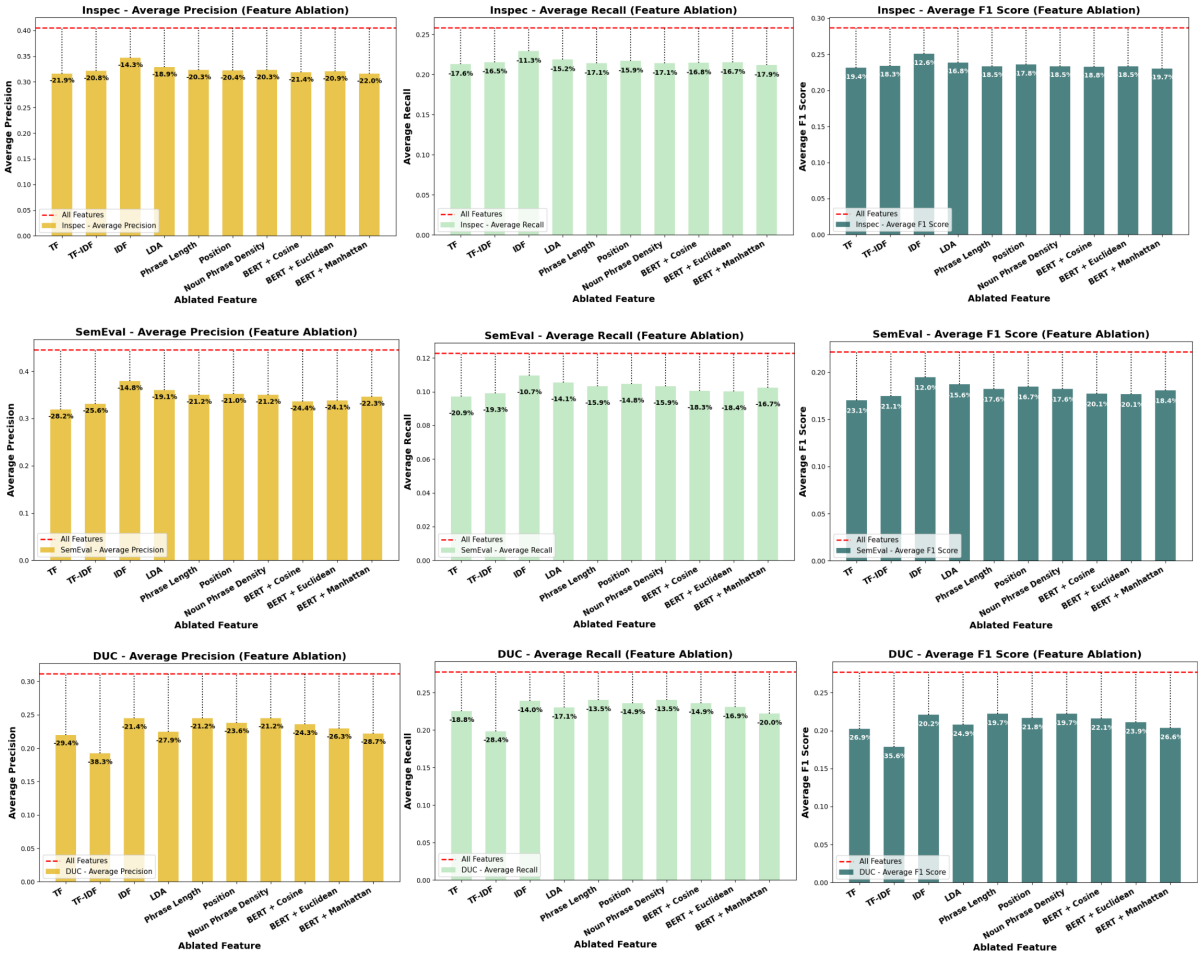

Figure 3.12: Feature-level ablation results showing the impact of removing individual TopSKey features across benchmark datasets

ence, with BERT + Manhattan distance producing the greatest performance degradation when removed. This confirms that fine-grained contextual semantic alignment serves as the dominant feature within TopS-Key's ranking framework.

Frequency-based features such as TF and IDF also produced substantial declines, reinforcing the continued importance of statistical prominence alongside contextual semantics. Structural features contributed comparatively smaller but consistently positive gains, indicating that positional and phrase-structure cues serve as complementary refinements rather than primary ranking drivers.

No individual feature removal improved performance, confirming that the full feature space provides the most balanced integration of semantic, statistical, and structural evidence.

Model Component Ablation Analysis To further assess architectural contribution, TopS-Key's three primary components, Contextual PCA, Adaptive Shannon Entropy Weights (ASEW),

and Fuzzy TOPSIS, were individually excluded and are evaluated using Eq. 3.60.

$$\Delta F_1^i = \frac{F_1^{\text{full}} - F_1^i}{F_1^{\text{full}}} \times 100\% \tag{3.60}$$

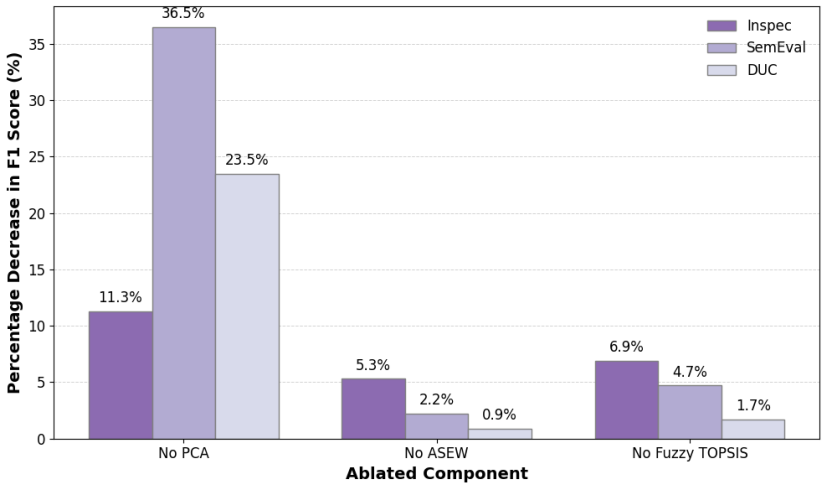

Figure 3.13: Percentage decrease in average F1-score resulting from exclusion of individual TopSKey components

As illustrated in Figure 3.13, Contextual PCA produced the most substantial performance decline across all datasets, establishing dimensionality reduction as the primary architectural driver. This demonstrates that PCA is essential not only for compression but also for preserving informative variance while suppressing redundancy.

ASEW provided moderate but consistent improvements through adaptive weighting, indicating that objective criterion weighting significantly enhances ranking discrimination. Fuzzy TOPSIS contributed smaller yet stable gains by refining final ranking decisions.

Collectively, these findings confirm that TopS-Key's effectiveness arises from the layered interaction of semantic feature engineering, variance-preserving dimensionality reduction, adaptive weighting, and fuzzy ranking refinement.

3.5.6 Comparative Performance Analysis

To comprehensively assess the effectiveness of the proposed frameworks, FuBAKE and TopS-Key were compared against a broad range of benchmark keyphrase extraction methods, including RAKE [39] using GloVe embeddings [55], PositionRank [46], MultipartiteRank [129], EmbedRank [130], YAKE [38], RoBERTa [57], KeyBERT [58], TripleRank [51], MDERank [131],PromptRank [132], HGUKE [71], BRYT [133], AdaptiveUKE [43], HCUKE [52], and

SDRank [53]. These baselines represent major statistical, graph-based, embedding-driven, transformer-based, and hybrid methodological paradigms, thereby ensuring comprehensive comparative coverage.

All models were evaluated under identical experimental settings using semantic similarity-based matching rather than strict lexical overlap to ensure fair and practically meaningful comparison. Performance was measured using Precision (P), Recall (R), and F1-score (F1) at Top@3, Top@5, and Top@10 across Inspec, SemEval2017 Task 10, and DUC2001.

This analysis therefore evaluates not only comparative performance superiority but also how differing methodological architectures behave across varying retrieval depths and document complexities.

3.5.6.1 Performance Comparison at Top@3

Table 3.4: Comparative performance of FuBAKE and TopS-Key against benchmark methods at Top@3

Method	**Inspec @N=3**			**SemEval2017 @N=3**			**DUC2001 @N=3**		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
FuBAKE	0.2958	0.1061	0.1472	0.3833	0.0731	0.1198	0.1476	0.0676	0.0908
TopS-Key	0.4242	0.1479	0.2081	0.4653	0.0890	0.1457	0.3241	0.1795	0.2205
RAKE + GloVe	0.1405	0.0583	0.0764	0.0980	0.0258	0.0397	0.0141	0.0149	0.0129
PositionRank	0.3772	0.1464	0.2109	0.4273	0.1030	0.1620	0.3231	0.1667	0.2200
MultipartiteRank	0.3330	0.1424	0.1847	0.3147	0.0778	0.1214	0.2884	0.1736	0.2036
EmbedRank	0.1673	0.0683	0.0906	0.1520	0.0349	0.0556	0.0519	0.0293	0.0359
YAKE	0.3602	0.1565	0.2018	0.3513	0.0857	0.1338	0.1883	0.1059	0.1292
RoBERTa	0.2217	0.1005	0.1256	0.2527	0.0596	0.0940	0.1407	0.0783	0.0955
KeyBERT	0.3347	0.1477	0.1887	0.3867	0.0938	0.1474	0.1672	0.0975	0.1156
TripleRank	0.2033	0.0874	0.1128	0.1920	0.0470	0.0737	0.0519	0.0300	0.0361
MDERank	0.2833	0.1199	0.1568	0.2847	0.0714	0.1110	0.1163	0.0663	0.0803
PromptRank	0.3680	0.1565	0.2040	0.3527	0.0881	0.1296	0.1461	0.0795	0.0992
HGUKE	0.2360	0.0852	0.1180	0.2706	0.0502	0.0828	0.2045	0.1092	0.1345
BRYT	0.2345	0.0895	0.1223	0.2239	0.0444	0.0725	0.1834	0.1060	0.1265
AdaptiveUKE	0.2563	0.0983	0.1327	0.2799	0.0531	0.0873	0.2911	0.1558	0.1958
HCUKE	0.3446	0.1228	0.1710	0.3766	0.0737	0.1205	0.2754	0.1493	0.1832
SDRank	0.3414	0.1151	0.1645	0.3650	0.0672	0.1110	0.2684	0.1377	0.1774

At Top@3, evaluation primarily emphasizes early-stage ranking precision, where only the highest-priority candidate keyphrases are retrieved. Under this setting, performance reflects a model's ability to identify the most semantically salient phrases with minimal tolerance for ranking error. As shown in Table 3.4, TopS-Key consistently demonstrated the strongest precision across all benchmark datasets while maintaining highly competitive F1-scores, indicating superior effectiveness in prioritizing dominant keyphrases at shallow retrieval depth.

On Inspec, TopS-Key achieved exceptionally strong precision and remained highly competitive with PositionRank in overall F1-score. This trend suggests that within concise technical

abstracts, where semantic density is high and lexical ambiguity is comparatively constrained, TopS-Key's integration of contextual feature scoring, dimensionality reduction, and entropy-guided ranking provides highly effective selective prioritization. PositionRank's close performance indicates that graph-based positional information remains useful in shorter structured documents; however, TopS-Key's broader methodological integration yields stronger consistency.

On SemEval2017, TopS-Key again achieved the highest precision, although PositionRank demonstrated slightly stronger F1-score due to marginally broader recall. This divergence implies that while TopS-Key is particularly precision-oriented at shallow depth, graph-based traversal may occasionally retrieve slightly broader semantic coverage in medium-length scientific paragraphs. Nevertheless, TopS-Key's consistently strong precision-recall balance suggests greater overall ranking reliability under semantically richer contexts.

The most substantial advantage emerged on DUC2001, where TopS-Key clearly outperformed nearly all benchmark methods across all metrics. Since DUC2001 contains longer and more structurally heterogeneous news articles, this result indicates that TopS-Key's feature integration and ranking framework remains highly robust even under substantial semantic complexity. In contrast, the comparatively weaker performance of statistical, static embedding, and several transformer-only methods highlights their limitations in maintaining semantic discrimination when document heterogeneity increases.

FuBAKE also demonstrated clear superiority over most baseline approaches, outperforming traditional statistical, embedding-only, and many recent hybrid systems. This confirms that its fuzzy Bayesian ant colony optimization architecture provides strong adaptive semantic prioritization. However, compared with TopS-Key, FuBAKE exhibited relatively lower shallow-depth precision, suggesting that its architecture may emphasize broader exploratory optimization over highly selective early-stage ranking.

Several broader methodological insights emerge from Top@3 analysis:

- TopS-Key provides the strongest early-stage ranking precision across heterogeneous textual environments.
- PositionRank remains a highly competitive graph-based baseline, particularly in shorter or moderately structured documents.
- FuBAKE offers robust optimization-driven retrieval but comparatively lower shallow-

depth ranking precision than TopS-Key.

- The persistent underperformance of RAKE + GloVe, EmbedRank, and similar approaches suggests that purely statistical or static embedding paradigms remain insufficient when contextual semantic prioritization is required.

Overall, the Top@3 comparative results indicate that TopS-Key offers the most effective framework for highly selective keyphrase extraction, particularly in scenarios where precision among the top-ranked outputs is the dominant objective.

3.5.6.2 Performance Comparison at Top@5

Table 3.5: Comparative performance of FuBAKE and TopS-Key against benchmark methods at Top@5

Method	Inspec @N=5			SemEval2017 @N=5			DUC2001 @N=5		
	P	R	F1	P	R	F1	P	R	F1
FuBAKE	0.2644	0.1553	0.1832	0.3304	0.1024	0.1515	0.1407	0.1048	0.1176
TopS-Key	0.4034	0.2291	0.2768	0.4553	0.1444	0.2123	0.2892	0.2591	0.2629
RAKE + GloVe	0.1442	0.0972	0.1089	0.0977	0.0417	0.0560	0.0183	0.0231	0.0191
PositionRank	0.3445	0.2354	0.2613	0.3980	0.1601	0.2209	0.2810	0.2557	0.2581
MultipartiteRank	0.2898	0.1937	0.2176	0.2929	0.1190	0.1635	0.2453	0.2303	0.2272
EmbedRank	0.1481	0.0990	0.1114	0.1444	0.0562	0.0785	0.0466	0.0486	0.0441
YAKE	0.2935	0.2022	0.2249	0.3033	0.1235	0.1693	0.1742	0.1639	0.1616
RoBERTa	0.2035	0.1383	0.1526	0.2312	0.0898	0.1254	0.1229	0.1120	0.1124
KeyBERT	0.2748	0.1920	0.2103	0.3129	0.1243	0.1729	0.1400	0.1300	0.1289
TripleRank	0.1912	0.1277	0.1432	0.1770	0.0697	0.0969	0.0441	0.0412	0.0411
MDERank	0.2618	0.1733	0.1958	0.2578	0.1076	0.1466	0.1148	0.1065	0.1068
PromptRank	0.3800	0.2339	0.2732	0.3710	0.1376	0.1939	0.1340	0.1174	0.1206
HGUKE	0.2095	0.1277	0.1482	0.2404	0.0734	0.1091	0.1617	0.1402	0.1439
BRYT	0.2485	0.1500	0.1761	0.2432	0.0815	0.1182	0.1752	0.1557	0.1583
AdaptiveUKE	0.2494	0.1545	0.1780	0.2684	0.0848	0.1250	0.2366	0.2058	0.2124
HCUKE	0.3300	0.1928	0.2287	0.3419	0.1088	0.1597	0.2547	0.2210	0.2279
SDRank	0.3155	0.1656	0.2067	0.3370	0.1012	0.1513	0.2242	0.1858	0.1980

At Top@5 as observed in Table 3.5, TopS-Key established a clearer overall performance lead across all datasets by consistently outperforming both FuBAKE and competing baselines in precision, recall, and F1-score. This demonstrates stronger retrieval scalability and broader ranking consistency. FuBAKE remained competitive but displayed comparatively lower expansion efficiency as candidate retrieval depth increased.

3.5.6.3 Performance Comparison at Top@10

In Table 3.6, broader extraction depth places significantly greater emphasis on ranking scalability, semantic consistency, and the ability to maintain precision-recall balance across larger candidate pools. Under this setting, TopS-Key emerged as the strongest overall performer

Table 3.6: Comparative performance of FuBAKE and TopS-Key against benchmark methods at Top@10

Method	Inspec @N=10			SemEval2017 @N=10			DUC2001 @N=10		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
FuBAKE	0.2023	0.2247	0.2000	0.2543	0.1551	0.1855	0.1181	0.1757	0.1384
TopS-Key	0.3875	0.3975	0.3755	0.4144	0.2575	0.3061	0.3208	0.3935	0.3472
RAKE + GloVe	0.1874	0.2021	0.1887	0.1037	0.0850	0.0906	0.0285	0.0383	0.0314
PositionRank	0.3361	0.3509	0.3337	0.3406	0.2627	0.2876	0.3174	0.3913	0.3460
MultipartiteRank	0.2834	0.2966	0.2817	0.2417	0.1899	0.2060	0.2567	0.3260	0.2816
EmbedRank	0.1658	0.1747	0.1658	0.1366	0.1051	0.1154	0.0506	0.0708	0.0569
YAKE	0.2678	0.2865	0.2699	0.2510	0.1991	0.2150	0.2075	0.2623	0.2268
RoBERTa	0.2045	0.2133	0.2025	0.1918	0.1453	0.1602	0.1400	0.1803	0.1587
KeyBERT	0.2498	0.2691	0.2515	0.2296	0.1768	0.1943	0.1566	0.2072	0.1738
TripleRank	0.2033	0.2074	0.1994	0.1727	0.1293	0.1437	0.0629	0.0765	0.0680
MDERank	0.2579	0.2622	0.2533	0.2267	0.1792	0.1947	0.1612	0.2000	0.1758
PromptRank	0.3692	0.3091	0.3251	0.3969	0.2340	0.2802	0.1707	0.2000	0.1805
HGUKE	0.1957	0.2374	0.2010	0.2219	0.1365	0.1627	0.1430	0.1778	0.1548
BRYT	0.2396	0.2808	0.2431	0.2467	0.1593	0.1864	0.2083	0.2568	0.2255
AdaptiveUKE	0.2182	0.2546	0.2209	0.2347	0.1472	0.1738	0.2201	0.2686	0.2379
HCUKE	0.3008	0.3359	0.2997	0.3109	0.1954	0.2308	0.2868	0.3533	0.3109
SDRank	0.2828	0.2365	0.2432	0.3119	0.1626	0.2054	0.2163	0.2569	0.2302

across all three benchmark datasets, consistently achieving the highest Precision, Recall, and F1-score. This dominant performance indicates that TopS-Key's integrated architecture combining contextual feature engineering, PCA-based dimensionality reduction, Shannon entropy-adaptive weighting, and Fuzzy TOPSIS ranking remains highly stable even when retrieval scope expands substantially.

On Inspec, TopS-Key achieved an F1-score of 0.3755, substantially outperforming FuBAKE (0.2000) and all benchmark methods. Although PositionRank remained the closest competitor with an F1-score of 0.3337, TopS-Key preserved a notable advantage, suggesting that its ranking-centric architecture is particularly effective in concise technical documents where semantic precision and feature integration are critical.

On SemEval2017, which contains more contextually complex scientific paragraphs, TopS-Key again demonstrated clear superiority with an F1-score of 0.3061, outperforming PositionRank (0.2876), PromptRank (0.2802), and FuBAKE (0.1855). This indicates that TopS-Key's dimensionality reduction and adaptive weighting mechanisms remain robust under medium-complexity semantic environments.

The strongest comparative evidence emerged on DUC2001, where TopS-Key achieved an F1 score of 0.3472, narrowly surpassing PositionRank (0.3460) while substantially outperforming FuBAKE (0.1384) and most transformer or hybrid baselines. Since DUC2001 contains the longest and most structurally diverse documents, this result highlights TopS-Key's ability to

preserve ranking precision even under substantial semantic and structural heterogeneity.

FuBAKE, while not matching TopSKey's broader ranking dominance, continued to outperform many benchmark methods, particularly classical statistical and several neural baselines. Its comparatively lower Top@10 performance suggests that while its fuzzy Bayesian ant colony optimization is highly effective for adaptive semantic exploration, it is relatively less optimized for large-scale ranking breadth than TopS-Key's multi-criteria decision framework.

Several broader trends are evident from the Top@10 analysis:

- TopS-Key consistently provides the strongest large-scale retrieval performance across short, medium, and long documents.
- PositionRank remains a strong graph-based competitor, particularly at larger retrieval depths, but is generally surpassed by TopS-Key.
- FuBAKE demonstrates strong robustness but exhibits comparatively lower retrieval scalability at expanded extraction depth.
- Traditional statistical, embedding-only, and several transformer-based methods show clear performance limitations relative to the proposed frameworks.

Overall, the Top@10 comparative results confirm that TopS-Key offers the most scalable and globally consistent ranking architecture within this chapter, particularly when broader candidate retrieval is required, while FuBAKE remains a robust optimization driven alternative with strengths more concentrated in adaptive search than large-scale ranking expansion.

3.5.7 Statistical Significance Analysis

Although comparative evaluation metrics such as Precision, Recall, and F1-score provide direct quantitative evidence of effectiveness, descriptive performance superiority alone does not fully establish whether the observed gains of the proposed frameworks are statistically reliable. Because benchmark improvements may occasionally arise from dataset-specific variability, methodological coincidence, or retrieval fluctuations, formal statistical validation is necessary to determine whether FuBAKE and TopS-Key represent genuine and reproducible advancements in automatic keyphrase extraction. Accordingly, statistical significance analysis was conducted separately for both frameworks using validation strategies aligned with their distinct architectural objectives.

For FuBAKE, statistical reliability was evaluated using the two-tailed Z-test and the Wilcoxon signed-rank test. The Z-test, a parametric significance measure for mean-level comparison [134], was used to determine whether FuBAKE's aggregate F1-score differed significantly from the benchmark mean. In parallel, the Wilcoxon signed-rank test, a non-parametric paired significance test [135], examined whether the observed improvements remained consistent across individual document instances. Together, these tests provide both global and document-level validation of FuBAKE's optimization-centric architecture.

For TopS-Key, statistical validation was conducted using the two-tailed Z-test and one-way Analysis of Variance (ANOVA) followed by Tukey's Honestly Significant Difference (HSD) post-hoc analysis [136]. While the Z-test established mean-level superiority relative to benchmark averages, ANOVA [137] was employed to determine whether statistically significant performance variance existed across the broader population of competing methods. Tukey HSD subsequently enabled pairwise identification of TopS-Key's specific methodological distinction relative to alternative systems. This combination provides both aggregate and large-scale comparative validation of TopS-Key's ranking-centric framework.

3.5.7.1 Statistical Validation of FuBAKE

For FuBAKE, the two-tailed Z-test was first applied across Inspec, SemEval2017, and DUC2001 to determine whether its aggregate F1-score significantly exceeded the collective benchmark mean. Across all benchmark datasets, the observed Z-scores consistently surpassed the critical threshold of 1.96 at the 95% confidence level, thereby rejecting the null hypothesis and confirming statistically significant aggregate level superiority.

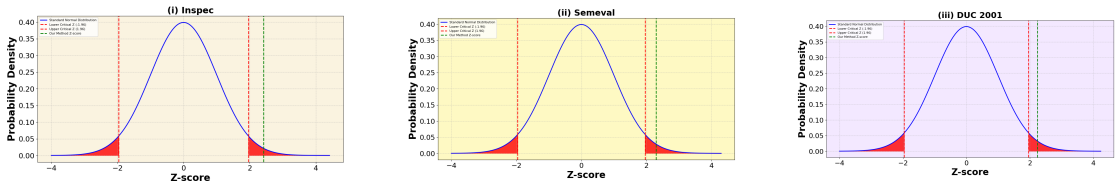

Figure 3.14: Two-tailed Z-test distributions illustrating FuBAKE's aggregate statistical superiority over benchmark methods across Inspec, SemEval2017, and DUC2001.

Figure 3.14 illustrates the Z-test distributions for all three benchmark datasets and confirms that FuBAKE's observed mean-level improvements are systematic rather than incidental, suggesting that its fuzzy Bayesian ant colony optimization framework provides statistically reliable aggregate performance enhancement across datasets of varying structural complexity.

To further examine whether these gains extend beyond aggregate averages into document-

level consistency, the Wilcoxon signed-rank test was subsequently employed. This test is particularly appropriate for optimization-driven systems because it directly evaluates paired per-document performance differences without imposing strict normality assumptions. Across all Top@N configurations and benchmark datasets, extremely small p-values consistently rejected the null hypothesis, thereby confirming that FuBAKE's superiority remains stable across individual document instances.

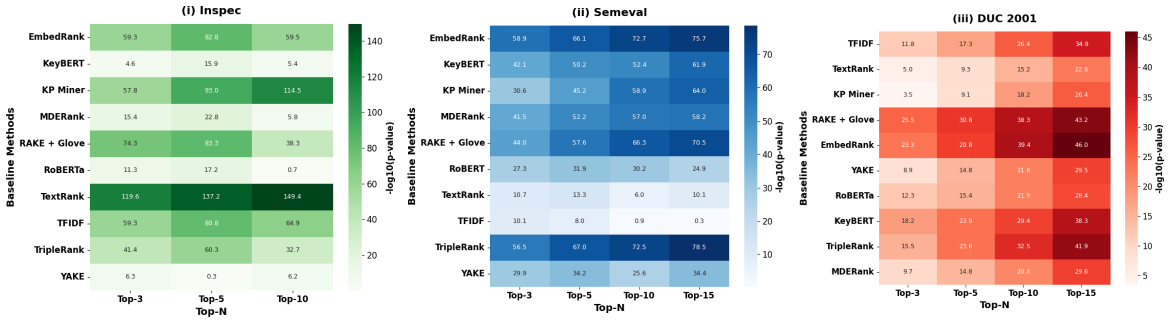

Figure 3.15: Wilcoxon signed-rank significance analysis illustrating FuBAKE's paired document-level superiority across benchmark datasets.

As illustrated in Figure 3.15, the combined Z-test and Wilcoxon analyses demonstrate that FuBAKE's statistical superiority is not confined to benchmark-level averages alone. Rather, its improvements remain both globally significant and consistently reliable across individual retrieval instances. This dual validation substantially strengthens confidence in FuBAKE's optimization stability and confirms that its observed gains reflect genuine methodological advancement rather than random fluctuation.

3.5.7.2 Statistical Validation of TopS-Key

For TopS-Key, statistical analysis similarly began with the two-tailed Z-test to determine whether its mean F1-score significantly exceeded the collective benchmark average across Inspec, SemEval2017, and DUC2001. In each dataset, observed Z-scores exceeded the critical threshold of 1.96, thereby rejecting the null hypothesis and confirming statistically significant aggregate-level superiority.

As shown in Figure 3.16, TopS-Key consistently achieves statistically significant aggregate-level improvements across all benchmark datasets. These results reinforce the robustness of its PCA-driven, entropy-weighted Fuzzy TOPSIS architecture and validate the effectiveness of its ranking-centric design.

However, mean-level superiority alone does not establish whether TopS-Key is statistically distinguishable from the broader methodological landscape. To address this, one-way ANOVA

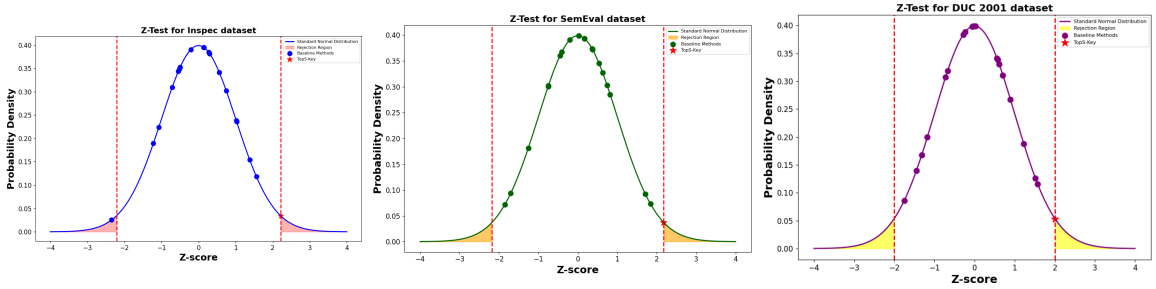

Figure 3.16: Two-tailed Z-test distributions illustrating TopS-Key's aggregate statistical superiority over benchmark methods across benchmark datasets.

was employed to evaluate whether statistically significant performance variance existed across all competing methods collectively. Across Inspec, SemEval2017, and DUC2001, the observed F-statistics substantially exceeded critical thresholds, thereby rejecting the null hypothesis that all methods perform equivalently. These findings indicate that TopS-Key's superiority extends beyond average benchmark advantage into broader methodological distinction.

Because ANOVA identifies only the presence of significant variance, Tukey HSD post-hoc analysis was subsequently applied to determine pairwise distinctions. Across all benchmark datasets, TopS-Key consistently demonstrated statistically significant superiority over competing methods under Tukey HSD criteria, thereby confirming that it occupies a statistically distinct performance tier rather than merely outperforming weaker baselines.

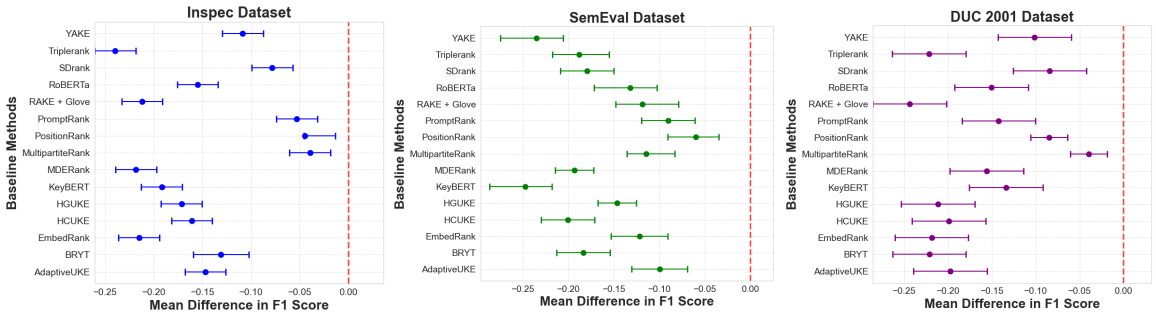

Figure 3.17: ANOVA and Tukey HSD significance analysis illustrating TopS-Key's statistically distinct performance tier across benchmark datasets.

As illustrated in Figure 3.17, the Z-test, ANOVA, and Tukey HSD analyses collectively demonstrate that TopS-Key's observed gains are not simply benchmark-specific numerical improvements, but statistically reliable methodological distinctions maintained across heterogeneous competing paradigms.

Overall, the statistical analyses conducted for both frameworks provide strong empirical evidence that the performance gains of FuBAKE and TopS-Key are statistically significant, methodologically meaningful, and reproducible across diverse benchmark environments.

FuBAKE demonstrates particularly strong optimization consistency at both aggregate and paired document levels, while TopS-Key exhibits robust ranking superiority and statistically distinct large-scale comparative performance. Together, these findings substantially reinforce the empirical credibility of the proposed frameworks as genuine contributions to the advancement of automatic keyphrase extraction.

3.5.8 Evaluation of ForestRank

This subsection evaluates ForestRank by examining three core aspects: feature importance, the impact of feature refinement, and comparative performance against benchmark keyphrase extraction methods. Unlike FuBAKE and TopS-Key, which primarily emphasize optimization and fuzzy ranking, ForestRank focuses on feature-driven ranking through the integration of Random Forest-based feature importance estimation with the WASPAS framework. Consequently, its evaluation centers on identifying the most informative features and assessing how their selection influences extraction effectiveness.

Experiments were conducted on the SemEval2017 Task 10 dataset [127], which provides a balanced scientific benchmark across multiple technical domains. This dataset is particularly suitable for ForestRank because its moderate document complexity allows clearer assessment of feature relevance in ranking performance.

3.5.8.1 Feature Importance Analysis

To determine which features contribute most significantly to keyphrase extraction, nine statistical, structural, and semantic features were initially computed for each candidate noun phrase. Random Forest was then employed to estimate feature importance based on each feature's contribution to classification quality.

Table 3.7: Feature importance scores in ForestRank

Feature	**Importance**
Term Frequency	0.086505
Inverse Document Frequency	0.015408
TF-IDF	0.090780
Phrase Length	0.036940
Noun Phrase Density	0.086007
Position Score	0.149263
Semantic Similarity	0.176357
Word2Vec Score	0.143981
BERT Score	0.214760

As shown in Table 3.7, BERT similarity emerged as the most influential feature, followed by semantic similarity, position score, and Word2Vec similarity. This pattern indicates that contextual-semantic representations contribute more strongly to extraction quality than conventional statistical measures. In particular, the dominance of BERT suggests that deep contextual understanding plays a central role in distinguishing meaningful keyphrases from less informative candidates.

Based on these findings, the four highest-ranked features were retained for subsequent ranking, while lower-contributing features were excluded to reduce redundancy.

3.5.8.2 Performance Impact of Feature Refinement

To evaluate the practical value of feature selection, ForestRank was tested using both the full nine-feature set and the optimized four-feature subset within the WASPAS framework. Performance was measured at Top@5, Top@10, and Top@15 using fuzzy similarity matching.

Table 3.8: Performance comparison using all features and optimized features

Top-N	**All Features**	**Best Features**
Top@5	95.00%	95.20%
Top@10	98.60%	99.20%
Top@15	99.20%	99.60%

Table 3.8 demonstrates that the optimized feature subset consistently outperformed the complete feature space across all retrieval depths. Although the numerical improvements are moderate, they confirm that selective feature refinement enhances ranking precision by reducing noise and preserving only the most discriminative information. This supports the methodological premise that feature quality is more influential than feature quantity in ranking-based extraction systems.

3.5.8.3 Comparative Performance Analysis

To position ForestRank within the broader keyphrase extraction landscape, its performance was compared against TF-IDF [138], TextRank [14], RoBERTa [57], RAKE + GloVe [39], KeyBERT [58], and PositionRank [46].

As illustrated in Figure 3.18, ForestRank consistently achieved the highest performance across all evaluation depths. This superiority suggests that combining machine learning-based feature prioritization with WASPAS ranking offers a more balanced and effective extraction strategy than purely statistical, graph-based, or embedding-only methods.

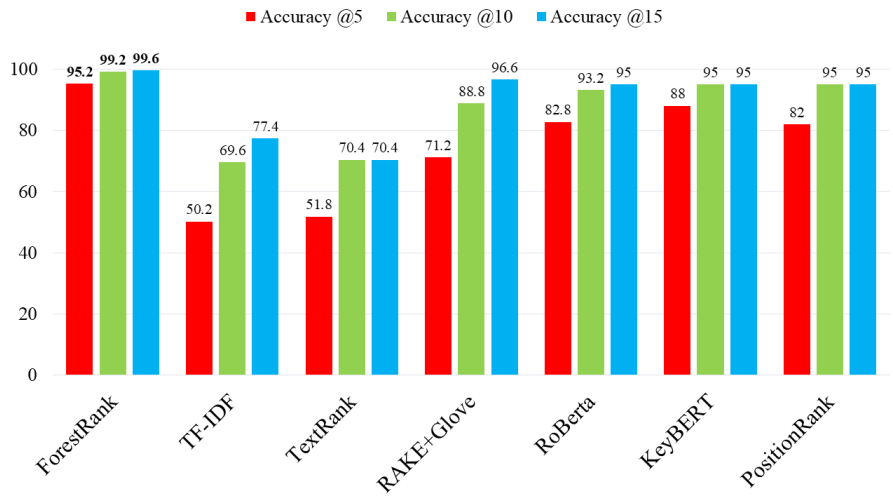

Figure 3.18: Comparative performance of ForestRank against benchmark methods across Top@5, Top@10, and Top@15.

Overall, the evaluation confirms that ForestRank's strength lies in its ability to identify and prioritize the most informative features before ranking. By integrating feature importance estimation with structured multi-criteria decision-making, ForestRank achieves strong extraction effectiveness while maintaining interpretability and computational efficiency.

3.6 Conclusion

This chapter presented three complementary keyphrase extraction frameworks, namely ForestRank, TopS-Key, and FuBAKE, each designed to address limitations of conventional statistical and graph-based extraction methods through progressively advanced semantic modeling and optimization strategies.

ForestRank introduced a feature-driven extraction framework combining Random Forest-based feature importance estimation with WASPAS ranking, demonstrating the effectiveness of contextual and semantic feature prioritization for improving extraction quality. TopS-Key further enhanced this direction through PCA-based dimensionality reduction, Shannon entropy weighting, and Fuzzy TOPSIS ranking, thereby reducing feature redundancy while improving ranking robustness through uncertainty-aware multi-criteria decision making.

FuBAKE extended the framework toward optimization-centric extraction by integrating fuzzy logic, Bayesian belief updating, and ant colony optimization within a unified semantic extraction architecture. The framework demonstrated strong robustness across structurally diverse and semantically complex datasets through adaptive exploration and probabilistic refinement

mechanisms.

Overall, the progression from ForestRank to TopS-Key and FuBAKE reflects a clear methodological evolution from feature-centric ranking toward adaptive semantic optimization. The experimental findings consistently demonstrate that semantic contextualization, uncertainty-aware ranking, and optimization-driven selection substantially improve keyphrase extraction effectiveness compared with conventional approaches, thereby establishing a strong foundation for subsequent social media and multimodal keyword extraction frameworks presented in the thesis.

Chapter 4

Keyword Extraction from Social Media Text

Keyword extraction from social media text presents a distinct and increasingly important challenge in natural language processing, aimed at identifying salient terms that capture the underlying intent, semantics, and communicative significance of short, informal, and dynamically evolving user-generated content. Unlike conventional textual documents, social media data is characterized by brevity, linguistic irregularity, contextual ambiguity, and domain-specific expression, which limit the effectiveness of traditional keyword extraction frameworks. Existing methods often struggle to adapt to noisy short-text environments and frequently overlook the functional diversity of social discourse, resulting in reduced contextual relevance and suboptimal keyword identification.

This chapter focuses on developing socially adaptive methodologies for keyword extraction by integrating semantic modeling, contextual intelligence, and domain-specific computational strategies tailored to diverse social media environments.

The chapter begins with social media text preprocessing, which establishes the foundation for effective candidate keyword generation by addressing the unique structural and linguistic challenges of short-text data. Following this, three proposed frameworks, *COM-Key*, *QuToX*, and *HumorKey*, are introduced to address keyword extraction across progressively specialized social media contexts. COM-Key focuses on commonsense-augmented keyword discov-

ery for general social communication, QuToX extends this framework toward toxicity-aware and interpretable keyword extraction in harmful discourse, and HumorKey addresses humor-sensitive keyword extraction through computational modeling of incongruity and expressive variation.

4.1 Social Media Text Preprocessing

Preprocessing is a critical stage in social media keyword extraction, as raw social media text is often characterized by noise, brevity, informal linguistic structures, and platform specific elements that can hinder meaningful semantic analysis. Unlike conventional textual documents, social media content frequently contains mentions, hashtags, hyperlinks, contractions, emojis, abbreviations, and irregular expressions that require specialized normalization. Consequently, preprocessing in this domain must not only reduce structural and lexical noise but also preserve contextually relevant social signals.

This section presents the preprocessing framework adopted for social media text in this chapter. It first outlines standard preprocessing techniques adapted for noisy short-text environments, followed by an extended preprocessing strategy designed to further refine semantic quality and contextual relevance for more effective keyword extraction.

4.1.1 Standard Preprocessing and Linguistic Normalization for Social Media Text

Building upon the general preprocessing principles previously established for textual documents, social media text requires additional normalization to address platform-specific structural irregularities and informal linguistic conventions. While foundational operations such as lowercasing, tokenization, stopword removal, and lemmatization remain consistent with conventional preprocessing, social media environments introduce unique artifacts such as mentions, hashtags, hyperlinks, contractions, and abbreviated expressions that require specialized handling [139, 140].

A primary distinction in social media preprocessing involves the treatment of platform-specific markers. User mentions (e.g., `@username`) are removed because they primarily function as social identifiers rather than semantically meaningful lexical units. Hashtags, however, fre-

quently encode topical significance; therefore, the symbolic prefix is discarded while preserving the textual component to retain contextual relevance. Similarly, hyperlinks, retweet markers, and extraneous punctuation are filtered to reduce structural noise while preserving linguistic clarity.

Social media discourse also commonly includes contractions and compressed informal forms, which can increase lexical inconsistency. To address this, contractions are expanded into their canonical equivalents to standardize vocabulary and improve downstream semantic representation [139]. Where syntactic structure is beneficial, Part-of-Speech tagging may additionally be employed to preserve grammatically informative categories such as nouns, adjectives, and modifiers, which often carry stronger keyword salience in short-text environments [141].

Figure 4.1 presents a representative tweet from the IPL 2019 dataset prior to preprocessing, illustrating the noisy and structurally irregular nature of raw social media text. The example contains user mentions, hashtags, punctuation irregularities, and redundant lexical components that may hinder semantic interpretation if left unprocessed.



Figure 4.1: Sample raw tweet from the IPL 2019 dataset before preprocessing

The tweet is subsequently preprocessed to produce a cleaner and more semantically interpretable textual representation suitable for keyword extraction and downstream analysis.

Pre-processed Tweet: Knows Ego Boost Everyone Politics Trump Election

This example demonstrates how standard preprocessing transforms a noisy, platform-dependent post into a normalized textual representation by removing structurally irrelevant elements such as mentions, hashtags, punctuation, emojis, and stopwords while preserving semantically meaningful lexical content. This process establishes a linguistically consistent basis for subsequent semantic refinement and domain-specific keyword extraction.

4.1.2 Extended Preprocessing Strategy for Social Media Text

Despite standard linguistic normalization, social media text frequently remains affected by lexical sparsity, irregular recurrence patterns, and semantically weak token variation due to its short length, informal expression, and highly dynamic discourse structure. In such environments, structurally valid tokens may still contribute limited communicative value, while socially meaningful keywords may be obscured by sparse noise, repetitive platform expressions, or idiosyncratic lexical artifacts. Consequently, social media keyword extraction benefits from an additional refinement stage specifically designed to enhance semantic salience within noisy short-text environments.

To address this challenge, extended preprocessing introduces recurrence-aware semantic filtering through Standard Instance Recurrence (SIR) [84], which evaluates the normalized recurrence strength of each token across the social media corpus. Let $f(w_i)$ denote the frequency of token w_i across all social media posts, and let N represent the total number of token occurrences in the dataset. The SIR score for token w_i is defined as Eq. 4.1.

$$\mathrm{SIR}(w_i) = \frac{f(w_i)}{N} \tag{4.1}$$

This score measures the proportional recurrence of each lexical unit relative to the broader corpus, thereby identifying tokens that demonstrate stronger distributional significance within social discourse.

To distinguish semantically meaningful tokens from weak or sparse lexical elements, the mean SIR score across all unique tokens is computed as Eq. 4.2.

$$\mu_{\mathrm{SIR}} = \frac{1}{M} \sum_{i=1}^{M} \mathrm{SIR}(w_i) \tag{4.2}$$

where M denotes the total number of unique tokens. Tokens satisfying the Eq. 4.3.

$$\mathrm{SIR}(w_i) \geq \mu_{\mathrm{SIR}} \tag{4.3}$$

are retained for further analysis, while tokens below this threshold are filtered out.

The underlying rationale is that socially relevant lexical units are more likely to recur with meaningful proportional consistency, whereas sparse or anomalous tokens often reflect struc-

tural noise, isolated user behavior, or semantically weak variation. Unlike simple frequency filtering, SIR-based refinement emphasizes normalized recurrence, thereby preserving contextually representative vocabulary while minimizing distortive lexical outliers.

This refinement is particularly valuable in social media environments because short-text discourse often contains disproportionate lexical variation, where noisy expressions may coexist alongside highly informative social cues. By selectively preserving semantically recurrent tokens, extended preprocessing improves candidate consistency, strengthens contextual reliability, and reduces redundancy without excessively suppressing socially meaningful diversity.

Overall, this extended strategy functions as a social-media-specific semantic enrichment layer that complements standard preprocessing by improving lexical relevance and contextual precision. Through recurrence aware refinement, social media text is transformed into a more semantically stable and extraction-ready representation, thereby providing a stronger foundation for subsequent communication aware, toxicity sensitive, and humor-oriented keyword extraction frameworks.

4.2 COM-Key: Commonsense-Augmented Fuzzy Graphs with Multi-Centrality for Keyword Discovery and Knowledge Mining in Social Media Data

COM-Key formulates keyword extraction in social media text as a semantically structured graph modeling problem, where candidate terms are represented as concept-level nodes enriched through contextual embeddings and external commonsense knowledge. Rather than relying on isolated lexical features, the framework constructs a fuzzy conceptual graph in which relationships between candidate keywords are modeled through graded semantic associations derived from both embedding similarity and ConceptNet-based knowledge expansion. This enables the representation of candidate terms within a unified semantic topology that captures both contextual and commonsense relationships.

The framework is organized into four sequential stages: semantic embedding and clustering of candidate tokens, commonsense-enriched fuzzy graph construction, fuzzy centrality-based importance modeling, and final keyword ranking through multi-criteria decision-making. Through this design, COM-Key transforms semantically normalized tokens into a structured conceptual graph, allowing keyword importance to be evaluated based on both structural connectivity and

contextual relevance. Figure 4.2 presents the complete methodological workflow of COM-Key.

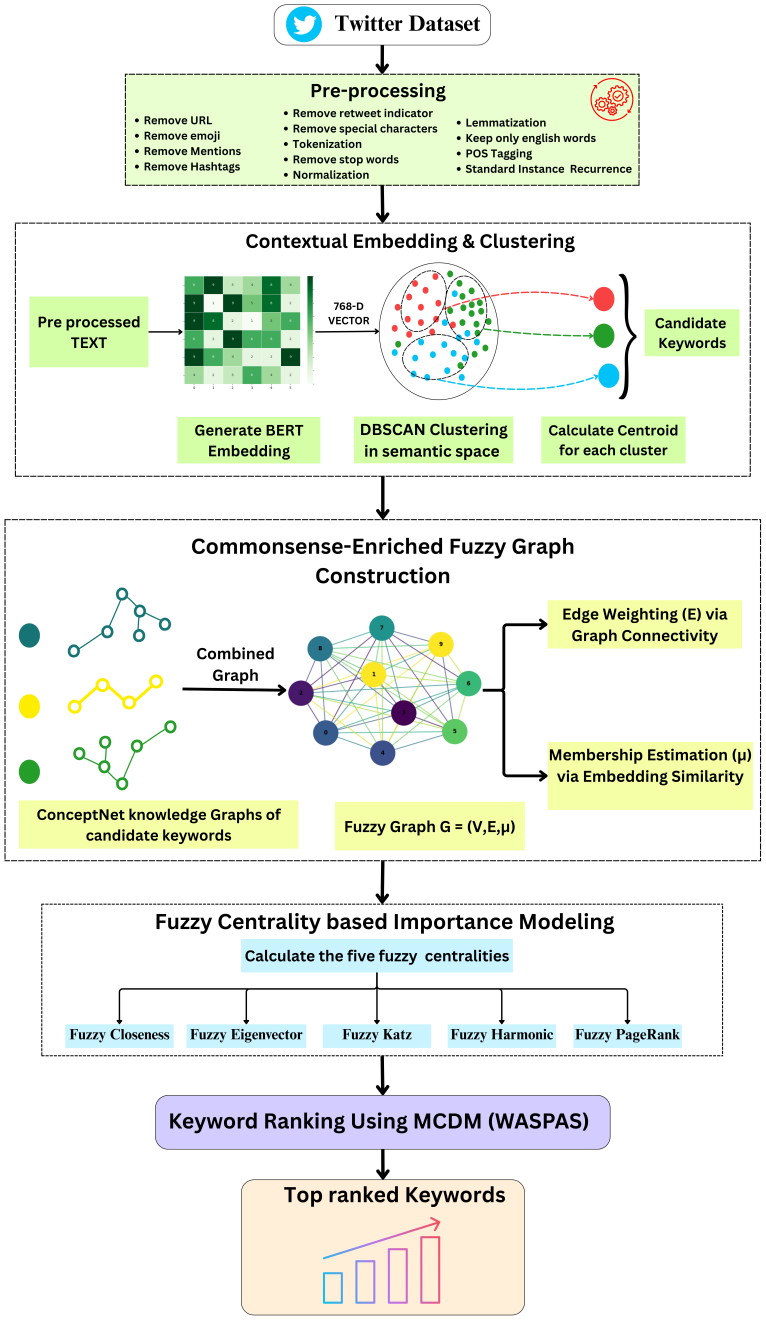

Figure 4.2: Overall methodological framework of COM-Key for social media keyword extraction

4.2.1 Semantic Embedding and Clustering of Candidate Tokens

The informal, compressed, and linguistically irregular nature of social media discourse presents substantial semantic ambiguity, making direct keyword identification from raw short-text en-

vironments inherently challenging. Even after preprocessing, social media content often contains fragmented lexical variation, abbreviated expressions, and context-sensitive semantics that require deeper representational grounding before reliable keyword extraction can occur. Consequently, the first stage of COM-Key focuses on transforming normalized social media text into semantically structured candidate concepts suitable for higher-level graph construction.

Since preprocessing and semantic normalization strategies for social media text have already been established in the preceding section, this stage begins with a linguistically refined token space in which structural noise has been minimized while socially relevant lexical signals are preserved. The objective is therefore not further normalization, but semantic contextualization and concept discovery.

Let the filtered token set be represented as $W = \{w_1, w_2, \ldots, w_m\}$. To capture contextual meaning beyond surface lexical similarity, each token $w_i \in W$ is encoded using BERT [56], which generates high-dimensional contextual embeddings $\mathbf{e}(w_i) \in \mathbb{R}^d$ capable of modeling bidirectional semantic dependencies, where d denotes the dimensionality of the embedding space (e.g., $d = 768$ for BERT-base). This contextual embedding process is particularly important in social media environments, where lexical meaning is often highly dependent on surrounding discourse and informal usage patterns.

While token-level embeddings provide semantic richness, effective keyword extraction requires broader conceptual organization through the identification of semantically coherent lexical groupings. To achieve this, COM-Key employs Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [142], which groups semantically similar tokens into dense conceptual neighborhoods while filtering sparse outliers and noise.

DBSCAN is particularly suitable for short-text social media environments because it does not require prior specification of cluster count and can effectively distinguish dense semantic manifolds from isolated lexical irregularities. Let the resulting semantic clusters be defined as $\mathscr{C} = \{C_1, C_2, \ldots, C_k\}$, where each cluster C_j represents a semantically coherent grouping of contextually related lexical units. To obtain a representative concept-level embedding for each cluster, the semantic centroid is computed as Eq. 5.14.

$$\mathbf{c}_j = \frac{1}{|C_j|} \sum_{w_i \in C_j} \mathbf{e}(w_i) \tag{4.4}$$

These centroids function as semantically aggregated candidate concepts, enabling the transition from fragmented lexical units to structured conceptual representations. This transformation is methodologically significant because social media communication often conveys salient ideas through dispersed but semantically related expressions rather than repeated exact terminology.

Overall, semantic embedding and clustering establish the conceptual-semantic substrate of COM-Key by converting normalized short-text inputs into coherent concept-level candidates. These centroid-based conceptual structures subsequently serve as the foundational nodes for commonsense enrichment and fuzzy graph construction in the next stage of the framework.

4.2.2 Commonsense-Enriched Fuzzy Graph Construction

Although semantic embedding and density-based clustering provide an effective foundation for identifying semantically coherent candidate concepts, contextual similarity alone is often insufficient for modeling the deeper conceptual associations required in noisy social media environments. Social media discourse frequently contains culturally contextualized references, implicit semantics, domain-specific terminology, and metaphorical expressions whose significance may not be fully recoverable through statistical language representations alone. While contextual embedding models such as BERT capture distributional semantics effectively, they primarily reflect lexical co-occurrence patterns and may underrepresent broader commonsense relationships, particularly when meaningful associations are implicit, sparse, or socially inferred. Therefore, to enhance semantic interpretability beyond contextual proximity, COM-Key extends candidate concept modeling through the integration of external commonsense knowledge using ConceptNet [143].

ConceptNet is a large-scale semantic knowledge graph that encodes human commonsense through semantically typed relationships between concepts. Unlike conventional embedding spaces, where semantic relatedness is inferred from usage patterns, ConceptNet explicitly represents structured conceptual relations such as *IsA*, *PartOf*, *UsedFor*, *HasA*, and *RelatedTo*. This distinction is particularly important in social media analysis because socially relevant concepts are often implied through cultural or contextual references rather than explicit lexical repetition. Consequently, ConceptNet enables candidate keywords to be expanded into broader semantic neighborhoods that incorporate taxonomic, functional, and associative knowledge.

For example, a candidate concept such as *Cricket* may not only relate to directly co-occurring

lexical terms, but may also connect to semantically broader concepts such as *Sport*, *Match*, *Wicket*, *India*, and *Sachin Tendulkar*. Figure 4.3 illustrates this structured semantic expansion, demonstrating how ConceptNet transforms a socially contextualized term into a richer conceptual network.

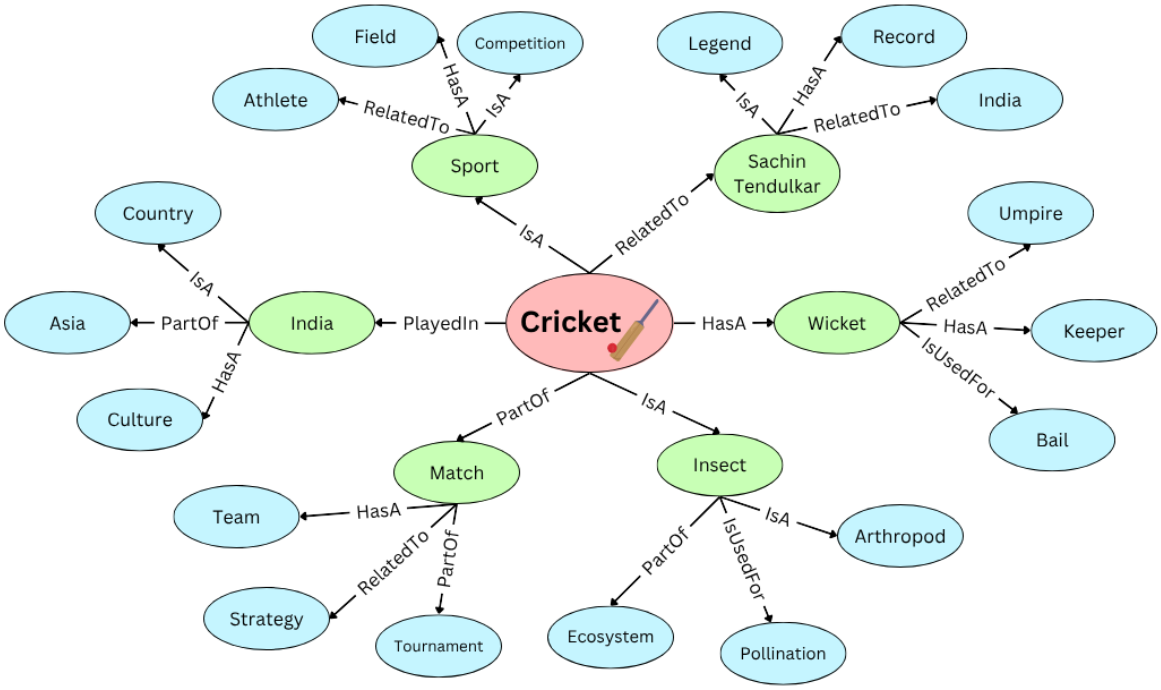

Figure 4.3: Illustrative ConceptNet semantic neighborhood for the keyword *Cricket*, demonstrating semantically typed commonsense relations across associated concepts

This form of semantic expansion is methodologically significant because it allows COM-Key to move beyond isolated lexical similarity and instead model candidate concepts within a broader commonsense-aware conceptual topology. However, because real-world semantic relationships are rarely binary and often exhibit gradual or uncertain association, COM-Key further incorporates fuzzy graph theory to represent conceptual relationships in graded form. Rather than simply determining whether concepts are connected, fuzzy graph modeling captures the degree and resilience of conceptual association.

Accordingly, the commonsense-enriched fuzzy graph framework integrates three complementary stages:

1. Semantic expansion through ConceptNet-based conceptual neighborhoods,
2. Structural edge weighting through graph-theoretic connectivity,
3. Fuzzy membership estimation through contextual semantic similarity.

Formally, the fuzzy graph is defined as $G = (V, E, \mu)$, where $V = \{c_1, c_2, \ldots, c_k\}$ denotes the set of centroid-based candidate concepts obtained from semantic clustering, E represents weighted conceptual relationships, and $\mu : V \to [0, 1]$ denotes the fuzzy membership function representing contextual semantic centrality.

4.2.2.1 Semantic Expansion via ConceptNet Neighborhoods

To enrich each candidate centroid $c_i \in V$ with structured external semantics, a local semantic neighborhood is constructed by traversing ConceptNet up to a specified depth d. This process yields a relational neighborhood $R(c_i) = \{r_1, r_2, \ldots, r_m\}$, where each related concept r_j is connected to c_i through semantically typed relations belonging to Eq. 4.5.

$$\mathscr{R} \subseteq \{\text{IsA}, \text{PartOf}, \text{UsedFor}, \text{HasA}, \text{RelatedTo}, \ldots\} \tag{4.5}$$

The induced ConceptNet subgraph is therefore represented as $G_i = (R(c_i), E_i)$.

This expansion transforms each semantically clustered candidate into a broader conceptual neighborhood, thereby embedding candidate keywords within curated commonsense structures rather than relying solely on distributional language patterns.

4.2.2.2 Edge Weighting via Structural Graph Connectivity

Following semantic expansion, conceptual relatedness between any two candidate centroids $c_i, c_j \in V$ is modeled by combining their respective ConceptNet-induced subgraphs using Eq. 4.6.

$$G_{ij} = G_i \cup G_j = (V_{ij}, E_{ij}) \tag{4.6}$$

where: $V_{ij} = R(c_i) \cup R(c_j), \quad E_{ij} = E_i \cup E_j$.

To quantify the structural robustness of this conceptual association, COM-Key defines fuzzy edge weight using edge connectivity as in Eq. 4.7.

$$w(c_i, c_j) := \lambda(G_{ij}) = \min\{|E'| : G_{ij} \setminus E' \text{ is disconnected}\} \tag{4.7}$$

where $E' \subseteq E_{ij}$ denotes the minimum edge cut required to disconnect the union graph.

This formulation captures semantic resilience by measuring the extent to which multiple redundant conceptual pathways connect two candidate concepts. Stronger edge connectivity

therefore reflects richer and more structurally stable semantic association, providing a graph-theoretically grounded alternative to simplistic binary adjacency or direct similarity metrics.

4.2.2.3 Fuzzy Membership Estimation via Contextual Embedding Similarity

While edge connectivity models structural conceptual strength, node-level semantic importance is quantified through fuzzy membership estimation. Let $e(c_i) \in \mathbb{R}^d$ denote the contextual BERT embedding of centroid concept c_i. The fuzzy membership value of each node is defined as the average pairwise cosine similarity between c_i and all other centroid concepts

$$\mu(c_i) = \frac{1}{|V|-1} \sum_{\substack{j=1 \\ j \neq i}}^{|V|} \frac{e(c_i) \cdot e(c_j)}{\|e(c_i)\| \|e(c_j)\|} \tag{4.8}$$

This measure quantifies the degree to which each concept is contextually integrated within the broader semantic space. Concepts with higher membership values demonstrate stronger global semantic representativeness, thereby balancing local structural semantics with broader contextual coherence.

Overall, by integrating contextual embeddings, structured commonsense knowledge, structural graph connectivity, and fuzzy membership estimation, COM-Key constructs a semantically enriched fuzzy graph capable of representing social media concepts as graded, uncertain, and context-sensitive entities. This conceptual-semantic graph subsequently forms the structural basis for fuzzy centrality-based importance modeling, where candidate keyword salience is assessed through multiple graph-theoretic centrality perspectives.

4.2.3 Valuation of Node Importance via Fuzzy Centrality Modeling

Following the construction of the commonsense-enriched fuzzy graph $G = (V, E, \mu)$, the next stage focuses on evaluating the relative importance of candidate concepts within this semantic structure. While the preceding stage establishes conceptual relationships through contextual embeddings, commonsense expansion, and fuzzy graph modeling, effective keyword extraction additionally requires identifying which candidate nodes exhibit the greatest semantic and structural significance. In social media environments, lexical frequency alone is often insufficient, as keyword relevance frequently depends on broader factors such as semantic accessibility, contextual coherence, and conceptual influence. Therefore, candidate valuation must reflect the topology of the enriched semantic graph rather than surface occurrence alone [144].

Conventional centrality measures are typically designed for crisp graph structures in which relationships are binary. In contrast, COM-Key operates on a fuzzy graph, where node importance and conceptual relationships are represented through graded semantic values. This necessitates fuzzy centrality formulations capable of preserving uncertainty and structural nuance. Accordingly, COM-Key employs multiple complementary fuzzy centrality measures to capture diverse dimensions of candidate importance rather than relying on a single structural perspective.

Accordingly, five fuzzy centrality measures are incorporated:

1. Fuzzy Closeness Centrality,
2. Fuzzy Eigenvector Centrality,
3. Fuzzy Katz Centrality,
4. Fuzzy Harmonic Centrality,
5. Fuzzy PageRank Centrality.

Each measure contributes a distinct structural perspective, enabling a more comprehensive assessment of candidate importance across the fuzzy semantic network.

4.2.3.1 Fuzzy Closeness Centrality

Fuzzy closeness centrality [145] evaluates how proximally a candidate concept is positioned relative to all other concepts in the fuzzy graph. Nodes that maintain shorter average fuzzy shortest-path distances are considered more centrally accessible within the conceptual structure. This measure is defined as Eq. 4.9.

$$C_{\text{close}}(c_i) = \frac{|V| - 1}{\sum_{j \neq i} d_f(c_i, c_j)} \tag{4.9}$$

where $|V|$ denotes the number of candidate concepts and $d_f(c_i, c_j)$ represents the fuzzy shortest-path distance between nodes c_i and c_j. Higher closeness values indicate stronger conceptual accessibility across the graph.

4.2.3.2 Fuzzy Eigenvector Centrality

While closeness emphasizes accessibility, fuzzy eigenvector centrality captures conceptual influence by assigning greater importance to nodes connected to other highly important nodes.

Thus, a concept derives significance not solely from the number of connections it possesses, but from the prominence of its conceptual neighbors. This recursive formulation is expressed as Eq. 4.10.

$$C_{\text{eig}}(c_i) = \frac{1}{\lambda} \sum_{j \in V} w(c_i, c_j) C_{\text{eig}}(c_j) \tag{4.10}$$

where $w(c_i, c_j)$ denotes fuzzy edge weight and λ is the principal eigenvalue. This measure identifies concepts embedded within structurally influential semantic neighborhoods.

4.2.3.3 Fuzzy Katz Centrality

Fuzzy Katz centrality [146] extends eigenvector centrality by incorporating both direct and indirect semantic pathways, thereby accounting for broader conceptual propagation across the graph. This is particularly useful in sparse short-text environments, where concept relevance may emerge through extended semantic pathways rather than immediate adjacency. The formulation is given by Eq. 4.11.

$$C_{\text{katz}}(c_i) = \sum_{k=1}^{\infty} \sum_{j \in V} \alpha^k P_k(c_i, c_j) \tag{4.11}$$

where α is the attenuation factor $(0 < \alpha < 1)$ and $P_k(c_i, c_j)$ denotes the weighted number of fuzzy paths of length k between concepts c_i and c_j. This formulation captures semantic diffusion through both local and extended conceptual structures.

4.2.3.4 Fuzzy Harmonic Centrality

Because fuzzy conceptual graphs may contain weakly connected or structurally sparse regions, harmonic centrality provides a more stable alternative to closeness by avoiding instability associated with unreachable nodes. It is defined as Eq. 4.12.

$$C_{\text{harm}}(c_i) = \sum_{\substack{j \neq i \\ d_f(c_i, c_j) \neq 0}} \frac{1}{d_f(c_i, c_j)} \tag{4.12}$$

This measure rewards candidate concepts that maintain broad structural reachability, even when portions of the graph are only weakly connected.

4.2.3.5 Fuzzy PageRank Centrality

Fuzzy PageRank centrality models conceptual importance through recursive semantic flow across the fuzzy graph. Concepts receiving consistent structural influence from other important nodes are assigned higher importance scores. Adapted from the original PageRank framework [147], the formulation is Eq. 4.13.

$$C_{\mathrm{pr}}(c_i) = (1-\beta) + \beta \sum_{j \in V} \frac{w(c_j, c_i)}{\sum_{k \in V} w(c_j, c_k)} C_{\mathrm{pr}}(c_j) \tag{4.13}$$

where $\beta \in (0,1)$ denotes the damping factor controlling recursive propagation. This measure captures the diffusion of semantic relevance across interconnected conceptual structures.

Collectively, these five fuzzy centrality measures provide a multidimensional valuation framework that captures structural accessibility, conceptual influence, semantic propagation, reachability, and recursive prominence. This multi-centrality strategy is methodologically valuable because no single centrality measure can fully characterize keyword importance within socially dynamic and semantically uncertain environments.

By integrating these complementary perspectives, COM-Key constructs a more balanced and structurally grounded importance space for candidate concepts. These fuzzy centrality valuations subsequently serve as the decision matrix for the final ranking stage, where multi-criteria optimization is applied to identify the most semantically significant keywords.

4.2.4 Multi-Criteria Keyword Ranking via WASPAS Aggregation of Fuzzy Centralities

Following the valuation of candidate concepts through multiple fuzzy centrality measures, the final stage of COM-Key focuses on transforming these diverse structural assessments into a unified and interpretable keyword ranking. Although each fuzzy centrality measure captures an important dimension of conceptual significance, no individual metric alone can comprehensively represent the multidimensional nature of keyword importance within a commonsense-enriched fuzzy semantic graph. For instance, closeness centrality emphasizes semantic accessibility, eigenvector centrality captures conceptual prestige, Katz centrality models extended semantic propagation, harmonic centrality reflects broad structural reachability, and PageRank centrality evaluates recursive semantic influence. Because these measures encode complemen-

tary rather than redundant perspectives, direct reliance on a single criterion may bias keyword selection toward only one structural property while overlooking other meaningful dimensions of semantic relevance. Consequently, a principled aggregation mechanism is required to synthesize these diverse semantic signals into a balanced final ranking [148].

To address this challenge, COM-Key adopts the Weighted Aggregated Sum Product Assessment (WASPAS) method, a multi-criteria decision-making (MCDM) framework that integrates additive and multiplicative evaluation strategies within a unified ranking structure. WASPAS is particularly suitable in this context because it combines the interpretability of weighted summation with the consistency sensitivity of weighted product modeling. As a result, candidate keywords are evaluated not only according to their cumulative performance across multiple criteria, but also according to the degree to which they maintain consistently strong performance across all semantic dimensions.

Let $K = \{k_1, k_2, \ldots, k_n\}$, denote the set of candidate keywords corresponding to the nodes of the fuzzy graph, and let $\mathscr{F} = \{F_1, F_2, \ldots, F_m\}$, represent the set of fuzzy centrality criteria, where $m = 5$ in the COM-Key framework. For each candidate keyword k_i, its score under criterion F_j is denoted by x_{ij}.

Since the five fuzzy centrality measures may operate on different numerical scales, normalization is first performed to ensure cross-criterion comparability as in Eq. 4.14.

$$\hat{x}_{ij} = \frac{x_{ij}}{\max_k x_{kj}} \tag{4.14}$$

where $\hat{x}_{ij}$ denotes the normalized score of keyword k_i under criterion F_j.

The final WASPAS score for keyword k_i is then computed as Eq. 4.15 [148].

$$Q_i = \lambda \sum_{j=1}^{m} w_j \hat{x}_{ij} + (1-\lambda) \prod_{j=1}^{m} \left(\hat{x}_{ij}\right)^{w_j} \tag{4.15}$$

where:

- w_j denotes the weight assigned to criterion F_j,
- $\lambda \in [0,1]$ controls the balance between additive and multiplicative aggregation,
- $\sum_{j=1}^{m} w_j = 1$.

In this formulation, the first term represents the Weighted Sum Model (WSM), which captures

average aggregated semantic performance across all criteria, while the second term represents the Weighted Product Model (WPM), which rewards candidates demonstrating consistently strong performance across all semantic dimensions. In COM-Key, $\lambda = 0.5$ is typically adopted to ensure balanced emphasis between these complementary aggregation principles.

The use of WASPAS is methodologically significant because it transforms fuzzy centrality measures from isolated structural indicators into an integrated semantic decision space. Rather than privileging one graph property over another, WASPAS allows COM-Key to evaluate candidate concepts through a holistic multi-criteria framework that simultaneously accounts for accessibility, influence, diffusion, reachability, and recursive prominence. This is particularly important in social media environments, where semantic importance is inherently multidimensional and where structurally noisy discourse may otherwise distort singular ranking strategies.

Furthermore, the multiplicative component of WASPAS introduces robustness by penalizing candidates that perform strongly in one dimension but weakly across others. This characteristic is especially beneficial in short-text environments, where semantically meaningful keywords should ideally demonstrate broad conceptual significance rather than isolated structural prominence.

Once Q_i is computed for all candidate keywords, concepts are ranked in descending order as $Q_{(1)} \geq Q_{(2)} \geq \cdots \geq Q_{(n)}$.

The highest-ranked concepts are then selected as the final extracted keywords.

Overall, the WASPAS-based aggregation stage serves as the final decision-making mechanism of COM-Key, synthesizing the semantic richness of fuzzy centrality modeling into a unified, interpretable, and structurally balanced keyword ranking. By integrating multiple dimensions of conceptual importance into a single optimization framework, this stage enhances both the theoretical rigor and practical robustness of keyword extraction in semantically sparse and socially dynamic environments.

4.2.5 Algorithmic Summary of COM-Key

For clarity, Algorithm 4 summarizes the complete workflow of the proposed COM-Key methodology. The process begins with semantically normalized microtext input, followed by contextual embedding generation and semantic clustering. Subsequently, commonsense-enriched fuzzy graph construction and fuzzy centrality-based importance modeling are performed to capture semantic relevance and contextual relationships. Finally, the WASPAS-based multi-

criteria ranking mechanism is applied to identify and extract the most significant keywords.

Algorithm 4 Stage-wise Schematic of the Proposed COM-Key Framework

Input: Preprocessed microtext corpus $\mathscr{D} = \{d_1, d_2, \ldots, d_n\}$
Output: Ranked keyword list $\mathscr{K}^*$
Stage 1: Semantic Embedding $\rightarrow$ Clustering
Each token $w_i \in \mathscr{D}$ is encoded using BERT: $\mathbf{x}_i \leftarrow \text{BERT}(w_i)$
DBSCAN is applied on $\mathscr{X} = \{\mathbf{x}_i\}$ to form clusters $\mathscr{C} = \{C_1, \ldots, C_k\}$
Cluster centroids are computed as $c_j \leftarrow \frac{1}{|C_j|} \sum_{\mathbf{x} \in C_j} \mathbf{x}$ for each $C_j \in \mathscr{C}$
Stage 2: Fuzzy Graph Construction $\rightarrow$ Commonsense Enrichment
A fuzzy graph $G = (V, E, \mu)$ is initialized with $V \leftarrow \{c_j\}$ and $E = \emptyset$
For each pair (c_i, c_j), semantic membership is $\mu_{ij} \leftarrow \cos(c_i, c_j)$
Commonsense relatedness is obtained as $r_{ij} \leftarrow \text{ConceptNet}(c_i, c_j)$
Fuzzy edge weight is computed: $w(c_i, c_j) \leftarrow r_{ij} \cdot \mu_{ij}$
Stage 3: Fuzzy Centrality-Based Importance Modeling
For each node $c_i \in V$, compute centrality scores:
$$C_{\text{close}}(c_i), C_{\text{eig}}(c_i), C_{\text{katz}}(c_i), C_{\text{harm}}(c_i), C_{\text{pr}}(c_i)$$
Stage 4: Multi-Criteria Aggregation $\rightarrow$ Keyword Ranking
Normalize centralities as $x_{ij} \leftarrow \frac{C_j(c_i)}{\max_k C_j(c_k)}$
Compute WASPAS score as $Q_i \leftarrow \lambda \sum_j w_j x_{ij} + (1-\lambda) \prod_j x_{ij}^{w_j}$
Return $\mathscr{K}^*$ by sorting nodes in descending order of Q_i

4.3 Experimental Evaluation and Analysis of COM-Key

This section presents the empirical evaluation of the proposed keyword extraction framework. It includes a description of the datasets used, the evaluation metrics, and the experimental protocol. The performance of the method is then analyzed through comparative evaluation with baseline approaches, followed by a discussion of the results.

4.3.1 Datasets

The evaluation is conducted using three publicly available Twitter datasets that differ in size and domain, allowing assessment of the proposed method under varied microtext conditions.

- **IPL 2019 Dataset:** A collection of 1,502 tweets related to the Indian Premier League (IPL) 2019 season, representing a small, domain-specific dataset focused on sports discussions.
- **Elections 2019 Dataset:** A medium-scale dataset comprising 46,040 tweets associated with the 2019 Indian general elections, containing diverse political content along with

metadata such as timestamps, user information, and engagement indicators.

- **GOT 8 Dataset:** A large-scale dataset consisting of 760,626 tweets related to the eighth season of Game of Thrones, reflecting high lexical variability and noise typical of social media data.

These datasets provide a range of evaluation settings, from focused event-based discussions to large-scale trending topics.

All datasets were preprocessed using the pipeline described in the previous section, including normalization, token filtering, and linguistic processing to obtain a consistent representation for keyword extraction.

4.3.2 Evaluation Protocol

The evaluation of the proposed COM-Key framework is conducted using both qualitative and quantitative measures. A gold standard of reference keywords is constructed through human annotations, and annotation consistency is assessed using inter-annotator agreement. The extraction performance is then evaluated using standard information retrieval metrics.

4.3.2.1 Gold Standard Annotations

To establish a reference for evaluation, keywords are manually annotated by three independent annotators. Each annotator selects a set of salient keywords for each dataset. All annotations are normalized to lowercase to ensure consistency. Table 4.1 presents the aggregated keyword sets across datasets.

4.3.2.2 Inter-Annotator Agreement

To assess the consistency of manual annotations, inter-annotator agreement is measured using the Sørensen–Dice coefficient [149].

Given two keyword sets K_i and K_j, the agreement is computed as Eq. 4.16.

$$\text{Dice}(K_i, K_j) = \frac{2 \cdot |K_i \cap K_j|}{|K_i| + |K_j|} \tag{4.16}$$

Pairwise agreement is computed across annotators for each dataset. This provides an indication of the reliability of the annotated keyword sets.

Table 4.1: Keywords identified by three independent annotators across the three datasets

Annotator	Dataset	Annotated Keywords
Annotator 1	IPL2019	dhoni, CSK, chennai, kkr, rcb, Indian, Cricbuzz, ipl, srh, Scored, India, vivoIPL, cricket, season, kohli, dream11, kkrider, IPLDhamakaa, playoffs
	Elections2019	election, Gandhi, bjp, modi, congress, terrorist, delhi, PM, government, cell, voting, china, indian, pulwama, pakistan, rahul, russia, democratic, tmc, rss, attack, chowkidar
	GOT8	GOT, DRAGONGLASS, GAME, THRONES, STARK, RAVVER, SOPHIE, HARINGTON, WINTERFELL, CERSEI, TARGARYEN, LANNISTER, DINKLAGE, DAENERYS, HBO, WESTEROS, SANSA'S
Annotator 2	IPL2019	Indian, ipl, Kohli, dhoni, CSK, match, win, game, wicket, season, cricket, cricbuzz, live, bowlers, bat, score, six, Thala, century, runs, predict
	Elections2019	modi, India, RSS, government, politics, Congress, country, youth, nation, opposition, votes, national, elections, PM, BJP, Police, soldier, Gandhi, pakistan, campaign, president
	GOT8	Game, daenerys, theories, spoiler, thrones, snow, dragon, season, watch, premier, best, worst, walkers, stark, amazing, released, star, actor, finished, episode, kill, HBO
Annotator 3	IPL2019	Indian, ipl, player, season, kohli, dhoni, KKR, Chennai, cricbuzz, Will, RCB, cricket, jadeja, team, watson, match, run, jadhav, cricbuzzLIVE, SRH
	Elections2019	bjp, congress, election, govt, former, gandhi, india, modi, special, win, unemployment, undemocratic, delhi, tukde, surgical, rahulgandhi, economy, opposition, pakistan
	GOT8	throne, fantasy, stark, got, Daenerys, watch, agree, podcasted, game, premier, clarke, fans, dragon, love, episode, watch, season, archery, harington, flicks

Figure 4.4 presents the pairwise Sørensen–Dice agreement scores between annotators across the evaluated datasets. The observed agreement values indicate a substantial level of consistency among annotators, suggesting that the manually assigned keyword sets are reliable and suitable for subsequent evaluation. High inter-annotator agreement further supports the validity of the annotation process and reduces the likelihood that experimental outcomes are influenced by annotation inconsistency.

4.3.3 Evaluation Metrics

The performance of keyword extraction is evaluated using standard information retrieval metrics computed at different values of n (top-n keywords), specifically $n = 3, 5, 10$. The definitions of the evaluation metrics are summarized in Table 4.2.

Table 4.2: Evaluation metrics used for assessing keyword extraction performance

Metric	Definition	Formula
Precision	Measures the proportion of predicted keywords that are relevant	$\frac{\lvert P \cap A \rvert}{\lvert P \rvert}$
Recall	Measures the proportion of relevant keywords that are successfully retrieved	$\frac{\lvert P \cap A \rvert}{\lvert A \rvert}$
F1-score	Harmonic mean of Precision and Recall, balancing accuracy and completeness	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

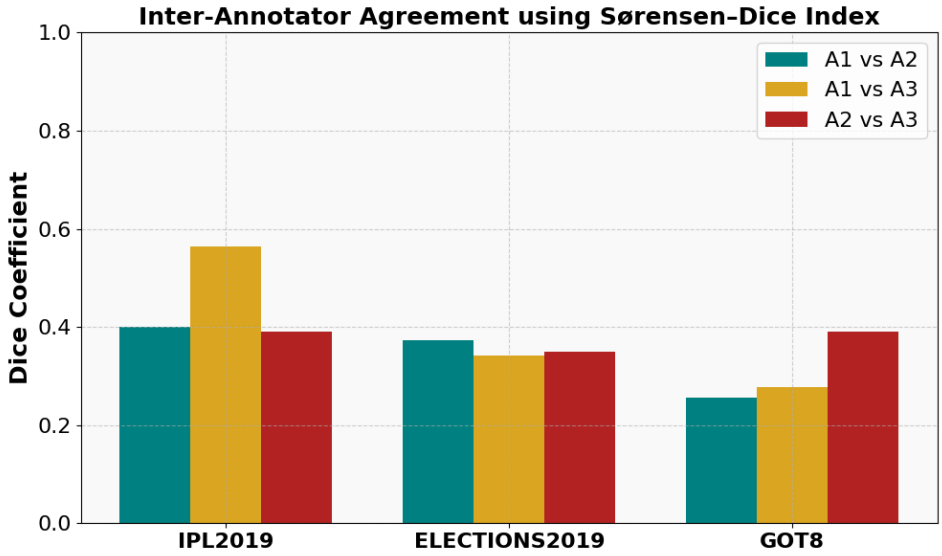

Figure 4.4: Pairwise Sørensen-Dice coefficients between annotators across datasets

In Table 4.2, P denotes the set of predicted keywords and A represents the set of annotated keywords.

4.3.4 Results and Discussion

The performance of the proposed COM-Key framework is evaluated across three datasets at different values of n (top-n keywords). The results, aggregated across annotators, are presented in Table 4.3.

Table 4.3: Evaluation metrics at top-$n = 3, 5, 10$ across datasets and annotators

Reader	**@3 (Pr, Re, F1)**	**@5 (Pr, Re, F1)**	**@10 (Pr, Re, F1)**
		IPL 2019	
Reader 1	(1.00, 0.158, 0.273)	(0.80, 0.211, 0.333)	(0.50, 0.263, 0.345)
Reader 2	(1.00, 0.143, 0.250)	(1.00, 0.238, 0.385)	(0.80, 0.381, 0.516)
Reader 3	(1.00, 0.150, 0.261)	(0.80, 0.200, 0.320)	(0.70, 0.350, 0.467)
		Elections 2019	
Reader 1	(1.00, 0.136, 0.240)	(1.00, 0.227, 0.370)	(0.80, 0.364, 0.500)
Reader 2	(1.00, 0.143, 0.250)	(1.00, 0.238, 0.385)	(0.70, 0.333, 0.452)
Reader 3	(1.00, 0.158, 0.273)	(0.80, 0.211, 0.333)	(0.70, 0.368, 0.483)
		GOT 8	
Reader 1	(1.00, 0.177, 0.300)	(0.80, 0.235, 0.364)	(0.50, 0.294, 0.370)
Reader 2	(1.00, 0.130, 0.231)	(0.80, 0.174, 0.286)	(0.50, 0.217, 0.303)
Reader 3	(1.00, 0.158, 0.273)	(0.60, 0.158, 0.250)	(0.30, 0.158, 0.207)

4.3.4.1 Performance Analysis

As observed in Table 4.3, the proposed method achieves perfect precision at $n = 3$ across all datasets and annotators, indicating that the top-ranked keywords are consistently relevant.

Recall increases with higher values of n, reflecting improved coverage of annotated keywords. Correspondingly, the F1-score improves from $n = 3$ to $n = 5$, indicating a better balance between precision and recall. Beyond $n = 5$, the improvement in F1-score becomes marginal, suggesting diminishing gains from additional keywords.

The method demonstrates consistent behavior across datasets. The IPL and Elections datasets show stable improvements with increasing n, whereas the GOT 8 dataset exhibits relatively lower gains, reflecting its higher variability and noise.

4.3.4.2 Discussion

The results highlight the trade-off between precision and recall in keyword extraction. Smaller values of n yield highly precise results with limited coverage, while larger values improve recall at the cost of reduced precision.

Figure 4.5 illustrates the variation of precision, recall, and F1-score with increasing values of n, averaged across annotators.

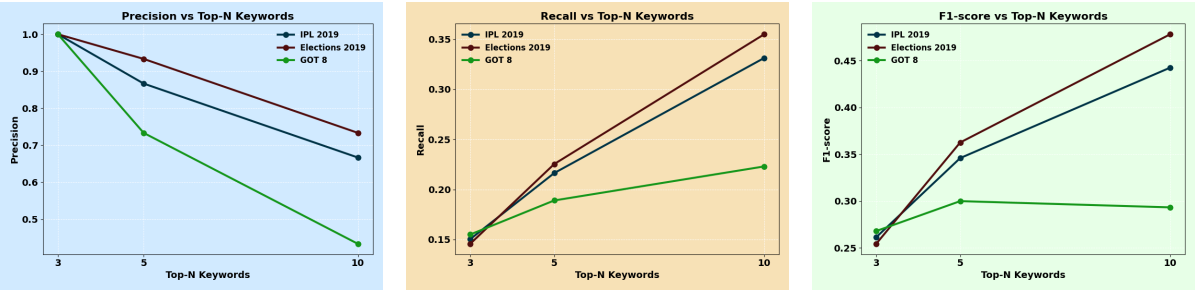

Figure 4.5: Trend analysis of Precision, Recall, and F1-score across datasets with increasing number of extracted keywords

The trends indicate that $n = 5$ provides a balanced performance across datasets, achieving a suitable compromise between relevance and coverage. Structured datasets such as IPL and Elections benefit from increasing n, while the GOT 8 dataset shows reduced gains beyond moderate keyword counts due to its informal and diverse nature.

Overall, the results demonstrate that the proposed framework effectively extracts relevant keywords across different social media contexts, maintaining consistent performance across datasets of varying scale and complexity.

4.3.5 Comparative Analysis

The performance of the proposed COM-Key framework is compared with several keyword extraction methods, including TF-IDF [138], TextRank [14], TopicRank [47], RAKE [39], MultipartiteRank [129], YAKE [38], TKG [80], NE Rank [81], KEGBA [83], SwiftRank [88], KECNW [49], and FLAKE [86].

The comparison is conducted using Precision, Recall, and F1-score for the top-10 extracted keywords in Table 4.4 and Table 4.5 for IPL 2019 dataset and Elections 2019 dataset respectively, evaluated against annotations from three readers under a consistent experimental setup.

Table 4.4: Comparison of keyword extraction methods on IPL 2019 dataset across three readers

Methods	**Reader 1**			**Reader 2**			**Reader 3**		
	P	R	F1	P	R	F1	P	R	F1
TF-IDF	0.3	0.1579	0.2069	0.6	0.2857	0.3871	0.5	0.25	0.3333
TextRank	0.4	0.2105	0.2759	0.6	0.2857	0.3871	0.4	0.2	0.2667
TopicRank	0.2	0.1053	0.1379	0.4	0.1905	0.2581	0.5	0.25	0.3333
RAKE	0.2	0.1053	0.1379	0.2	0.0952	0.1290	0.5	0.25	0.3333
MultipartiteRank	0.2	0.1053	0.1379	0.4	0.1905	0.2581	0.5	0.25	0.3333
YAKE	0.4	0.2105	0.2759	0.6	0.2857	0.3871	0.6	0.3	0.4
TKG	0.3	0.1579	0.2069	0.6	0.2857	0.3871	0.5	0.25	0.3333
NE Rank	0.4	0.2105	0.2759	0.5	0.2381	0.3226	0.4	0.2	0.2667
KEGBA	0.1	0.0526	0.0690	0.2	0.0952	0.1290	0.3	0.15	0.2
SwiftRank	0.4	0.2105	0.2759	0.5	0.2381	0.3226	0.6	0.3	0.4
KECNW	0.4	0.2105	0.2759	0.5	0.2381	0.3226	0.4	0.2	0.2667
FLAKE	0.2	0.1053	0.1379	0.4	0.1905	0.2581	0.5	0.25	0.3333
COM-Key	**0.5**	**0.2632**	**0.3448**	**0.8**	**0.3810**	**0.5161**	**0.7**	**0.35**	**0.4667**

As observed in Tables 4.4 and 4.5, the proposed COM-Key framework consistently achieves higher F1-scores across all readers compared to the baseline methods. The improvement is primarily due to better recall, indicating enhanced coverage of relevant keywords. Traditional frequency-based methods show comparatively lower performance, while graph-based approaches provide moderate improvements but remain limited in capturing deeper semantic relationships. The proposed method benefits from incorporating semantic clustering, commonsense knowledge, and fuzzy graph modeling, leading to more effective keyword extraction across datasets.

Table 4.5: Comparison of keyword extraction methods on Elections 2019 dataset across three readers

Methods	Reader 1			Reader 2			Reader 3		
	P	R	F1	P	R	F1	P	R	F1
TF-IDF	0.5	0.2273	0.3125	0.6	0.2857	0.3871	0.5	0.2632	0.3448
TextRank	0.5	0.2273	0.3125	0.5	0.2381	0.3226	0.4	0.2105	0.2759
TopicRank	0.3	0.1364	0.1875	0.4	0.1905	0.2581	0.4	0.2105	0.2759
RAKE	0.1	0.0455	0.0625	0.1	0.0476	0.0645	0.0	0.0	0.0
MultipartiteRank	0.3	0.1364	0.1875	0.4	0.1905	0.2581	0.4	0.2105	0.2759
YAKE	0.7	0.3182	0.4375	0.6	0.2857	0.3871	0.6	0.3158	0.4138
TKG	0.3	0.1364	0.1875	0.2	0.0952	0.1290	0.2	0.1053	0.1379
NE Rank	0.5	0.2273	0.3125	0.6	0.2857	0.3871	0.5	0.2632	0.3448
KEGBA	0.2	0.0909	0.1250	0.2	0.0952	0.1290	0.2	0.1053	0.1379
SwiftRank	0.5	0.2273	0.3125	0.6	0.2857	0.3871	0.6	0.3158	0.4138
KECNW	0.5	0.2273	0.3125	0.6	0.2857	0.3871	0.5	0.2632	0.3448
FLAKE	0.2	0.0909	0.1250	0.3	0.1429	0.1935	0.3	0.1579	0.2069
COM-Key	**0.8**	**0.3636**	**0.5**	**0.7**	**0.3333**	**0.4516**	**0.7**	**0.3684**	**0.4828**

4.3.6 Statistical Performance Analysis

To assess the statistical significance of the observed performance improvements, inferential statistical tests are conducted using the *paired t-test* [150] and *one-way Analysis of Variance (ANOVA)* [137], followed by *Tukey's Honestly Significant Difference (HSD)* post-hoc analysis [136].

4.3.6.1 Paired t-Test

The paired t-test is applied to evaluate whether the improvement in F1-score of the proposed method over baseline approaches is statistically significant. The hypotheses are defined as Eq. 4.17.

$$\begin{aligned} H_0 &: \mu_d = 0 \\ H_A &: \mu_d > 0 \end{aligned} \tag{4.17}$$

The test statistic is computed as Eq. 4.18.

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \tag{4.18}$$

where $\bar{d}$ is the mean difference in F1-scores, s_d is the standard deviation, and n is the number of samples.

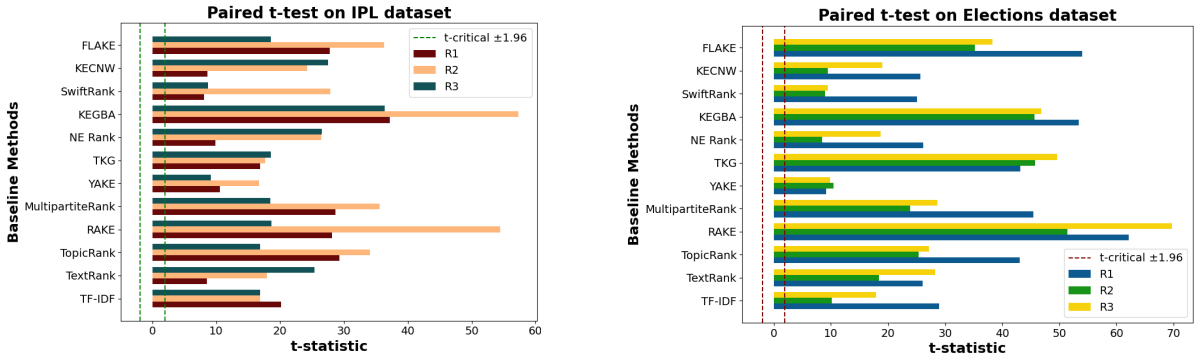

Figure 4.6: Paired t-test results on F1-scores for IPL and Elections datasets

As observed in Figure 4.6, the computed t-values exceed the critical threshold at the 95% confidence level in most comparisons, indicating that the performance improvements are statistically significant.

4.3.6.2 ANOVA Analysis

To evaluate differences across all methods simultaneously, one-way ANOVA is conducted on F1-scores. The hypotheses are defined as Eq. 4.19.

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \cdots = \mu_k \\ H_A &: \exists i, j \text{ such that } \mu_i \neq \mu_j \end{aligned} \tag{4.19}$$

The F-statistic is computed as Eq. 4.20.

$$F = \frac{MS_{\text{between}}}{MS_{\text{within}}} \tag{4.20}$$

The results yield a high F-statistic with p-value < 0.000001, indicating significant differences among the compared methods.

4.3.6.3 Tukey HSD Test

To identify pairwise differences between methods, Tukey's HSD test is applied. The test statistic is computed as Eq. 4.21.

$$Q = \frac{|\bar{x}_i - \bar{x}_j|}{\sqrt{\frac{MS_{\text{within}}}{n}}} \tag{4.21}$$

As shown in Figure 4.7, the proposed method achieves higher median F1-scores with lower

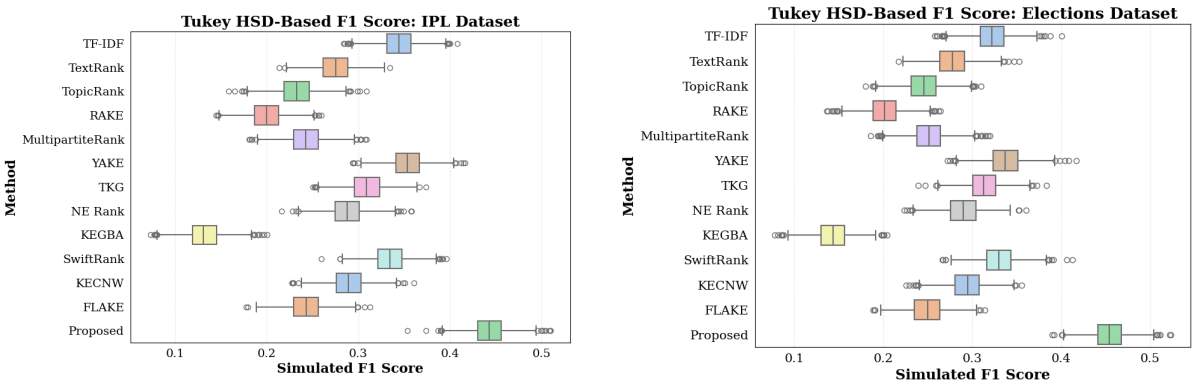

Figure 4.7: Tukey HSD-based boxplot analysis of F1-scores for IPL and Elections datasets

variance compared to baseline methods. The results confirm that the performance improvements are statistically significant and consistent across datasets.

4.4 QuToX: A Quantum-Inspired Game-Theoretic Computational Framework for Keyword Extraction and Interpretable Toxicity Detection in Short-Text Data

QuToX formulates keyword extraction in short-text environments as a strategic interaction process among candidate terms, modeled using principles from quantum game theory (QGT). Unlike classical game-theoretic formulations, where players select discrete strategies independently, QGT represents strategies within a quantum state space, enabling the modeling of complex and interdependent interactions among elements [151, 152]. This is particularly relevant for social media text, where word importance is often context-dependent and influenced by subtle semantic relationships.

Within this framework, each word in a tweet is treated as a strategic entity whose importance emerges through interaction with other words. Two key quantum properties enable this modeling. First, superposition allows a word to simultaneously represent multiple potential importance states rather than committing to a single deterministic role. Second, entanglement captures interdependence between words, enabling strongly related terms to influence each other's significance in a non-local and context-sensitive manner.

To operationalize this formulation, contextual semantic similarity between words is first derived using BERT-based embeddings, which are then used to construct a payoff matrix encoding pairwise interactions. This payoff structure governs the strategic behavior of candidate

terms within a quantum-inspired simulation process. Through repeated interaction and measurement, the system converges toward stable configurations in which highly relevant keywords emerge with higher selection probability.

The overall workflow of the QuToX keyword extraction process is illustrated in Figure 4.8. It comprises semantic similarity modeling, payoff matrix construction, and quantum-inspired interaction simulation, culminating in the extraction of contextually significant keywords.

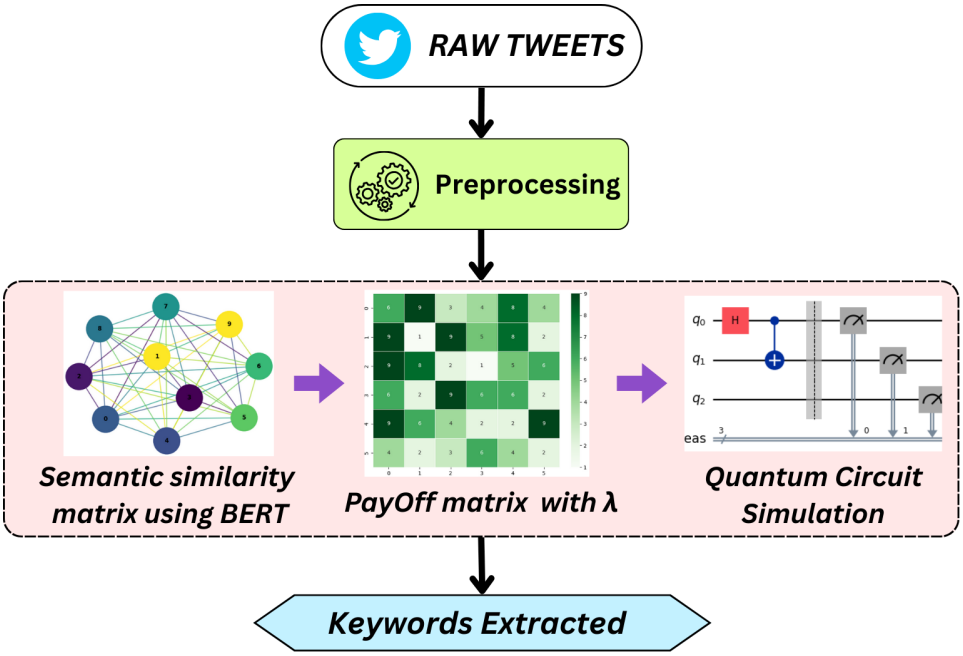

Figure 4.8: Workflow of the QuToX keyword extraction framework

4.4.1 Semantic Representation using Contextual Word Embeddings

As established in the preceding section, social media text is first transformed into a normalized and linguistically consistent token space through standard and extended preprocessing procedures. The resulting tokens provide the input for semantic representation, where each word is mapped to a contextual vector space to enable subsequent interaction modeling.

Let the set of preprocessed tokens for the ith tweet be denoted as $\mathbf{W}^{(i)} = \{w_1^{(i)}, w_2^{(i)}, \ldots, w_{m_i}^{(i)}\}$.

To capture context-dependent semantics, each token is encoded using the pre-trained BERT-base model. Unlike static embedding methods, contextual embeddings account for the influence of surrounding words, which is particularly relevant in short-text environments where meaning is often determined by local context rather than global frequency patterns.

The token sequence is first processed using the BERT tokenizer, which decompose words into

subword units $T^{(i)} = \{t_1^{(i)}, t_2^{(i)}, \ldots, t_L^{(i)}\}, \quad L \geq m_i$.

The tokenized sequence is then passed through the BERT model to obtain contextual representations as in Eq. 4.22.

$$H^{(i)} = \left[\mathbf{h}_1^{(i)}, \mathbf{h}_2^{(i)}, \ldots, \mathbf{h}_L^{(i)}\right]^\top \in \mathbb{R}^{L \times d} \tag{4.22}$$

where $d = 768$ denotes the embedding dimensionality, and each $\mathbf{h}_k^{(i)}$ corresponds to the contextual embedding of the kth subword token.

Since a single word may correspond to multiple subword units, a unified representation is obtained through aggregation. For a word $w_j^{(i)}$ mapped to subwords $t_a^{(i)}, \ldots, t_b^{(i)}$, the corresponding embedding is computed as Eq. 4.23.

$$\mathbf{v}_j^{(i)} = \frac{1}{b-a+1} \sum_{k=a}^{b} \mathbf{h}_k^{(i)} \tag{4.23}$$

The semantic representation of the tweet is thus defined as Eq. 4.24.

$$V^{(i)} = \left[\mathbf{v}_1^{(i)}, \mathbf{v}_2^{(i)}, \ldots, \mathbf{v}_{m_i}^{(i)}\right]^\top \in \mathbb{R}^{m_i \times d} \tag{4.24}$$

This embedding matrix provides a structured representation of contextual semantics for all candidate tokens. It serves as the foundational input for the subsequent interaction modeling stage, where pairwise relationships between words are quantified and incorporated into a payoff-based formulation for keyword extraction.

4.4.2 Similarity and Interaction Modeling

Following the construction of contextual embeddings, pairwise semantic interactions among words are modeled to establish a structured interaction space for the subsequent quantum formulation. This stage converts contextual representations into a numerical structure that captures dependencies between candidate tokens.

Given the embedding matrix $V^{(i)}$, semantic similarity between any two words $w_p^{(i)}$ and $w_q^{(i)}$ is computed using cosine similarity Eq. 6.30.

$$S_{p,q}^{(i)} = \frac{\mathbf{v}_p^{(i)} \cdot \mathbf{v}_q^{(i)}}{\|\mathbf{v}_p^{(i)}\| \|\mathbf{v}_q^{(i)}\|} \tag{4.25}$$

This results in a similarity matrix $S^{(i)} \in \mathbb{R}^{m_i \times m_i}$, where each entry quantifies the contextual relationship between a pair of words.

To incorporate tweet-level contextual influence, a scalar parameter $\lambda \in [0,1]$ is introduced as a global interaction modifier. This parameter captures the overall contextual intensity of the tweet and is used to adjust local pairwise interactions. The interaction matrix is defined as Eq. 4.26.

$$I_{p,q}^{(i)} = S_{p,q}^{(i)} + 0.5 \cdot \lambda \tag{4.26}$$

This formulation integrates local semantic similarity with a global contextual signal, allowing interactions to reflect both word-level relationships and tweet-level characteristics.

To ensure numerical stability and comparability, the interaction values are normalized using the softmax function. For each word $w_p^{(i)}$, the normalized interaction with other words is computed as Eq. 4.27.

$$\text{softmax}_p(q) = \frac{e^{I_{p,q}^{(i)}}}{\sum_{k \neq p} e^{I_{p,k}^{(i)}}}, \quad q \neq p \tag{4.27}$$

This normalization converts interaction scores into a probability distribution, emphasizing relatively stronger relationships while suppressing weaker ones. Self-interactions are excluded, as the focus is on relational dependencies between distinct words.

The normalized interaction values are then scaled to obtain the payoff matrix Eq. 4.28.

$$P_{p,q}^{(i)} = \pi \cdot \text{softmax}_p(q), \quad q \neq p \tag{4.28}$$

where π serves as a scaling constant for mapping interaction strengths into a suitable range for quantum parameterization.

The resulting payoff matrix $P^{(i)}$ encodes normalized and contextually adjusted interaction strengths between words. This matrix provides the numerical foundation for the quantum game-theoretic modeling stage, where each element $P_{p,q}^{(i)}$ represents the influence of one word on another within the interaction space.

Through this transformation, contextual semantic relationships are systematically converted into a structured payoff representation, enabling the modeling of interdependent word impor-

tance within a unified framework.

4.4.3 Quantum Game-Theoretic Keyword Selection

Building upon the interaction structure defined in the previous stage, keyword extraction is formulated as a quantum game in which candidate words are modeled as interacting entities within a shared semantic space. This formulation enables the representation of interdependent word importance, allowing keyword selection to emerge from collective contextual interactions rather than independent evaluation.

The operational flow of the quantum keyword extraction process is illustrated in Figure 4.9. It depicts the sequence of steps involved in transforming the payoff matrix into a quantum state, followed by superposition, entanglement, parameterized evolution, and measurement-based keyword selection.

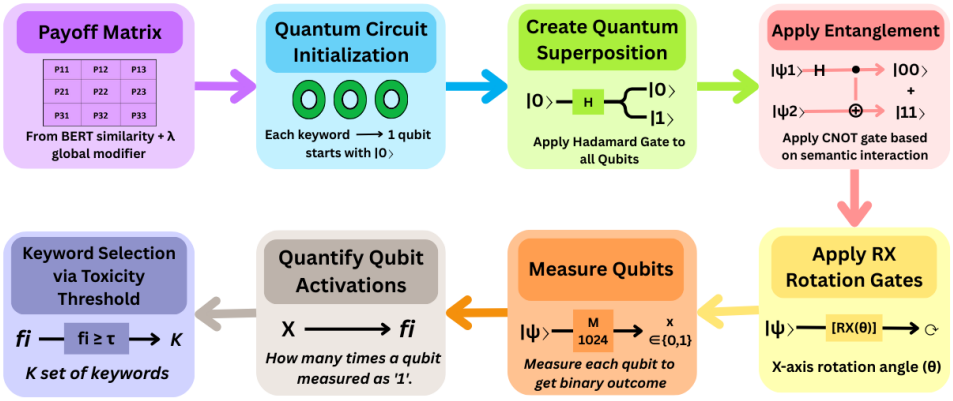

Figure 4.9: Flow of the quantum keyword extraction process from payoff-based state initialization to superposition, entanglement, parameterized evolution, and measurement-based keyword selection

For a given tweet $T^{(i)}$, let the corresponding token set be $W^{(i)} = \{w_1^{(i)}, \ldots, w_{n_i}^{(i)}\}$.

The quantum game is defined as Eq. 4.29.

$$\mathcal{G}_Q^{(i)} = (Q^{(i)}, \mathcal{H}^{(i)}, P^{(i)}, \mathcal{U}^{(i)}, \mathcal{M}^{(i)}) \tag{4.29}$$

where:

- $Q^{(i)} = \{q_1^{(i)}, \ldots, q_{n_i}^{(i)}\}$ denotes qubits corresponding to words,
- $\mathcal{H}^{(i)} = \mathbb{C}^{2^{n_i}}$ represents the Hilbert space of all possible keyword configurations,

- $P^{(i)}$ is the payoff matrix derived from semantic interactions,
- $\mathscr{U}^{(i)}$ represents the unitary evolution of the system,
- $\mathscr{M}^{(i)}$ denotes the measurement operation.

Each qubit $q_j^{(i)}$ encodes the selection state of word $w_j^{(i)}$, where $|0\rangle$ denotes non-selection and $|1\rangle$ denotes selection. Unlike classical binary decisions, the quantum representation allows a word to exist in a superposition of both states Eq. 4.30.

$$\alpha|0\rangle + \beta|1\rangle, \quad |\alpha|^2 + |\beta|^2 = 1 \tag{4.30}$$

This representation enables simultaneous consideration of multiple selection possibilities, reflecting uncertainty and contextual variability in word importance.

All qubits are initialized to the computational basis state $|0\rangle$, followed by the application of the Hadamard operator Eq. 4.31.

$$H|0\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle) \tag{4.31}$$

Applying $H^{\otimes n_i}$ yields Eq. 4.32.

$$|\psi_0^{(i)}\rangle = \frac{1}{\sqrt{2^{n_i}}} \sum_{x \in \{0,1\}^{n_i}} |x\rangle \tag{4.32}$$

This uniform superposition represents all possible keyword subsets, allowing the model to evaluate multiple configurations in parallel rather than sequentially.

Semantic dependencies between words are incorporated through entanglement. For contextually related word pairs, Controlled-NOT operations are applied as Eq. 4.33.

$$\text{CNOT}_{i \to j} : |x_i\rangle \left|x_j\right\rangle \mapsto |x_i\rangle \left|x_i \oplus x_j\right\rangle \tag{4.33}$$

Through entanglement, the selection state of one word becomes dependent on another, enabling the model to capture contextual associations such as co-occurring or semantically linked terms.

To incorporate semantic importance into the quantum state, parameterized rotations are applied to each qubit Eq. 4.34.

$$R_x(\theta_j^{(i)}) = \begin{bmatrix} \cos\left(\frac{\theta_j^{(i)}}{2}\right) & -i\sin\left(\frac{\theta_j^{(i)}}{2}\right) \\ -i\sin\left(\frac{\theta_j^{(i)}}{2}\right) & \cos\left(\frac{\theta_j^{(i)}}{2}\right) \end{bmatrix} \tag{4.34}$$

where $\theta_j^{(i)} = \alpha \cdot P_{jj}^{(i)}$

Here, the rotation angle is derived from the self-importance score of each word. Larger values of $P_{jj}^{(i)}$ increase the probability amplitude of the $|1\rangle$ state, thereby increasing the likelihood that the corresponding word is selected.

The combination of entanglement and rotation operations defines the unitary evolution $\mathscr{U}^{(i)}$, allowing each word to adjust its state based on both intrinsic importance and contextual relationships. This joint evolution captures cooperative behavior among words, leading to a stable distribution over possible keyword subsets.

After unitary evolution, the final quantum state $|\psi_f^{(i)}\rangle$ is measured in the computational basis, producing a binary vector $x^{(i)} \in \{0,1\}^{n_i}$.

To reduce the effect of stochastic variation in quantum measurement, the process is repeated γ times. The empirical selection frequency for each word is computed as Eq. 4.35.

$$f_j^{(i)} = \frac{1}{\gamma}\sum_{s=1}^{\gamma} \mathbb{I}[x_j^{(i,s)} = 1] \tag{4.35}$$

This frequency reflects the stability of a word's selection across multiple measurements and serves as an estimate of its contextual importance.

The final keyword set is determined by thresholding these frequencies given by Eq. 4.36.

$$\mathscr{K}^{(i)} = \{w_j^{(i)} \mid f_j^{(i)} \geq \tau\} \tag{4.36}$$

This selection mechanism ensures that words consistently identified as important across multiple quantum evaluations are retained as keywords.

Overall, this formulation integrates superposition for parallel exploration, entanglement for contextual dependency modeling, and parameterized evolution for importance weighting within a unified framework. As a result, keyword extraction is guided by collective semantic interactions, enabling a more context-aware and globally consistent selection process.

4.4.4 Algorithmic Summary of QuToX

For clarity, Algorithm 5 summarizes the workflow of the proposed QuToX framework, comprising semantic representation, interaction modeling, quantum evolution, and final keyword selection.

Algorithm 5 Stage-wise Schematic of the Proposed QuToX Framework

Input: Preprocessed microtext corpus $\mathscr{D} = \{d_1, \ldots, d_n\}$
Output: Keyword set $\mathscr{K}^*$
Stage 1: Semantic Representation
Encode tokens using BERT: $\mathbf{v}_j^{(i)} \leftarrow \text{BERT}(w_j^{(i)})$
Stage 2: Interaction Modeling
Compute similarity: $S_{p,q}^{(i)} \leftarrow \cos(\mathbf{v}_p^{(i)}, \mathbf{v}_q^{(i)})$
Context adjustment: $I_{p,q}^{(i)} \leftarrow S_{p,q}^{(i)} + 0.5\lambda$
Normalize: $\hat{I}_{p,q}^{(i)} \leftarrow \text{softmax}(I_{p,q}^{(i)})$
Payoff matrix: $P_{p,q}^{(i)} \leftarrow \pi \cdot \hat{I}_{p,q}^{(i)}$
Stage 3: Quantum Initialization
Initialize qubits: $q_j^{(i)} \leftarrow |0\rangle$
Apply superposition: $|\psi_0^{(i)}\rangle \leftarrow H^{\otimes n_i}$
Stage 4: Quantum Evolution
Apply entanglement (CNOT) based on interactions
Rotation: $\theta_j^{(i)} \leftarrow \alpha P_{jj}^{(i)}$
Update states: $q_j^{(i)} \leftarrow R_x(\theta_j^{(i)})$
Stage 5: Measurement and Selection
Measure γ times: $x^{(i)} \in \{0,1\}^{n_i}$
Frequency: $f_j^{(i)} \leftarrow \frac{1}{\gamma} \sum \mathbb{I}[x_j^{(i)} = 1]$
Extract keywords: $\mathscr{K}^{(i)} \leftarrow \{w_j^{(i)} \mid f_j^{(i)} \geq \tau\}$
Return $\mathscr{K}^*$

4.5 Experimental Evaluation

This section presents the evaluation of the keyword extraction component of the QuToX framework. The analysis focuses on examining the behavior and consistency of the extracted keywords across tweets with varying lengths and linguistic characteristics.

4.5.1 Dataset

The evaluation is conducted using a dataset of tweets obtained by combining multiple publicly available sources containing toxic and non-toxic content. The datasets are merged without

modifying the original tweets in order to preserve the natural linguistic properties of social media text, including informal expressions, abbreviations, and variations in writing style.

The resulting dataset consists of 56,744 tweets and covers a wide range of language usage patterns. The distribution of toxic and non-toxic tweets is illustrated in Figure 4.10.

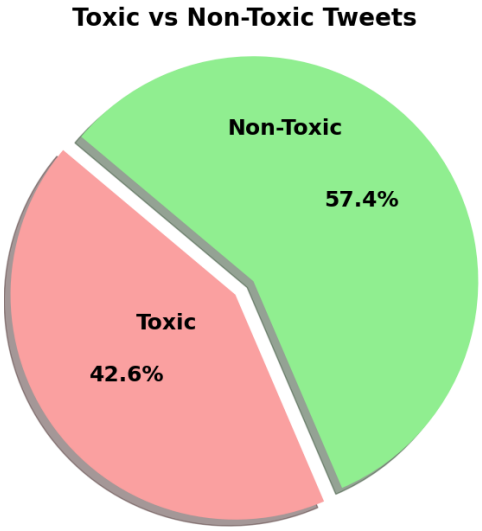

Figure 4.10: Distribution of toxic and non-toxic tweets in the dataset

4.5.2 Keyword Extraction Analysis

The effectiveness of the keyword extraction process is examined by analyzing the relationship between tweet length and the number of extracted keywords. The statistical properties of both are summarized in Table 4.6.

Table 4.6: Statistical summary of tweet length and extracted keywords

Metric	Tweet Length	Keyword Count
Average	18.87	7.40
Median	18.00	7.00
Minimum	1	1
Maximum	321	26
Std. Deviation	9.39	3.33

As shown in Table 4.6, tweet lengths vary significantly, reflecting the inherent diversity of user-generated content. Despite this variation, the keyword extraction process produces a consistent number of keywords, with an average of 7.40 keywords per tweet and no instances of zero-keyword extraction.

This consistency can be attributed to the interaction-based modeling of tokens within the quan-

tum framework, where contextual relationships influence the selection of keywords. The resulting keyword sets provide a compact representation of the original tweet content while preserving relevant contextual information.

Since the extracted keywords serve as an intermediate representation for downstream toxicity detection, comparative and statistical evaluation is performed in the subsequent section where the complete framework is assessed.

4.6 HumorKey: A Novel Keyword Extraction Framework Using Bit-Level Encoding and Fisher–Yates Shuffling for Humor Detection

The HumorKey framework approaches keyword extraction through a transformation-based representation strategy, where candidate tokens are encoded into a binary space and reorganized to reveal structurally significant patterns. Rather than relying on explicit semantic graphs or interaction models, this method operates by restructuring token representations and identifying salient keywords based on their stability across transformed configurations.

Following preprocessing (as discussed in the previous section), tokens are converted into bit-level representations and subsequently permuted using the Fisher-Yates shuffling algorithm. This controlled reordering preserves the underlying token distribution while enabling the emergence of consistent patterns across multiple transformations.

The methodology consists of three stages: bit-level encoding, structured permutation, and frequency-based keyword selection. Figure 4.11 presents the overall workflow of the HumorKey keyword extraction process.

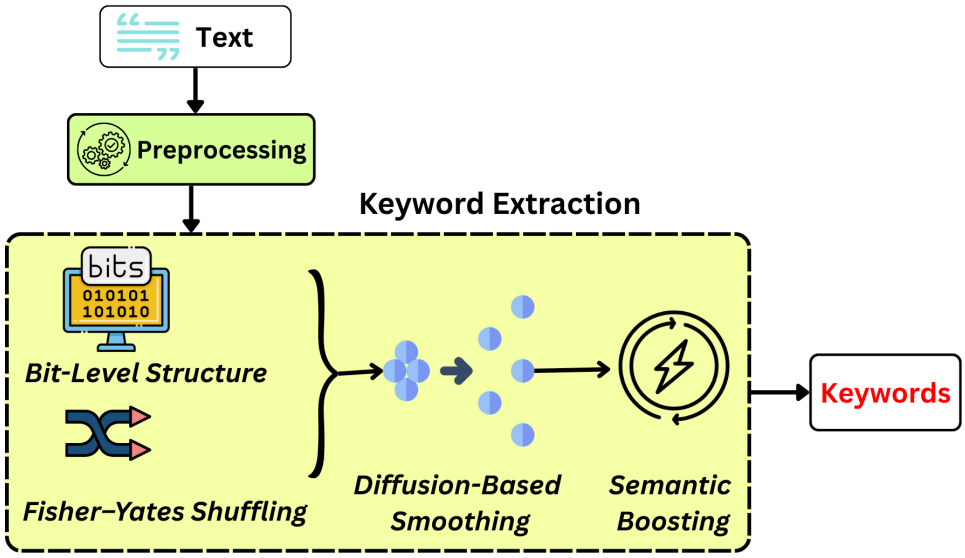

Figure 4.11: Workflow of the HumorKey keyword extraction framework based on bit-level encoding and structured permutation

4.6.1 Bit-Level Structural Encoding

Following preprocessing, each text instance is represented as a sequence of tokens, $T = \{w_1, w_2, \ldots, w_n\}$. Each token w is encoded at the character level by converting its constituent characters into corresponding 8-bit binary representations, thereby forming a binary sequence $\mathbf{b}_w$. Subsequently, the structural complexity of each token is computed using Eq. 4.37.

$$B(w) = \frac{\sum_{i=1}^{L-1} |b_i - b_{i+1}|}{L \cdot \log(|w|+1)} \cdot \frac{|\text{unique}(w)|}{|w|} \tag{4.37}$$

where L denotes the total number of bits in the word and $|\text{unique}(w)|$ represents the number of distinct characters. This score captures structural variability and character diversity within the token.

4.6.2 Permutation via Fisher-Yates Shuffling

To reduce positional bias, the token sequence is permuted using the Fisher-Yates algorithm. Given $T = \{w_1, \ldots, w_n\}$, a shuffled sequence T_s is generated as Eq. 4.38.

$$j \sim \text{Uniform}(1, i), \quad \text{swap}(w_i, w_j), \quad i = n, \ldots, 2 \tag{4.38}$$

Bit-level scores are recomputed on the shuffled sequence, and the combined structural score is

obtained as Eq. 4.39.

$$B_{\text{combined}}(w_i) = \frac{B(w_i) + B_s(w_i)}{2} \tag{4.39}$$

where $B_s(w_i)$ denotes the score computed in the shuffled configuration.

4.6.3 Contextual Diffusion

To incorporate local contextual influence, scores are smoothed using a diffusion process as in Eq. 4.40.

$$\tilde{B}(w_i) = \alpha B_{\text{combined}}(w_i) + (1-\alpha)\frac{1}{|\mathcal{N}_i|}\sum_{j \in \mathcal{N}_i} B_{\text{combined}}(w_j) \tag{4.40}$$

where $\mathcal{N}_i$ denotes the neighborhood of token w_i, and $\alpha \in [0,1]$ controls the balance between local and contextual contributions.

4.6.4 Semantic Adjustment and Keyword Selection

Semantic importance is incorporated by applying a boosting factor $\beta > 1$ to named entities and proper nouns as in Eq. 4.41.

$$\hat{B}(w_i) = \tilde{B}(w_i) \cdot \begin{cases} \beta, & w_i \in E \cup P \\ 1, & \text{otherwise} \end{cases} \tag{4.41}$$

where E and P denote the sets of named entities and proper nouns, respectively.

Tokens are ranked in descending order of $\hat{B}(w_i)$, and the top-k tokens are selected as keywords using Eq. 4.42.

$$\mathcal{K} = \text{Top-}k\{\hat{B}(w_i)\} \tag{4.42}$$

In this work, $k = 5$ is used to maintain a compact yet informative keyword set.

4.6.5 Algorithmic Summary of HumorKey

For clarity, Algorithm 6 summarizes the workflow of the proposed HumorKey framework, including bit-level encoding, structured permutation, contextual diffusion, semantic adjustment, and keyword selection.

Algorithm 6 Stage-wise Schematic of the Proposed HumorKey Framework

Input: Preprocessed token sequence $T = \{w_1, \ldots, w_n\}$
Output: Keyword set $\mathscr{K}$
Stage 1: Bit-Level Encoding
for each token $w_i \in T$ **do**
    Convert w_i into binary representation
    Compute structural score $B(w_i)$
end for
Stage 2: Fisher–Yates Permutation
Generate shuffled sequence T_s
for each token w_i **do**
    Compute shuffled score
    Compute combined structural score
end for
Stage 3: Contextual Diffusion
for each token w_i **do**
    Update score using neighboring tokens
end for
Stage 4: Semantic Adjustment
for each token w_i **do**
    Boost scores of named entities and proper nouns
end for
Stage 5: Keyword Selection
Rank tokens using final scores
Select top-k tokens as $\mathscr{K}$
return $\mathscr{K}$

The methodology described in this section focuses on keyword extraction for social media text using the proposed HumorKey framework. Since the extracted keywords are subsequently utilized as feature representations for humor classification, their effectiveness is evaluated in the application chapter, where the complete humor detection framework is assessed.

4.7 Conclusion

This chapter presented three complementary frameworks for keyword extraction from social media text, namely COM-Key, QuToX, and HumorKey, each designed to address the unique

challenges posed by short, noisy, and contextually sparse microtext environments.

COM-Key introduced a semantically enriched graph-based framework that integrates contextual embeddings, commonsense knowledge, and fuzzy graph modeling. By combining multiple fuzzy centrality measures within a multi-criteria decision-making framework, COM-Key demonstrated the ability to capture both local and global semantic importance. The experimental evaluation showed consistent improvements in keyword relevance across datasets of varying scale and complexity, highlighting the effectiveness of incorporating external knowledge and uncertainty-aware graph structures.

The keyword extraction component of QuToX approached the problem from a different perspective by modeling token interactions through a quantum-inspired framework. By representing words as interdependent entities and capturing contextual relationships through interaction matrices, the method enabled a more holistic assessment of word importance. The analysis of extracted keywords indicated stable performance across tweets of varying lengths, demonstrating the robustness of the approach in handling the variability inherent in social media text.

HumorKey further extended the methodological scope by introducing a lightweight yet structurally informed keyword extraction strategy based on bit-level encoding and structured permutation. By capturing intrinsic lexical patterns and contextual dependencies through diffusion and semantic adjustment, the framework provided compact and informative keyword representations suitable for downstream tasks such as humor detection.

Collectively, the progression from COM-Key to QuToX to HumorKey reflects a shift from semantically enriched graph-based modeling to interaction-driven representations and finally to efficient structural encoding mechanisms. Across all three approaches, the results emphasize that effective keyword extraction in social media contexts requires a combination of semantic awareness, contextual interaction modeling, and robustness to linguistic variability.

Overall, this chapter establishes that keyword extraction from social media text benefits from hybrid methodologies that integrate semantic representations, structural analysis, and adaptive modeling techniques. These findings provide a foundation for the subsequent application-oriented analysis, where the extracted keywords are utilized for downstream tasks such as toxicity and humor detection.

Chapter 5

Multimodal Keyword Extraction

The preceding chapters addressed keyword extraction in text-only settings, focusing on structured documents and social media data. In these contexts, semantic interpretation relies exclusively on linguistic information, where keywords are derived from textual features, contextual embeddings, and graph-based or structural representations. While these approaches have demonstrated effectiveness, they are inherently limited in scenarios where meaning is distributed across multiple modalities.

In many real-world applications, textual content is accompanied by visual information in the form of images, memes, or multimedia posts. In such cases, textual cues alone may be insufficient to fully capture the underlying semantic intent. For instance, social media posts often rely on images to convey context, sentiment, or implicit meaning that is not explicitly expressed in the accompanying text. This necessitates the development of keyword extraction techniques that can jointly interpret both textual and visual signals.

Multimodal keyword extraction extends traditional approaches by integrating information from multiple modalities to generate more informative and contextually grounded keywords. Rather than treating text in isolation, the proposed framework models the interaction between textual features and visual representations, enabling the extraction of keywords that reflect the combined semantic content. This requires mechanisms for cross-modal alignment, where representations from different modalities are mapped into a shared semantic space [153].

In this chapter, we address the problem of keyword extraction in multimodal environments by

developing methods that jointly analyze textual and visual inputs. The proposed approaches aim to enhance keyword relevance by incorporating visual context, thereby improving semantic coverage in complex communication scenarios. This transition from unimodal to multimodal processing represents a natural progression toward more comprehensive and context-aware keyword extraction systems.

5.1 Multimodal Content

This section provides the conceptual foundation for multimodal keyword extraction by examining the nature of multimodal data and its relevance in social media environments. Unlike traditional text-centric data, multimodal content conveys information through the combination of multiple modalities, where meaning emerges from their interaction rather than from any single modality in isolation [153].

Multimodal data typically involves the integration of heterogeneous sources such as text, images, audio, and video. These modalities complement each other by contributing different aspects of semantic information, thereby enabling richer and more expressive representations. In recent years, multimodal learning has gained significant attention due to its ability to improve performance in tasks where unimodal approaches are insufficient [154, 155].

In the context of social media, multimodal communication is particularly prevalent. Users frequently combine short textual expressions with images to convey meaning, context, and emotion. In such scenarios, textual content alone may not fully capture the intended message, as visual elements often provide essential contextual cues or implicit information. This necessitates the development of methods that jointly model textual and visual modalities for more accurate semantic interpretation.

The focus of this chapter is on image-text multimodal content, where both modalities contribute to the overall meaning. Among various forms of multimodal data, this category is especially relevant for keyword extraction in social media, as it reflects common patterns of user-generated content such as posts, advertisements, and memes. In particular, memes represent a complex and highly contextual form of multimodal communication, where meaning arises from the interplay between visual and textual components.

The subsequent sections further categorize different types of multimodal data and examine the

characteristics of image-text content in social media, providing the necessary foundation for the proposed multimodal keyword extraction framework.

5.1.1 Types of Multimodal Content and Image-Text Data in Social Media

Multimodal data can be categorized based on the combination of modalities involved, such as text-image, audio-visual, and text-audio content. Each form involves the integration of heterogeneous information sources, where different modalities contribute complementary aspects of meaning.

Among these, text-image content is particularly relevant in the context of online communication. This category includes social media posts, advertisements, and other forms of user-generated content where textual and visual elements coexist. The interaction between these modalities may vary, with images either reinforcing, modifying, or implicitly extending the meaning conveyed by text.

In social media environments, the role of image-text content becomes more prominent due to the brevity of textual expressions. Users often rely on images to provide additional context, convey emotions, or introduce implicit meaning that is not explicitly stated in text. Consequently, analyzing text in isolation may lead to incomplete interpretation of the overall message.

From the perspective of keyword extraction, this introduces additional challenges. Keywords derived solely from textual content may fail to capture concepts that are visually implied. Therefore, effective keyword extraction in such settings requires integrating visual information with textual cues to obtain a more comprehensive semantic representation.

Among different forms of image-text content, memes represent a particularly complex and expressive category, where meaning is strongly dependent on cross-modal interaction. The following subsection examines memes in detail.

5.1.2 Memes as Multimodal Communication Units

Memes have emerged as a dominant form of communication in online environments, combining visual and textual elements to convey meaning in a compact and often implicit manner. Unlike traditional forms of content, memes rely on the interaction between modalities to express humor, satire, and contextual commentary. The meaning of a meme is therefore not

contained within the text or image independently, but arises from their combined interpretation.

In social media platforms, memes are widely used to communicate ideas that range from everyday experiences to sociocultural and political narratives. The textual component typically consists of short, informal phrases, often incorporating humor, exaggeration, or sarcasm. The visual component, on the other hand, provides contextual grounding through expressions, symbolic imagery, or widely recognized templates. Together, these elements create a multimodal representation in which meaning is distributed across modalities.



Figure 5.1: Examples of humorous memes illustrating how meaning emerges from the interaction between visual context and textual content

As illustrated in Figure 5.1, the interpretation of memes is inherently dependent on cross-modal relationships. The visual component establishes a scenario or emotional tone, while the textual component provides a reinterpretation, punchline, or contextual shift. This interaction enables the expression of humor, irony, and relatability in a concise format.

A key characteristic of memes is their reliance on implicit context and shared cultural knowledge. Many memes derive meaning from widely recognized templates, symbolic references, or recurring formats that are understood within specific communities. Consequently, interpreting a meme often requires background knowledge beyond the explicit content presented.

From the perspective of keyword extraction, memes introduce significant challenges. Important concepts may not be explicitly stated in the text but instead implied through visual elements or contextual associations. As a result, traditional text-based keyword extraction methods may fail to capture the full semantic representation of the content. This highlights the

need for multimodal approaches that jointly analyze textual and visual information to identify meaningful keywords.

5.1.3 Offensive and Contextual Memes

While memes are often associated with humor and entertainment, a significant subset contains offensive or sensitive content. In such cases, the meaning is not always explicitly stated but emerges through subtle interactions between textual and visual elements. This implicit nature makes offensive memes particularly challenging to interpret and analyze.

Offensiveness in memes is frequently context-dependent and may rely on cultural references, symbolic associations, or implicit linguistic cues. The textual component alone may appear neutral or ambiguous, while the visual component introduces contextual signals that alter the interpretation. Conversely, the image may seem benign in isolation, but when combined with text, it conveys a harmful or derogatory meaning.



Figure 5.2: Examples of offensive memes where meaning emerges from cross-modal interaction between image and text

As illustrated in Figure 5.2, offensive meaning often arises through indirect or implicit mechanisms. In the first example, the textual statement appears exaggerated and satirical, but its interpretation depends on the contextual association with the visual representation. In the second example, the textual reference may seem descriptive, yet its combination with the visual content introduces a derogatory undertone linked to identity-based interpretation. In the third example, the phrase used in the text appears harmless in isolation but acquires offensive connotation when interpreted in relation to the depicted scene.

Such cases highlight the limitations of unimodal analysis. Traditional text-based approaches may fail to detect harmful intent when explicit offensive terms are absent, while image-only

analysis may overlook contextual linguistic cues. The semantic signal is therefore distributed across modalities and requires integrated interpretation.

From the perspective of keyword extraction, offensive memes present additional challenges. Keywords indicative of harmful content may not be explicitly present but instead inferred through cross-modal relationships. This necessitates the development of multimodal keyword extraction techniques capable of capturing implicit semantics, contextual dependencies, and culturally grounded interpretations.

5.1.4 Need for Multimodal Reasoning in Offensive Meme Interpretation

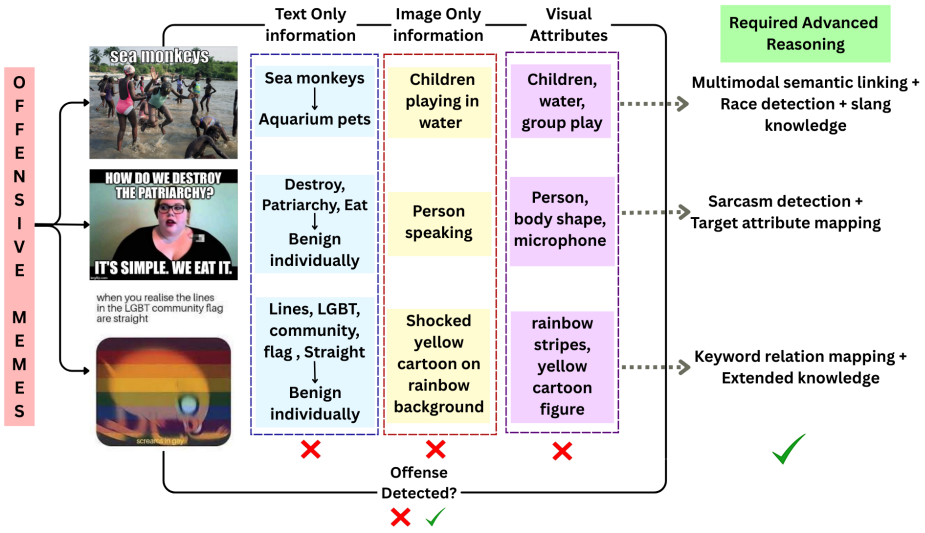

Figure 5.3: Illustration of how meme interpretation evolves from text-only, image-only, and visual attribute-level analysis to integrated multimodal reasoning.

Figure 5.3 demonstrates that the interpretation of offensive memes cannot be reliably achieved through isolated analysis of individual information sources. Each example is examined at three distinct levels: *text-only information*, *image-only information*, and *visual attribute-level interpretation*. When considered independently, all three levels yield benign or non-offensive interpretations.

At the text-only level, the extracted words or phrases appear harmless. For instance, terms such as “sea monkeys,” references to social or political ideas, or descriptions of identity-related concepts do not inherently indicate offensive intent when interpreted in isolation. Similarly, at the image-only level, the visual content presents ordinary scenes such as children playing, a person speaking, or symbolic imagery, none of which directly convey harmful meaning. At the visual attribute level, features such as objects, entities, colors, or contextual descriptors

provide additional semantic cues; however, these attributes remain individually neutral and insufficient to signal offensiveness.

The limitation becomes evident when all three levels are analyzed independently, as none of them alone captures the implicit intent embedded in the meme. In each example shown in Figure 5.3, the textual content, visual content, and extracted attributes individually suggest non-offensive interpretations, leading to incorrect or incomplete conclusions when analyzed in isolation.

Offensive meaning emerges only when these levels are jointly interpreted through multimodal reasoning. This involves linking textual cues with visual context, incorporating cultural and contextual knowledge, and analyzing relationships between attributes across modalities. For example, a seemingly neutral phrase may acquire a different interpretation when aligned with specific visual elements, while symbolic imagery may alter the meaning of otherwise benign text through contextual association. Such reasoning requires understanding not only the presence of features but also their interactions and underlying implications.

These observations highlight that offensive meme interpretation is fundamentally a cross-modal reasoning problem rather than a unimodal or attribute-level classification task. From the perspective of keyword extraction, this further emphasizes the need for approaches that capture not only explicit textual signals but also implicit semantics arising from visual context and multimodal interactions. Effective keyword extraction in this domain must therefore integrate information across modalities to accurately represent the underlying meaning of the content.

5.1.5 Multimodal Keyword Extraction

Keyword extraction aims to identify a set of representative terms that summarize the core semantic content of input data. In traditional settings, this process is primarily performed on textual data, where keywords are selected based on lexical, statistical, or contextual properties. However, in multimodal environments such as memes, textual information alone is often insufficient to capture the complete meaning of the content.

As discussed in the preceding sections, the interpretation of multimodal content is inherently dependent on the interaction between textual and visual modalities. Important semantic cues may be visually implied rather than explicitly expressed in text, while textual elements may modify or reinterpret the visual context. Consequently, keywords derived solely from text may

fail to represent implicit meanings, and those derived from visual attributes alone may lack contextual grounding.

This limitation becomes particularly significant in the analysis of memes, where meaning is distributed across modalities and often relies on contextual or culturally grounded interpretations. Keywords that accurately represent such content must therefore reflect both explicit textual information and implicit semantic associations arising from visual context.

Multimodal keyword extraction addresses this challenge by integrating textual and visual representations to identify semantically meaningful terms. This involves capturing cross-modal dependencies, aligning features across modalities, and incorporating contextual relationships that influence interpretation. Rather than treating modalities independently, the extraction process considers their joint contribution to the overall meaning.

In the context of offensive or context-sensitive memes, this approach is especially critical. Keywords indicative of harmful or implicit content may not be directly observable in text or image alone but emerge through their interaction. Therefore, effective keyword extraction must incorporate multimodal reasoning to ensure that the extracted keywords provide an accurate and comprehensive representation of the underlying semantics.

5.2 Knowledge-Enriched Semantic Keyword Extraction with NSGA-II for Offensive Meme Detection via Thermodynamic Modeling

This section presents the proposed multimodal keyword extraction framework designed for meme-based content. The framework integrates textual and visual semantic representations to identify contextually meaningful keywords that capture both explicit and implicit semantic cues. Unlike conventional keyword extraction approaches that primarily rely on textual information, the proposed framework incorporates cross-modal interactions and contextual reasoning to better represent the underlying meaning of multimodal content.

The methodology consists of multiple sequential stages involving textual processing, visual feature representation, semantic enrichment, cross-modal contextual alignment, and optimized keyword extraction. Initially, textual and visual modalities are processed independently to obtain structured semantic representations. These representations are subsequently integrated to capture relationships between linguistic expressions and visual context within the meme.

To improve semantic coverage, the framework further incorporates knowledge-enriched contextual expansion mechanisms that extend the representation beyond directly observable features. The extracted semantic cues are then refined through an optimization-driven keyword selection strategy based on Non-dominated Sorting Genetic Algorithm II (NSGA-II), enabling the identification of representative multimodal keywords that preserve semantic relevance and contextual consistency.

Figure 5.4 illustrates the overall workflow of the proposed framework for multimodal keyword extraction from meme-based content.

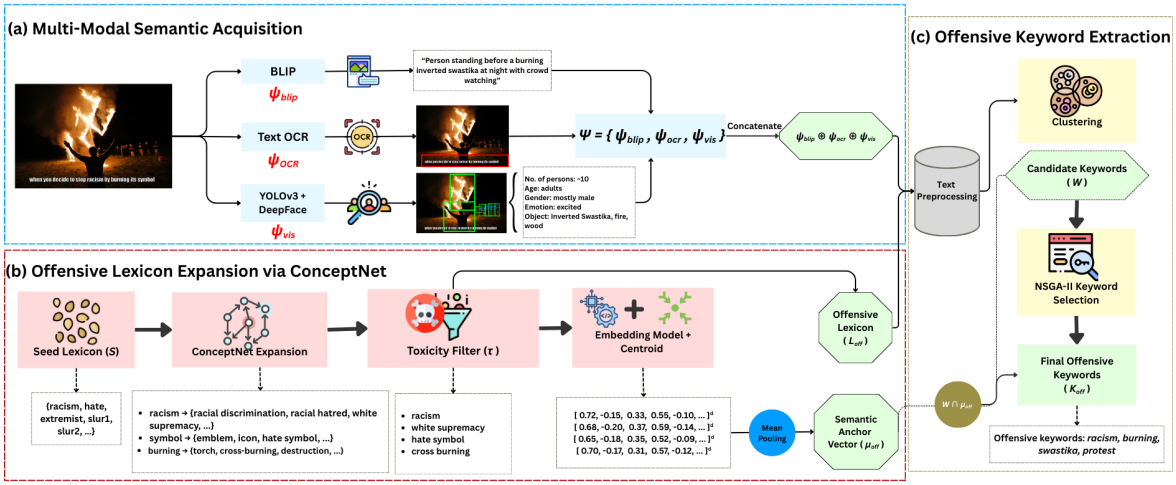

Figure 5.4: Overall workflow of the proposed knowledge-enriched multimodal keyword extraction framework for meme-based content

The following subsections describe each stage of the proposed framework in detail.

5.2.1 Multi-modal Semantic Acquisition

The first stage of the proposed framework focuses on acquiring semantically rich information from multiple complementary modalities present within meme-based content. Unlike conventional textual documents, memes communicate meaning through the interaction of visual scenes, embedded textual expressions, symbolic imagery, demographic cues, and contextual associations. Consequently, analyzing a single modality often leads to incomplete semantic interpretation, since important contextual signals may remain implicit within other modali-

ties. To address this limitation, the framework constructs a unified semantic representation by jointly extracting scene-level descriptions, embedded textual content, and social contextual visual attributes.

Let the meme image be represented as an image tensor $\mathscr{M} \in \mathbb{R}^{H \times W \times 3}$, where H and W denote the spatial dimensions of the image and the third dimension corresponds to RGB color channels.

The objective of this stage is to transform the raw meme image into a structured multimodal semantic representation capable of preserving contextual relationships across visual and textual modalities. To achieve this, the framework derives three complementary semantic components represented collectively as Eq. 5.1.

$$\Psi = \left\{ \psi_{\text{blip}}, \psi_{\text{ocr}}, \psi_{\text{vis}} \right\} \in \mathscr{T}^3 \tag{5.1}$$

where:

- ψ_{blip} represents the scene-level semantic caption,
- ψ_{ocr} denotes the textual information extracted from the meme,
- ψ_{vis} corresponds to demographic and emotive visual attributes.

The meme instance considered in Figure 5.4(a) depicts a burning symbolic structure surrounded by multiple individuals, accompanied by the textual statement *"when you decide to stop racism by burning its symbol"*. Individually, the image, textual content, and visual attributes provide only partial semantic information. However, when integrated, these modalities collectively reveal the broader contextual meaning embedded within the meme.

To construct the multimodal semantic representation, the framework sequentially extracts scene-level semantics, embedded textual information, and socially contextual visual attributes from the meme image.

5.2.1.1 Scene-Level Semantic Captioning using BLIP

The first semantic modality focuses on extracting scene-level contextual understanding from the image using the pretrained vision-language transformer *Bootstrapping Language-Image Pre-training (BLIP)* [156]. BLIP is a unified multimodal transformer framework designed to bridge the semantic gap between visual and linguistic representations through large-scale

vision-language pretraining. Unlike traditional image captioning approaches that depend on handcrafted visual descriptors or template-based representations, BLIP learns contextual associations between image regions and language through transformer-based cross-modal learning mechanisms.

Architecturally, BLIP integrates a visual encoder with a language transformer to jointly model image-text correspondence. During pretraining, the model employs objectives such as image-text contrastive learning, image-text matching, and language generation. These learning strategies enable the framework to generate semantically meaningful captions capable of describing objects, actions, spatial interactions, and contextual scene relationships.

Formally, the BLIP caption generation process is represented as $\psi_{\text{blip}} = \mathscr{F}_{\text{BLIP}}(\mathscr{M})$, where $\mathscr{F}_{\text{BLIP}}$ denotes the pretrained caption generation function and ψ_{blip} corresponds to the generated semantic representation.

For the meme shown in Figure 5.4(a), the generated caption describes a person standing before a burning symbolic structure while several individuals observe the event in the background. This representation captures high-level semantic concepts such as fire, symbolic imagery, crowd presence, and environmental context. Such scene-level understanding is important because memes frequently communicate implicit meaning through symbolic visual arrangements rather than explicit textual statements alone.

5.2.1.2 Embedded Text Extraction using OCR

Although scene-level captioning provides rich contextual semantics, memes often contain embedded textual overlays that convey sarcasm, humor, symbolic references, or contextual intent. In many cases, the textual component substantially alters the interpretation of the visual scene. Therefore, extracting embedded textual information constitutes an essential step in multimodal semantic acquisition.

To recover textual content from the meme image, the framework employs the EasyOCR framework [157], a deep learning-based Optical Character Recognition (OCR) system designed for robust multilingual text extraction from visually complex environments.

EasyOCR operates through a two-stage pipeline consisting of text detection and text recognition. Initially, candidate text regions are localized using a Character Region Awareness for Text Detection (CRAFT) based detector capable of identifying irregular and stylized textual patterns. Subsequently, recurrent neural network-based sequence recognition models decode

the detected text while preserving contextual dependencies between adjacent characters and symbols.

Prior to OCR processing, the image is transformed into a contrast-enhanced grayscale representation with increased spatial resolution $\mathscr{M}' \in \mathbb{R}^{2H \times 2W}$.

This preprocessing step improves textual visibility and enhances recognition performance for noisy, low-resolution, or stylistically distorted meme text.

The OCR extraction process is formally defined as Eq. 5.2.

$$\psi_{\text{ocr}} = \mathscr{F}_{\text{OCR}}\left(\mathscr{M}'\right) \tag{5.2}$$

where $\mathscr{F}_{\text{OCR}}$ denotes the OCR recognition function and ψ_{ocr} represents the extracted textual information.

The OCR module extracts the statement *"when you decide to stop racism by burning its symbol"*, introducing explicit semantic concepts such as *racism*, *burning*, and *symbol*. These linguistic cues provide contextual information that cannot be fully inferred through visual scene understanding alone. Consequently, OCR-based text extraction complements the scene-level semantic representation by incorporating embedded linguistic semantics central to meme interpretation.

5.2.1.3 Visual Demographic and Emotive Analysis

While scene captioning and OCR extraction capture contextual and linguistic information, memes also communicate meaning through socially contextual visual attributes such as group composition, emotional tone, facial expressions, and demographic indicators. Such information becomes particularly important in context-sensitive memes where interpretation depends on identity-related associations, collective interaction, or emotionally grounded symbolism.

To incorporate these cues, the framework first performs object-level person detection using the *You Only Look Once Version 3 (YOLOv3)* object detection model [158]. YOLOv3 is a single-stage real-time object detection architecture that simultaneously performs object localization and classification within a unified convolutional neural network framework. Unlike region proposal-based detectors that separately generate candidate object regions before classification, YOLOv3 divides the image into spatial grids and directly predicts bounding boxes and class probabilities in a single forward pass. This enables efficient detection while maintaining strong localization accuracy across complex visual environments.

Let c_i denote the detected class label corresponding to the ith detected object instance. The number of individuals present within the meme image is computed as Eq. 5.3.

$$\eta = \sum_{i=1}^{K} 1\{c_i = \text{ person }\} \tag{5.3}$$

where K represents the total number of detected object instances and η denotes the number of detected persons.

The framework identifies multiple individuals surrounding the burning symbolic structure, thereby capturing group-level contextual information associated with public gathering and collective visual activity.

Following person detection, facial regions are analyzed using the DeepFace framework [159]. DeepFace is a deep learning-based facial analysis framework that integrates multiple convolutional neural network architectures for demographic estimation, facial representation learning, and emotion recognition.

For each detected facial region $f_j \subset \mathscr{M}$, the framework estimates a structured attribute tuple $d_j = (a_j, g_j, r_j, em_j)$.

where:

- a_j denotes estimated age,
- g_j represents gender,
- r_j corresponds to ethnicity,
- em_j indicates the dominant emotional expression.

The ethnicity label space is represented as $\mathscr{C}_{\text{eth}} = \{\text{White}, \text{Black}, \text{Asian}, \text{Hispanic}, \text{Indian}, \text{Other}\}$, following demographic categorization schemes commonly adopted in visual bias analysis studies such as Gender Shades [160].

DeepFace utilizes pretrained convolutional neural networks trained on large-scale facial datasets to generate discriminative facial embeddings capable of estimating demographic and emotional characteristics. The extracted attributes are subsequently aggregated into a unified visual semantic representation defined as Eq. 5.4.

$$\psi_{\text{vis}} = \text{Summarize}\left(\{d_j\}_{j=1}^{m}, \eta\right) \in \mathscr{T} \tag{5.4}$$

where m denotes the number of analyzable facial regions within the image.

The resulting representation captures socially contextual information such as crowd composition, emotional expressions, and visually significant entities including fire and symbolic imagery. These cues complement the scene-level caption and embedded textual information, thereby contributing additional semantic signals necessary for deeper contextual interpretation.

5.2.1.4 Unified Multimodal Semantic Representation

The semantic outputs obtained from the three modalities are integrated to construct a unified multimodal representation of the meme content. This integration enables the framework to jointly preserve contextual visual semantics, embedded linguistic information, and social-contextual visual cues within a common semantic space.

Individually, each modality captures only a limited aspect of the meme. Scene-level captioning primarily describes objects, environmental context, and symbolic visual interactions, whereas OCR recovers explicit textual expressions embedded within the image. Similarly, visual attribute analysis contributes demographic and emotive information associated with facial expressions, group composition, and social context. When considered independently, these representations may overlook implicit relationships emerging from cross-modal interactions.

The integration of these modalities therefore enables the framework to capture semantic associations that remain inaccessible under isolated modality analysis. BLIP contributes scene-level contextual understanding, OCR introduces linguistic intent, and visual attribute analysis incorporates emotional and demographic context. Collectively, these complementary representations preserve both explicit and implicit semantic cues, thereby enabling richer interpretation of multimodal meme content.

The final integrated representation is formally defined in Eq. 5.5.

$$\Psi = \left\{ \psi_{\text{blip}}, \psi_{\text{ocr}}, \psi_{\text{vis}} \right\} \in \mathscr{T}^3 \tag{5.5}$$

The resulting multimodal representation serves as the foundational semantic input for the subsequent stages involving semantic enrichment and optimization-driven keyword extraction.

5.2.2 Offensive Lexicon Expansion via ConceptNet

Following multimodal semantic acquisition, the framework performs semantic lexicon expansion to enrich the contextual representation of potentially offensive concepts present within meme content. In multimodal memes, offensive meaning is often expressed indirectly through sarcasm, coded language, symbolic references, or semantically associated expressions rather than explicit toxic keywords alone. Consequently, relying solely on a fixed vocabulary of offensive terms limits the ability of the framework to capture implicit and context-dependent semantics.

To address this limitation, the proposed framework constructs a dynamically enriched offensive lexicon, denoted as $\mathscr{L}_{\text{off}}$, which incorporates not only explicitly toxic terms but also semantically affiliated concepts associated with offensive intent. This lexicon serves as a semantic prior that guides subsequent keyword extraction toward linguistically and contextually relevant indicators of harmful content.

The construction of $\mathscr{L}_{\text{off}}$ begins with a curated seed vocabulary $\mathscr{S} = \{s_1, s_2, \ldots, s_k\}$.

where each seed term s_i corresponds to a lexical unit empirically associated with toxic or offensive discourse. The seed vocabulary is compiled from annotated hate-speech datasets, offensive language corpora, and social media discourse repositories to ensure alignment with real-world expressions of harmful content.

Although the seed vocabulary captures explicitly offensive terminology, restricting the representation to a static set of terms reduces generalization capability and limits sensitivity toward semantically related or contextually implied expressions. Therefore, the framework expands the seed vocabulary using ConceptNet [161], a large-scale commonsense knowledge graph containing semantic and relational associations between concepts.

ConceptNet represents knowledge as a graph structure where nodes correspond to concepts and edges encode semantic relationships such as *RelatedTo*, *IsA*, *HasContext*, *Synonym*, and *UsedFor*. Unlike traditional lexical resources that primarily model synonymy, ConceptNet captures broader commonsense and contextual relationships between concepts, thereby enabling semantic expansion beyond direct lexical similarity.

For each seed term $s_i \in \mathscr{S}$ and relation type $r \subset \mathscr{R}_{\text{CN}}$, the framework queries ConceptNet to retrieve semantically related concepts connected through contextually meaningful relations. The expanded candidate concept set is formally defined as Eq. 5.6.

$$\mathscr{E} = \bigcup_{s_i \in \mathscr{S}} \bigcup_{r \subset \mathscr{R}_{\text{CN}}} \text{ConceptNet}\,(s_i, r) \tag{5.6}$$

where:

- $\mathscr{R}_{\text{CN}}$ denotes the set of semantic relations within the ConceptNet schema,
- $\text{ConceptNet}(s_i, r)$ represents the set of concepts linked to s_i through relation r.

The selected relation types are chosen to preserve semantic relevance while enabling contextual expansion of offensive concepts. This expansion process allows the framework to capture semantically related terms that may not be explicitly offensive in isolation but acquire harmful connotations under specific contextual settings.

For the meme considered in Figure 5.4(a), the OCR module extracts concepts such as *racism*, *burning*, and *symbol*. Through ConceptNet expansion, these terms become associated with semantically related concepts involving discrimination, hatred, violence, protest symbolism, and socially contextual expressions. Such expansion enables the framework to internalize contextual semantic relationships that remain inaccessible through direct lexical matching alone.

However, graph-based expansion may also introduce unrelated or weakly associated concepts due to the broad connectivity of commonsense knowledge graphs. To reduce semantic drift and preserve offensive contextual relevance, each candidate concept $e_j \in \mathscr{E}$ is evaluated using a pretrained neural toxicity classifier as in Eq. 5.7.

$$f_{\text{tox}} : \mathscr{V} \rightarrow [0, 1] \tag{5.7}$$

where $\mathscr{V}$ denotes the vocabulary space and the output represents the toxicity likelihood associated with a candidate term.

A toxicity threshold $\tau = 0.3$ is employed to retain only those expanded concepts exhibiting sufficient alignment with offensive discourse. The final offensive lexicon is therefore constructed as defined in Eq. 5.8.

$$\mathscr{L}_{\text{off}} = \mathscr{S} \cup \left\{ e_j \in \mathscr{E} \mid f_{\text{tox}}\,\left(e_j\right) \geq \tau \right\} \tag{5.8}$$

This filtering stage suppresses unrelated expansions while preserving contextually relevant offensive associations.

5.2.3 Offensive-Aware Keyword Extraction via NSGA-II

Following multimodal semantic acquisition and offensive lexicon expansion, the proposed approach identifies a compact set of informative keywords capable of representing both the contextual meaning and offensive characteristics of meme content. Unlike conventional keyword extraction techniques that rely primarily on lexical importance or frequency statistics, the proposed methodology formulates keyword extraction as a constrained multi-objective optimization problem involving diversity preservation, representational coverage, and offensive relevance.

To simultaneously optimize these competing objectives, the methodology employs the *Non-dominated Sorting Genetic Algorithm II (NSGA-II)* [162], an evolutionary optimization technique capable of exploring multiple trade-offs while preserving diverse candidate solutions. The extracted keywords subsequently serve as high-level indicators for downstream multimodal offense analysis.

5.2.3.1 Candidate Keyword Generation and Semantic Representation

Using the multimodal abstraction $\Psi = \{\psi_{\text{blip}}, \psi_{\text{ocr}}, \psi_{\text{vis}}\}$ and the expanded offensive lexicon $\mathscr{L}_{\text{off}}$, the methodology first constructs a refined candidate keyword space for optimization-driven extraction. Since meme content frequently contains sparse textual cues, stylistic distortions, abbreviated expressions, and multimodal noise, direct keyword selection from raw outputs often produces weak or repetitive results. Therefore, an intermediate semantic consolidation stage is performed prior to optimization.

The semantic descriptors obtained from the acquisition stage are concatenated into a unified textual document as in Eq. 5.9.

$$D = \psi_{\text{blip}} \oplus \psi_{\text{ocr}} \oplus \psi_{\text{vis}} \tag{5.9}$$

where $\oplus$ denotes multimodal textual concatenation.

The resulting document incorporates scene descriptions, embedded linguistic content, and social-contextual visual attributes extracted from the meme. For the meme illustrated in Figure 5.4(a), the integrated description contains concepts associated with symbolic burning, racism-related textual cues, crowd presence, and emotionally charged visual elements. Such integration preserves cross-modal relationships that remain inaccessible under isolated

modality processing.

To improve linguistic consistency and suppress noisy expressions commonly observed in social media content, the document undergoes several preprocessing operations, including:

- lowercasing,
- profanity normalization,
- token-level filtering,
- stopword removal,
- elimination of non-informative symbols and numeric tokens,
- spelling correction using SymSpell,
- noun phrase chunking,
- named entity recognition.

Profanity normalization standardizes lexical variations of offensive expressions frequently used to evade moderation systems, whereas spelling correction reduces the impact of intentionally manipulated toxic terminology. Similarly, noun phrase chunking and named entity recognition preserve meaningful multi-word structures that may otherwise be fragmented during tokenization.

Then the refined candidate vocabulary is represented as $\mathscr{W} = \{w_1, w_2, \ldots, w_N\}$, where each term denotes an informative lexical unit extracted from the multimodal description.

To enable embedding-based analysis, each candidate term is mapped into a dense vector space using a pretrained SentenceTransformer encoder as in Eq. 5.10.

$$\mathbf{e}_i = f_{\text{emb}}(w_i) \tag{5.10}$$

where:

- $f_{\text{emb}}(\cdot)$ denotes the embedding function,
- $\mathbf{e}_i \in \mathbb{R}^d$ represents the vector embedding associated with token w_i.

These embeddings preserve contextual similarity relationships within high-dimensional vector spaces, thereby enabling semantically related concepts to occupy neighboring regions within the embedding manifold.

The embedding representation of the candidate vocabulary is defined as Eq. 5.11.

$$\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N]^\top \in \mathbb{R}^{N \times d} \tag{5.11}$$

Although the candidate vocabulary preserves informative concepts, terms extracted from different modalities may still convey overlapping meanings. For instance, concepts such as *burning*, *fire*, and *flames* may simultaneously appear within visual captions and OCR-derived text. Retaining all such overlapping terms reduces informational diversity and biases the extraction process toward repetitive concepts.

To address this issue, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is applied over the embedding space $\mathbf{E}$. DBSCAN groups closely related embeddings into density-connected clusters while treating isolated points as noise. Unlike centroid-based clustering techniques, DBSCAN does not require prior specification of cluster count and can effectively capture irregular structures within high-dimensional embedding spaces.

A representative concept from each cluster is retained to construct the reduced candidate vocabulary $\mathscr{W}' \subseteq \mathscr{W}$.

This reduction stage suppresses repetitive concepts while preserving informational diversity within the candidate space.

To guide extraction toward offensive contextual cues, the methodology additionally incorporates the offensive anchor embedding $\boldsymbol{\mu}_{\text{off}} \in \mathbb{R}^d$, computed as the centroid representation of the expanded lexicon $\mathscr{L}_{\text{off}}$. This embedding acts as a semantic reference point that biases selection toward concepts associated with offensive discourse.

Consequently, the optimization stage operates using:

- the reduced vocabulary $\mathscr{W}'$,
- the embedding matrix $\mathbf{E}$,
- the offensive anchor embedding $\boldsymbol{\mu}_{\text{off}}$.

These representations collectively provide the foundation for optimization-driven multimodal keyword extraction.

5.2.3.2 Multi-objective Keyword Optimization

After constructing the refined candidate space, keyword extraction is formulated as a constrained multi-objective optimization problem. The objective is to identify a compact subset of keywords capable of jointly preserving representational coverage, minimizing redundancy, and maintaining affinity with offensive discourse.

These objectives are inherently conflicting in multimodal meme analysis. A highly diverse keyword subset may fail to preserve contextual coherence, whereas maximizing representational coverage alone may result in repetitive selections. Similarly, excessive emphasis on offensive relevance may bias extraction toward explicitly toxic expressions while overlooking important contextual indicators contributing to implicit interpretation.

Let the reduced candidate vocabulary be represented as $\mathscr{W}' = \{w_1, w_2, \ldots, w_N\}$. The selection process is encoded using a binary decision vector $x \in \{0,1\}^N$ such that Eq. 5.12.

$$x_i = \begin{cases} 1, & \text{if } w_i \text{ is selected} \\ 0, & \text{otherwise} \end{cases} \tag{5.12}$$

subject to the cardinality constraint $\sum_{i=1}^{N} x_i = k$, where k denotes the number of keywords to be extracted.

Each feasible solution therefore corresponds to a candidate keyword subset containing exactly k selected terms.

The optimization simultaneously considers three complementary objectives:

- redundancy minimization,
- representational coverage maximization,
- offensive affinity maximization.

Redundancy Minimization The first objective minimizes average pairwise similarity among selected embeddings in order to suppress repetitive concepts and preserve informational diversity.

Let $\mathscr{S}_{\text{sel}} = \{i \mid x_i = 1\}$ denote the set of selected keyword indices. The redundancy objective is defined as Eq. 5.13.

$$f_1(x) = \frac{2}{k(k-1)} \sum_{\substack{i,j \in \mathscr{S}_{\text{sel}} \\ i<j}} \cos(\mathbf{e}_i, \mathbf{e}_j) \tag{5.13}$$

Lower values of $f_1(x)$ indicate greater diversity among the selected keywords.

Representational Coverage Maximization While diversity is essential, the selected keywords must also collectively preserve the overall meaning of the multimodal meme representation. Therefore, the second objective encourages the selected keywords to remain close to their collective centroid embedding.

The centroid representation is computed as Eq. 5.14.

$$\bar{\mathbf{e}} = \frac{1}{k} \sum_{i \in \mathscr{S}_{\text{sel}}} \mathbf{e}_i \tag{5.14}$$

The corresponding coverage objective is defined as Eq. 5.15.

$$f_2(x) = -\frac{1}{k} \sum_{i \in \mathscr{S}_{\text{sel}}} \cos(\mathbf{e}_i, \bar{\mathbf{e}}) \tag{5.15}$$

This objective ensures that the extracted keywords collectively preserve the central meaning of the multimodal representation.

Offensive Affinity Maximization Offensive meme interpretation frequently depends on indirect associations, symbolic references, and contextual implications rather than explicit toxic expressions alone. Consequently, the optimization process incorporates offensive affinity as the third objective.

Using the offensive anchor embedding $\boldsymbol{\mu}_{\text{off}}$, the methodology measures similarity between selected keywords and offensive discourse patterns. The corresponding objective is defined as Eq. 5.16.

$$f_3(x) = -\frac{1}{k} \sum_{i \in \mathscr{S}_{\text{sel}}} \mathbb{1}\{\cos(\mathbf{e}_i, \boldsymbol{\mu}_{\text{off}}) > \tau\} \tag{5.16}$$

where τ denotes a similarity threshold and $\mathbb{1}\{\cdot\}$ represents the indicator function.

This objective biases the extraction process toward concepts exhibiting stronger association with offensive discourse.

The complete optimization problem is therefore formulated as Eq. 5.17.

$$\min_{x\in\{0,1\}^N} [f_1(x), f_2(x), f_3(x)] \quad \text{s.t.} \quad \sum_{i=1}^{N} x_i = k \tag{5.17}$$

The resulting formulation jointly balances diversity, representational coverage, and offensive affinity during multimodal keyword extraction.

5.2.3.3 NSGA-II-Based Evolutionary Optimization and Final Keyword Selection

To solve the formulated optimization problem, the methodology employs NSGA-II, a population based evolutionary optimization algorithm capable of identifying diverse Pareto-optimal solutions under multiple competing objectives.

Within the proposed approach, each candidate solution corresponds to a possible subset of keywords selected from the reduced vocabulary space $\mathscr{W}'$. Consequently, multimodal keyword extraction is transformed into an evolutionary search process in which candidate keyword subsets iteratively evolve toward improved objective trade-offs.

Each candidate solution is represented using a binary chromosome $x = [x_1, x_2, \ldots, x_N]$, where $x_i \in \{0,1\}$ indicates whether keyword w_i is selected.

To maintain consistency in extraction, each chromosome is constrained to contain exactly k active keywords $\sum_{i=1}^{N} x_i = k$. The optimization process begins by randomly initializing a population of feasible chromosomes satisfying the cardinality constraint. Each chromosome is subsequently evaluated using the objective vector as in Eq. 5.18.

$$\mathbf{F}(x) = [f_1(x), f_2(x), f_3(x)] \tag{5.18}$$

Since the objectives are conflicting in nature, no single solution simultaneously optimizes all criteria. Therefore, NSGA-II identifies a collection of Pareto-optimal solutions rather than a single global optimum.

A solution is considered Pareto optimal when no other feasible solution improves one objective without degrading at least one remaining objective. The collection of such non-dominated solutions forms the Pareto front $\mathscr{P}$ representing alternative trade-offs between diversity, coverage, and offensive relevance.

NSGA-II identifies Pareto-optimal solutions through non-dominated sorting, where chromosomes are ranked according to dominance relationships. Solutions belonging to lower-ranked

fronts exhibit superior objective trade-offs relative to higher-ranked solutions.

To preserve exploration diversity during optimization, NSGA-II additionally employs a crowding distance mechanism. Crowding distance estimates the density of neighboring solutions within the objective space. Solutions located within sparse regions of the Pareto front are preferentially retained, thereby preventing premature convergence and encouraging exploration of diverse keyword combinations.

The optimization process further incorporates customized evolutionary operators designed specifically for constrained keyword selection.

To preserve the fixed keyword count during crossover, the methodology employs the *Half Uniform Crossover (HUX)* operator, which exchanges differing binary positions between parent chromosomes while maintaining balanced selection distributions across offspring solutions.

Similarly, mutation is performed using a constraint-aware strategy in which one active index is deactivated while one inactive index is simultaneously activated. This operation preserves the fixed keyword cardinality while enabling exploration of alternative candidate subsets.

Through repeated application of selection, crossover, mutation, and non-dominated sorting, the population gradually evolves toward improved keyword subsets balancing all optimization objectives.

After convergence, NSGA-II produces the Pareto-optimal solution set $\mathscr{P} = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$ where each solution represents a different trade-off between diversity, representational coverage, and offensive relevance.

Since multiple Pareto-optimal solutions may exist, the methodology performs final solution selection using a weighted scalarization strategy defined by Eq. 5.19.

$$\mathscr{S}(x) = \alpha \cdot (1 - f_1(x)) + \beta \cdot (-f_2(x)) + \gamma \cdot (-f_3(x)) \tag{5.19}$$

where:

- $\alpha = 0.15$ controls diversity,
- $\beta = 0.25$ controls representational coverage,
- $\gamma = 0.60$ prioritizes offensive affinity.

The larger weight assigned to offensive affinity emphasizes the importance of preserving offensive contextual cues during extraction.

The optimal solution is obtained as Eq. 5.20.

$$x^* = \arg\max_{x \in \mathscr{P}} \mathscr{S}(x) \tag{5.20}$$

The final keyword subset is therefore represented as Eq. 5.21.

$$\mathscr{K}_{\text{off}} = \{w_i \in \mathscr{W}' \mid x_i^* = 1\} \tag{5.21}$$

A final refinement stage additionally injects missing high-toxicity concepts from the expanded lexicon whenever necessary while removing weakly associated terms based on similarity with the offensive anchor embedding.

The resulting keyword subset $\mathscr{K}_{\text{off}}$ provides a compact yet information-rich representation of the meme content by jointly preserving diversity, contextual relevance, and offensive cues. These keywords subsequently serve as high-level semantic indicators for downstream multimodal offense analysis.

5.2.4 Algorithmic Summary

For clarity, Algorithm 7 summarizes the workflow of the proposed multimodal offensive-aware keyword extraction framework, comprising multimodal semantic acquisition, offensive lexicon expansion, semantic embedding generation, multi-objective optimization, and final keyword extraction. This chapter primarily focuses on the keyword extraction component of the framework, whereas the downstream application toward offensive meme detection and classification is discussed separately in the application chapter.

5.3 Conclusion

This chapter extended keyword extraction beyond conventional textual and social media content toward multimodal meme analysis, where semantic meaning emerges through complex interactions between visual and textual modalities. Unlike unimodal content, memes frequently convey implicit intent, symbolic associations, sarcasm, and culturally contextual information that cannot be adequately captured through isolated textual or visual analysis. To address these challenges, the chapter introduced a multimodal offensive-aware keyword extraction framework capable of jointly modeling scene semantics, embedded textual information, and

Algorithm 7 Stage-wise Schematic of the Proposed Multimodal Offensive-Aware Keyword Extraction Framework

Input: Meme image $\mathscr{M}$, offensive seed set $\mathscr{S}$
Output: Offensive-aware keyword set $\mathscr{K}_{\text{off}}$
Stage 1: Multi-modal Semantic Acquisition
Generate caption representation ψ_{blip}
Extract OCR representation ψ_{ocr}
Compute visual semantic representation ψ_{vis}
Construct unified representation Ψ
Stage 2: Offensive Lexicon Expansion
Expand seed vocabulary $\mathscr{S} \rightarrow \mathscr{L}_{\text{off}}$
Apply toxicity-based filtering
Compute offensive semantic centroid μ_{off}
Stage 3: Candidate Representation
Concatenate semantic descriptors to form D
Extract candidate set $\mathscr{W} = \{w_1, w_2, \ldots, w_N\}$
Generate embedding matrix $\mathbf{E}$
Apply DBSCAN to obtain refined set $\mathscr{W}'$
Stage 4: Multi-objective Formulation
Encode solution vector $x \in \{0,1\}^N$
Define objective functions:
    f_1: redundancy minimization
    f_2: semantic coverage maximization
    f_3: offensive relevance maximization
Stage 5: NSGA-II Optimization
Initialize population $\mathscr{P}$
Perform non-dominated sorting
Apply crossover and mutation operations
Generate Pareto-optimal solutions
Select optimal solution x^*
Stage 6: Keyword Extraction
Construct offensive-aware keyword subset $\mathscr{K}_{\text{off}}$
return $\mathscr{K}_{\text{off}}$

social-contextual visual cues.

The proposed methodology integrated multiple semantic acquisition components, including BLIP-based scene captioning, OCR-driven textual extraction, and demographic-emotive visual analysis, to construct a unified multimodal semantic representation. This representation enabled the preservation of both explicit and implicit semantic relationships distributed across image and text modalities. To further enhance contextual understanding, an offensive lexicon expansion strategy based on ConceptNet was incorporated to capture semantically related offensive concepts and contextual associations extending beyond explicit toxic terminology.

Building upon these representations, the chapter formulated multimodal keyword extraction as

a constrained multi-objective optimization problem involving diversity preservation, representational coverage, and offensive affinity. An NSGA-II-based evolutionary optimization strategy was subsequently employed to identify semantically balanced keyword subsets capable of preserving contextual meaning while maintaining alignment with offensive discourse patterns. Through embedding-based representation learning, redundancy reduction, and Pareto-optimal optimization, the framework generated compact and contextually informative multimodal keyword representations suitable for downstream semantic analysis.

Overall, this chapter demonstrated that effective keyword extraction from multimodal meme content requires the integration of cross-modal semantics, contextual knowledge expansion, and optimization-driven selection strategies. The resulting offensive-aware keyword subsets provide semantically enriched representations that serve as the foundation for subsequent offensiveness modeling and multimodal classification. The application of these extracted keywords toward offensive meme detection and thermodynamic offense modeling is presented in Chapter 6.

Chapter 6

Downstream Applications of Keyword Extraction

The preceding chapters primarily focused on the development of keyword extraction methodologies across textual, social media, and multimodal content domains. These methodologies addressed the challenges of identifying semantically informative and contextually relevant keywords from diverse forms of unstructured data using statistical, semantic, optimization-driven, quantum-inspired, and multimodal learning strategies. While effective keyword extraction is itself an important research objective, the practical significance of extracted keywords is ultimately determined by their usefulness in downstream analytical and predictive tasks.

In many intelligent information processing systems, extracted keywords serve as an interpretable semantic layer between raw content and predictive models. Unlike high-dimensional latent representations that often lack transparency, keyword-based representations provide compact and human-interpretable semantic abstractions capable of preserving contextual meaning, semantic intent, and task-specific linguistic cues. Such representations become particularly valuable in applications involving short, noisy, and highly contextual social media content, where conventional feature engineering and purely statistical approaches frequently struggle to capture implicit semantics, sarcasm, ambiguity, and multimodal interactions.

This chapter therefore investigates the downstream applications of the proposed keyword extraction methodologies across multiple real-world content analysis tasks. The extracted key-

words are utilized as semantically enriched representations for improving predictive performance, robustness, and interpretability in socially relevant applications involving textual and multimodal data.

The downstream applications investigated in this chapter are summarized as follows:

- **QuToX: Toxicity Detection in Tweets**
  This study investigates the application of quantum-inspired keyword extraction for toxicity detection in noisy social media text. The extracted keywords are utilized to identify contextually significant toxic indicators and improve semantic understanding of implicit toxic discourse in tweets.

- **HumorKey: Humor Detection from Text**
  This study focuses on humor detection using semantically enriched keyword representations extracted from short textual content. The extracted keywords capture contextual incongruity, semantic shifts, and linguistic patterns associated with humorous expressions.

- **Multimodal Offensive Meme Detection**
  This study examines offensive meme detection using offensive-aware multimodal keywords extracted from integrated image-text representations. The extracted keywords preserve contextual, visual, and semantic offensive cues for downstream offensiveness modeling and classification.

Through these application studies, the chapter evaluates how semantically enriched keyword representations contribute toward improved classification performance, robustness against noisy and implicit content, and enhanced interpretability of downstream decision-making systems. Comparative evaluation, statistical significance analysis, and application-specific performance assessment are presented to demonstrate the practical effectiveness of the proposed keyword extraction methodologies across diverse real-world scenarios.

6.1 Toxicity Detection in Tweets

Following the quantum-inspired keyword extraction stage of the QuToX framework, the extracted semantic keywords are utilized for downstream toxicity classification of tweets. Instead

of relying on computationally intensive end-to-end deep learning architectures, the proposed approach adopts an interpretable and statistically grounded toxicity detection strategy based on lexicon-guided semantic modeling of keyword-level n-gram patterns. The objective of this stage is to determine whether the semantic information preserved within the extracted keyword subsets is sufficient for accurately identifying toxic and non-toxic tweet content.

Figure 6.1 illustrates the downstream toxicity detection component of the proposed QuToX framework. The extracted quantum-inspired semantic keywords are transformed into statistically informative lexical patterns, which are subsequently utilized for toxicity-oriented semantic modeling and final tweet classification.

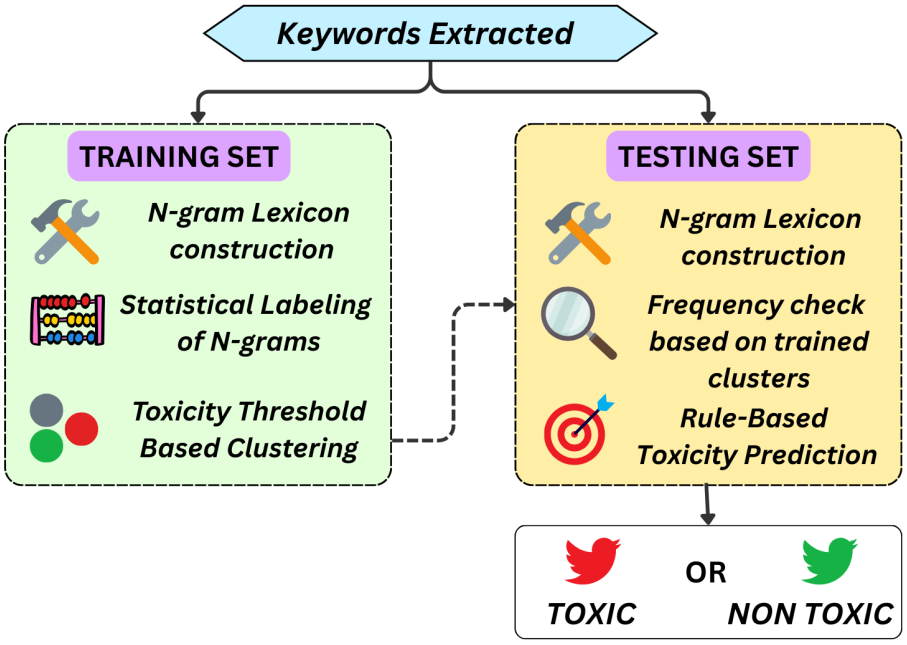

Figure 6.1: Downstream toxicity detection framework of QuToX using quantum-inspired semantically enriched keyword representations extracted from tweets.

In the proposed framework, the lexicon represents a structured collection of statistically informative n-grams associated with different toxicity categories. These n-grams are derived from the extracted keywords and grouped according to their statistical association with toxic, non-toxic, or neutral tweet semantics observed within the training corpus. Such a representation enables the framework to preserve contextual semantic information while maintaining interpretability of the downstream classification process.

Unlike conventional toxicity detection methods that frequently operate as black-box predictive systems, the proposed methodology emphasizes transparency and semantic interpretability. By utilizing semantically enriched keyword representations generated through quantum

interaction modeling, the framework captures contextual toxic indicators, implicit semantic relationships, and interaction-driven linguistic patterns commonly observed in noisy social media discourse.

To enable empirical evaluation, the keyword-labeled tweet dataset is partitioned into training and testing subsets using a stratified sampling strategy. Approximately 70% of the dataset is allocated to the training subset D_{train} for lexicon construction and statistical learning of toxicity-associated n-gram behavior, whereas the remaining 30% is reserved as the testing subset D_{test} for downstream evaluation. The stratified partitioning preserves balanced class distributions across both subsets and supports unbiased performance assessment of the proposed toxicity detection framework.

6.1.1 Lexicon Construction Using N-gram Expansion

After obtaining the semantically informative keyword subsets through the QuToX framework, the next stage focuses on transforming these extracted keywords into a structured lexical representation suitable for downstream toxicity prediction. Although individual keywords preserve important semantic cues, isolated lexical units often fail to fully capture local contextual dependencies and phrase-level linguistic behavior commonly observed in social media discourse. Toxic expressions in tweets frequently emerge through combinations of words rather than through single tokens alone. Consequently, the proposed approach extends the extracted keyword sequences into n-gram structures in order to preserve contextual phrase information associated with toxic and non-toxic communication patterns.

Let $\mathscr{K}_i = \{k_1, k_2, \ldots, k_n\}$ denote the ordered set of extracted keywords corresponding to the ith tweet. From this sequence, the framework generates all possible unigram, bigram, and trigram combinations using Eq. 6.1.

$$\text{NGram}(\mathscr{K}_i) = \bigcup_{n=1}^{3} \{\text{concat}(k_j, \ldots, k_{j+n-1}) \mid 1 \leq j \leq |\mathscr{K}_i| - n + 1\} \tag{6.1}$$

The operator concat$(\cdot)$ concatenates adjacent keywords to form phrase-level lexical units. Through this process, the extracted keywords are transformed into linguistically richer structures capable of preserving short-range contextual relationships.

The inclusion of multiple n-gram levels serves several important purposes. Unigrams preserve

isolated semantic indicators, whereas bigrams and trigrams capture contextual word interactions that may convey sarcasm, aggression, implicit abuse, or conversational hostility. In many social media scenarios, individual keywords may appear semantically harmless in isolation, while their sequential combinations reveal strong toxicity-oriented intent. Therefore, the n-gram expansion stage enables the methodology to preserve localized contextual structure that would otherwise be lost under independent keyword analysis.

Furthermore, the use of n-gram representations improves the robustness of the toxicity modeling process against the highly irregular linguistic characteristics of social media text, including abbreviations, fragmented syntax, slang, informal grammar, and context-dependent expressions.

6.1.2 Statistical Labeling of N-grams

Following n-gram generation, the framework constructs a statistical toxicity lexicon using the training subset D_{train}. The objective of this stage is to determine the degree to which each n-gram exhibits discriminatory association with toxic or non-toxic tweet categories.

For every unique n-gram $g \in \text{NGram}(\mathscr{K}_i)$ the framework computes its class-conditional occurrence frequencies across the training corpus.

Two frequency statistics are maintained for each n-gram:

- $c_T(g)$: number of occurrences of g in tweets labeled as toxic $(y_i = 1)$,
- $c_{NT}(g)$: number of occurrences of g in tweets labeled as non-toxic $(y_i = 0)$.

These statistics quantify how frequently a particular lexical pattern appears within each toxicity category. However, raw occurrence frequencies alone are insufficient for reliable discrimination because frequently occurring n-grams may appear across both classes with different proportions. Therefore, the framework employs a normalized differential frequency formulation to estimate the toxicity orientation of each n-gram.

The toxicity association score is defined using Eq. 6.2.

$$\text{toxicity}(g) = \frac{c_T(g) - c_{NT}(g)}{c_T(g) + c_{NT}(g)} \in [-1, 1] \tag{6.2}$$

It produces a bounded normalized score within the interval $[-1, 1]$, thereby enabling consistent comparison across n-grams with varying occurrence frequencies.

An n-gram receiving a high positive score occurs substantially more frequently within toxic tweets than within non-toxic discourse, indicating strong discriminatory relevance for toxicity identification. Conversely, strongly negative values correspond to n-grams predominantly associated with non-toxic communication. Scores approaching zero indicate semantically ambiguous patterns that exhibit relatively balanced occurrence across both classes.

The normalization term in the denominator plays an important role in preventing highly frequent n-grams from dominating the scoring process solely due to occurrence volume. Instead, the formulation emphasizes relative class-oriented behavior rather than absolute frequency magnitude.

Unlike manually curated toxicity vocabularies, the proposed lexicon construction process is entirely data-driven. Consequently, the generated lexical representation automatically captures domain-specific toxicity patterns emerging within the training corpus, including informal slang, abbreviated expressions, implicit insults, sarcastic constructions, and evolving social media terminology.

6.1.3 Threshold-Based Toxicity Clustering

Although the toxicity association score provides an estimate of class-oriented linguistic behavior, many n-grams exhibit weak or uncertain polarity due to contextual ambiguity or insufficient discriminatory strength. Directly incorporating such unstable patterns into the classification process may introduce noise and reduce predictive reliability. Therefore, the proposed methodology groups the generated n-grams into toxicity-oriented lexical clusters using a threshold-based decision strategy.

Let δ denote a predefined polarity threshold controlling the minimum discriminatory strength required for toxicity assignment. Using this threshold, each n-gram is assigned to one of three lexical categories according to Eq. 6.3.

$$\begin{aligned} \text{Toxic Cluster } (C_T) &\xrightarrow{\text{if}} \text{toxicity}(g) > \delta \\ \text{Non-Toxic Cluster } (C_{NT}) &\xrightarrow{\text{if}} \text{toxicity}(g) < -\delta \\ \text{Neutral Cluster } (C_0) &\xrightarrow{\text{if}} -\delta \leq \text{toxicity}(g) \leq \delta \end{aligned} \tag{6.3}$$

Under this formulation, strongly toxicity-oriented n-grams are assigned to the toxic cluster C_T, whereas strongly non-toxic expressions are grouped within C_{NT}. N-grams whose polarity

remains uncertain or statistically weak are assigned to the neutral cluster C_0.

The introduction of the neutral cluster serves an important stabilizing function within the proposed methodology. Rather than forcing ambiguous lexical patterns into hard toxicity categories, the framework isolates uncertain expressions to reduce noise propagation during downstream prediction. This mechanism improves classification robustness, particularly in social media environments where many phrases exhibit context-dependent meaning.

The complete lexical repository is represented as $\mathscr{L} = \{C_T, C_{NT}, C_0\}$, where each cluster contains n-grams exhibiting statistically similar toxicity-oriented behavior.

In the proposed methodology, the lexicon therefore acts as a structured statistical rule base linking lexical occurrence patterns with toxicity polarity. Since every toxicity decision can be directly traced back to explicit n-gram evidence and class-specific statistical behavior, the resulting prediction mechanism remains transparent, interpretable, and analytically explainable.

6.1.4 Rule-Based Toxicity Prediction

After constructing the toxicity-oriented lexical repository, the framework performs classification of unseen tweets using a rule-based voting mechanism grounded in lexical cluster matching.

Consider an unseen tweet represented through its extracted keyword subset $\mathscr{K}_j$. The framework first generates the corresponding n-gram representation $\text{NGram}(\mathscr{K}_j)$. These generated n-grams are subsequently matched against the previously constructed toxicity lexicon. The number of lexical matches belonging to the toxic cluster is computed as Eq. 6.1.

$$h_T = \left|\text{NGram}(\mathscr{K}_j) \cap C_T\right| \tag{6.4}$$

Similarly, the number of matches within the non-toxic cluster is defined as Eq. 6.5.

$$h_{NT} = \left|\text{NGram}(\mathscr{K}_j) \cap C_{NT}\right| \tag{6.5}$$

These quantities represent the degree to which the semantic content of the tweet aligns with toxicity-oriented and non-toxicity oriented lexical behavior, respectively.

The final toxicity label is then determined using the majority voting rule defined in Eq. 6.6.

$$\hat{y}_j = \begin{cases} 1, & \text{if } h_T > h_{NT} \\ 0, & \text{otherwise} \end{cases} \tag{6.6}$$

Under this decision strategy, a tweet is classified as toxic whenever toxicity-oriented lexical patterns dominate the extracted semantic representation. Conversely, tweets containing stronger alignment with non-toxic lexical patterns are assigned to the non-toxic category.

Importantly, neutral lexical patterns belonging to C_0 are excluded from the voting process. This exclusion prevents semantically ambiguous expressions from diluting the classification boundary and further improves prediction stability.

The proposed classification methodology therefore combines quantum-inspired semantic keyword extraction with statistically interpretable lexical decision modeling. Unlike opaque black-box architectures, the resulting toxicity detection process remains computationally efficient, transparent, and analytically explainable while still preserving the ability to capture implicit toxicity patterns within noisy social media discourse.

6.1.5 Toxicity Detection Evaluation

After extracting semantically informative keywords using the proposed quantum inspired framework, the next stage evaluates the effectiveness of these keywords for downstream toxicity detection. The primary objective of this evaluation is to examine whether the extracted keyword representations preserve sufficient contextual and discriminatory information for reliable identification of toxic and non-toxic tweet content.

Unlike conventional toxicity detection approaches that depend heavily on large supervised architectures and dense end-to-end training pipelines, the proposed methodology adopts a statistically grounded and interpretable classification strategy based on lexical toxicity modeling. Instead of directly utilizing raw tweet content, the framework operates on semantically enriched keyword representations and their contextual n-gram structures. This design enables the toxicity prediction process to remain transparent while still preserving important contextual linguistic patterns associated with abusive or supportive discourse.

To construct the toxicity-aware lexical model, the extracted keyword sequences are expanded into unigram, bigram, and trigram representations. These n-gram structures are subsequently analyzed over the training corpus to estimate their empirical association with toxic and non-

toxic tweet categories. Based on their statistical toxicity orientation, each n-gram is assigned to one of three semantic clusters:

- toxic,
- neutral,
- non-toxic.

The clustering process is guided using a toxicity threshold parameter of $\delta = 0.55$, which determines the minimum discriminatory strength required for toxicity assignment. The resulting toxicity-aware lexicon therefore captures recurring linguistic structures distributed across different toxicity levels and serves as the foundation of the downstream rule-based classification framework.

Figure 6.2 illustrates the distribution of n-grams according to their computed toxicity scores and occurrence frequencies. The visualization provides insight into the lexical behavior learned by the proposed framework and highlights the separation between toxic, neutral, and non-toxic linguistic regions.

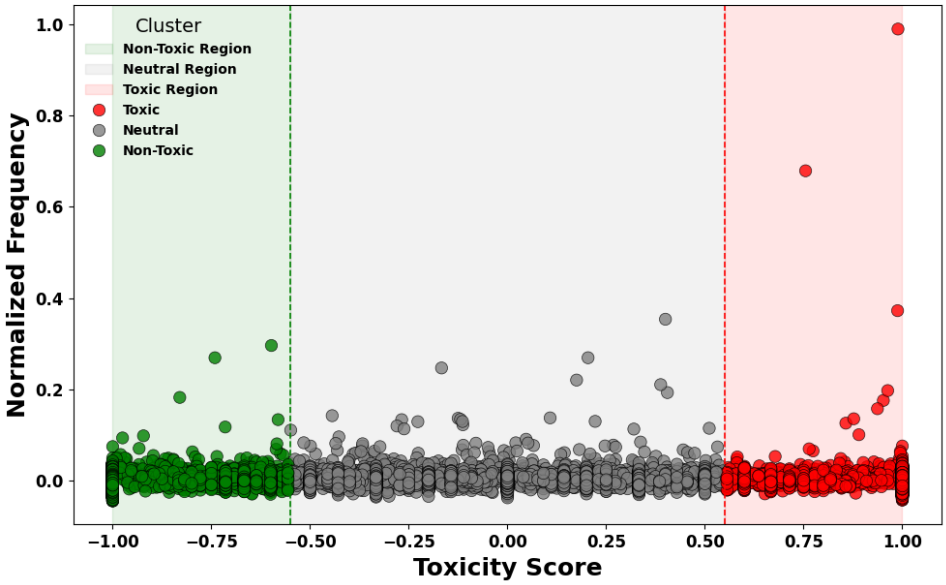

Figure 6.2: Distribution of n-gram clusters based on computed toxicity scores and occurrence frequency within the tweet corpus.

As illustrated in Figure 6.2, the majority of the generated n-grams are concentrated within the neutral region, indicating the presence of linguistically ambiguous or context-dependent expressions commonly observed in social media discourse. In contrast, n-grams positioned near

the extreme toxicity ranges exhibit strong discriminatory behavior and are more consistently associated with either toxic or non-toxic communication patterns. Several highly frequent n-grams additionally appear near the polarity extremes, suggesting that certain offensive or supportive linguistic expressions repeatedly occur within the dataset and therefore contribute strongly toward toxicity discrimination.

To further examine the semantic characteristics preserved by the extracted keywords, word cloud visualizations are generated separately for toxic and non-toxic lexical clusters. These visualizations qualitatively illustrate the dominant linguistic patterns learned by the framework and provide interpretive insight into the contrasting semantic behavior of toxic and non-toxic discourse.

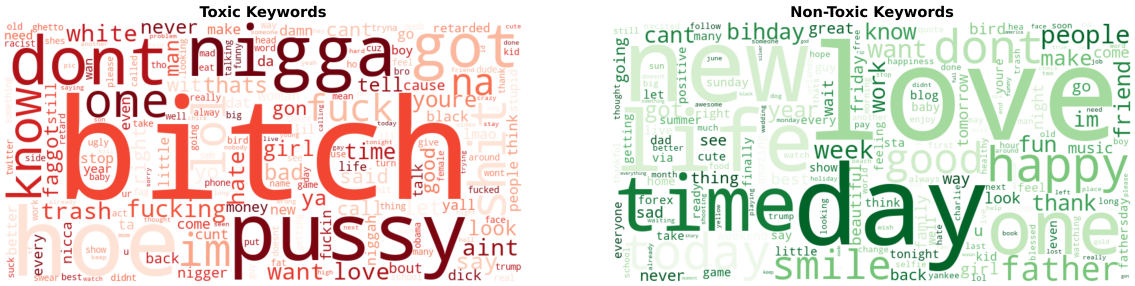

Figure 6.3: Word cloud visualization of dominant keywords present within (a) toxic lexical clusters and (b) non-toxic lexical clusters.

Figure 6.3 demonstrates clear semantic distinctions between the two lexical categories. The toxic word cloud predominantly contains aggressive, abusive, and offensive expressions frequently associated with hostile online interactions, whereas the non-toxic cluster highlights supportive, neutral, and conversational linguistic patterns. Together, these visualizations provide qualitative evidence that the proposed quantum-inspired keyword extraction methodology successfully preserves toxicity-oriented semantic structures necessary for downstream tweet classification.

6.1.6 Toxicity Threshold (δ) Sensitivity Analysis

The toxicity threshold parameter δ plays an important role in the proposed lexicon-driven classification framework, as it directly controls the assignment of n-grams into toxic, neutral, and non-toxic lexical clusters. Since the threshold determines the minimum discriminatory strength required for toxicity-oriented categorization, its selection significantly influences the balance between sensitivity and classification selectiveness during downstream prediction.

Therefore, to empirically justify the choice of the threshold value used in the proposed methodology, a detailed sensitivity analysis is performed by systematically varying δ and observing its impact on standard evaluation metrics.

The analysis evaluates the variation in Accuracy, Precision, Recall, and F1-Score across different threshold values, as illustrated in Figure 6.4. This experiment provides insight into the behavioral characteristics of the proposed toxicity detection framework under varying levels of classification strictness.

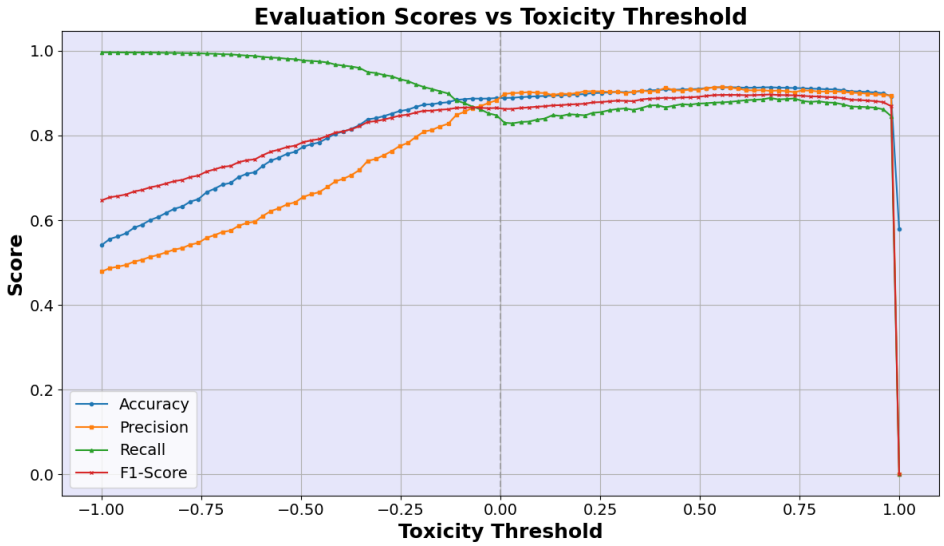

Figure 6.4: Variation of evaluation metrics with respect to the toxicity threshold parameter δ, illustrating the optimal performance balance near $\delta = 0.55$.

As observed in Figure 6.4, lower threshold values produce highly permissive toxicity assignments, causing a larger proportion of n-grams to be categorized within the toxic cluster. Under such conditions, the framework exhibits high Recall since the majority of toxic instances are successfully identified. However, this increased sensitivity simultaneously introduces a substantial number of false positive predictions, thereby reducing both Precision and overall classification Accuracy.

As the threshold value gradually increases, the classification process becomes increasingly selective. Consequently, Precision and Accuracy improve because only strongly toxicity oriented n-grams are retained within the toxic cluster. This stricter lexical filtering reduces false positive predictions and improves the reliability of toxicity assignments. However, the increased selectiveness also causes a gradual decline in Recall, indicating that certain borderline toxic expressions are no longer captured within the toxicity-oriented lexical space.

The behavior of the F1-Score further illustrates the trade-off between sensitivity and predic-

tion reliability. Initially, the F1-Score increases steadily as Precision improves while Recall remains relatively high. Beyond a certain threshold range, however, the reduction in Recall outweighs the gains in Precision, causing the F1-Score to decline. This behavior indicates diminishing returns associated with excessively restrictive toxicity thresholds.

The evaluation curves additionally demonstrate that the proposed methodology maintains relatively stable performance within the threshold interval approximately spanning 0.5 to 0.7. This stability suggests that the lexical clustering strategy remains robust across moderate variations in toxicity polarity assignment. Among these values, the threshold $\delta = 0.55$ provides the most balanced trade-off between sensitivity and selectiveness, yielding consistently strong performance across all evaluation metrics.

Therefore, based on the observed experimental behavior, the threshold value $\delta = 0.55$ is selected as the optimal operating point for the proposed toxicity detection framework. A more detailed evaluation of the classification performance corresponding to this threshold configuration is presented in the subsequent experimental analysis section.

6.1.7 Experimental Results

To comprehensively evaluate the effectiveness of the proposed toxicity detection framework, extensive experiments are conducted using the toxicity-aware lexical classification strategy derived from the quantum-inspired QuToX keyword extraction methodology. The objective of this evaluation is not only to measure predictive performance, but also to examine the robustness, reliability, and interpretability of the proposed lexicon-driven toxicity modeling approach under realistic social media conditions.

In the proposed methodology, toxicity classification decisions are generated through majority voting over the toxicity-oriented lexical clusters formed using statistically labeled n-grams. The toxicity threshold parameter was empirically fixed at $\delta = 0.55$ based on the sensitivity analysis presented in the previous subsection. This threshold configuration provided the most balanced trade-off between prediction sensitivity and classification reliability across multiple evaluation metrics.

To provide an intuitive understanding of the classification behavior of the proposed framework, the confusion matrix corresponding to the final prediction results is first analyzed. The confusion matrix offers a direct visualization of the relationship between actual tweet labels and predicted toxicity assignments across both toxic and non-toxic categories.

The confusion matrix is composed of four fundamental classification outcomes:

- **True Positives (TP):** Toxic tweets correctly identified as toxic.
- **True Negatives (TN):** Non-toxic tweets correctly classified as non-toxic.
- **False Positives (FP):** Non-toxic tweets incorrectly classified as toxic.
- **False Negatives (FN):** Toxic tweets incorrectly classified as non-toxic.

These quantities collectively form the basis for all subsequent evaluation metrics and provide detailed insight into the strengths and limitations of the proposed classification framework.

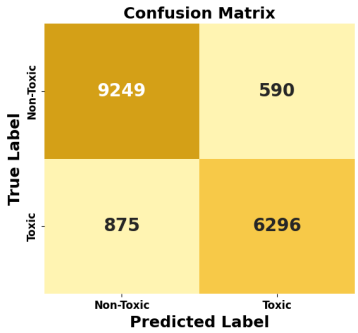

Figure 6.5: Confusion matrix illustrating the toxicity classification performance of the proposed QuToX framework.

As illustrated in Figure 6.5, the proposed framework demonstrates strong discriminatory capability across both toxicity classes. A substantial proportion of non-toxic tweets are correctly classified, while only a relatively small fraction of benign tweets are incorrectly flagged as toxic. This behavior indicates that the proposed lexical toxicity modeling approach effectively suppresses false positive predictions and maintains reliable classification boundaries.

Similarly, the framework successfully identifies the majority of toxic tweet instances, demonstrating strong sensitivity toward harmful or abusive content. The relatively low number of false negatives further indicates that semantically informative keywords extracted through the quantum-inspired interaction modeling process preserve important toxicity-oriented contextual cues necessary for downstream prediction.

Overall, the confusion matrix highlights the robustness of the proposed framework in handling noisy and imbalanced social media data while maintaining balanced predictive behavior across both classes.

6.1.7.1 Evaluation Metrics

To rigorously assess the performance characteristics of the proposed toxicity detection framework, multiple complementary evaluation metrics are employed. Since social media toxicity datasets frequently exhibit class imbalance and varying error costs, reliance on a single evaluation measure may provide an incomplete assessment of classification behavior. Therefore, the proposed methodology is evaluated using a diverse set of metrics capturing predictive accuracy, class discrimination capability, robustness, calibration quality, and agreement reliability.

All evaluation metrics are derived from the confusion matrix components TP, TN, FP, and FN.

Accuracy Accuracy measures the overall proportion of correctly classified tweet instances across both toxicity categories. It is defined using Eq. 6.7.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.7}$$

Although Accuracy provides a global estimate of classification correctness, it may become less informative under highly imbalanced class distributions.

Precision Precision evaluates the reliability of toxic predictions by measuring the proportion of tweets predicted as toxic that are actually toxic. It is defined using Eq. 6.8.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6.8}$$

High Precision indicates low false positive behavior and reflects the ability of the framework to avoid incorrectly flagging benign content as toxic.

Recall Recall, also referred to as Sensitivity, measures the proportion of actual toxic tweets that are successfully detected by the framework. It is defined using Eq. 6.9.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6.9}$$

This metric is particularly important in toxicity detection scenarios because failure to identify harmful content may lead to unsafe online interactions and ineffective moderation.

F1-Score The F1-Score provides a harmonic balance between Precision and Recall and is especially useful when both false positives and false negatives carry practical importance. It is defined using Eq. 6.10.

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6.10}$$

A high F1-Score indicates balanced predictive performance across both error categories.

Specificity Specificity evaluates the ability of the framework to correctly identify non-toxic tweet instances. It is defined using Eq. 6.11.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{6.11}$$

This metric is important for verifying that the proposed methodology does not excessively over-classify benign discourse as toxic.

Matthews Correlation Coefficient (MCC) The Matthews Correlation Coefficient provides a balanced evaluation measure incorporating all four confusion matrix components simultaneously. It is defined using Eq. 6.12.

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{6.12}$$

MCC values range from -1 to 1, where values approaching 1 indicate strong agreement between predictions and ground truth labels.

Cohen's Kappa Cohen's Kappa measures classification agreement while correcting for agreement occurring due to chance. It is defined using Eq. 6.13.

$$\mathscr{K} = \frac{p_0 - p_e}{1 - p_e} \tag{6.13}$$

where:

- p_0 represents the observed agreement,
- p_e denotes the expected agreement under random assignment.

Higher Kappa values indicate stronger prediction consistency beyond chance-level behavior.

ROC AUC The Receiver Operating Characteristic Area Under Curve (ROC AUC) evaluates the discriminative capability of the classifier across varying decision thresholds. It is defined using Eq. 6.14.

$$\text{ROC AUC} = \int_0^1 \text{Recall}(FPR)\, d(FPR) \tag{6.14}$$

where:

$$FPR = \frac{FP}{FP + TN} \tag{6.15}$$

A larger ROC AUC value indicates stronger class separability.

PR AUC The Precision–Recall Area Under Curve (PR AUC) measures the trade-off between Precision and Recall across varying classification thresholds. It is defined using Eq. 6.16.

$$\text{PR AUC} = \int_0^1 \text{Precision}(\text{Recall})\, d(\text{Recall}) \tag{6.16}$$

PR AUC is particularly informative under imbalanced toxicity datasets because it emphasizes minority-class prediction behavior.

Brier Score The Brier Score evaluates the calibration quality of probabilistic predictions generated by the classifier. It is defined using Eq. 6.17.

$$\text{Brier Score} = \frac{1}{N} \sum_{i=1}^{N} (p_i - y_i)^2 \tag{6.17}$$

where:

- p_i represents the predicted probability,
- y_i denotes the ground truth label.

Lower Brier Scores indicate better-calibrated and more reliable prediction confidence estimates.

6.1.7.2 Results

The evaluation metrics obtained using the proposed QuToX-based toxicity detection framework are presented in Table 6.1. The results demonstrate consistently strong performance

across multiple complementary evaluation criteria.

Table 6.1: Evaluation Metrics for Toxicity Classification Model

Metric	Value	Interpretation
Accuracy	91.39%	High overall correctness across both classes.
Precision	91.43%	Model is precise when predicting toxic tweets.
Recall	87.80%	Most toxic tweets are successfully detected.
F1-Score	89.58%	Balanced metric accounting for both FP and FN.
Specificity	94.00%	High capability to correctly identify non-toxic tweets.
MCC	≈ 0.82	Strong correlation between prediction and truth.
Cohen's Kappa	≈ 0.82	High agreement between model and ground truth beyond chance.
ROC AUC	0.909	Excellent class separability.
PR AUC	0.8542	Strong precision-recall trade-off, especially relevant for toxicity detection.
Brier Score	0.0861	Indicates confident and well-calibrated predictions.

To further analyze the robustness of the proposed methodology under imbalanced toxicity distributions, ROC and Precision–Recall curve analyses are additionally performed.

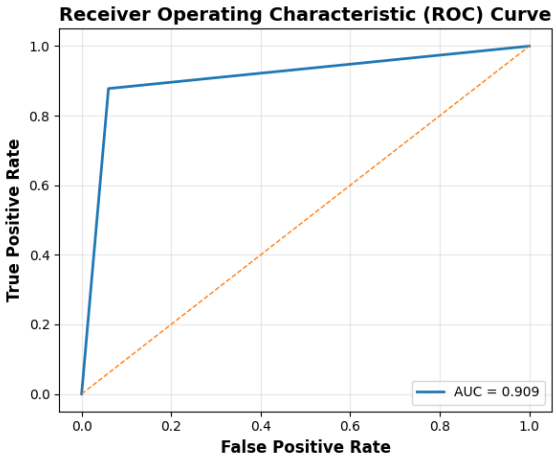

Figure 6.6: Receiver Operating Characteristic (ROC) curve of the proposed toxicity detection framework.

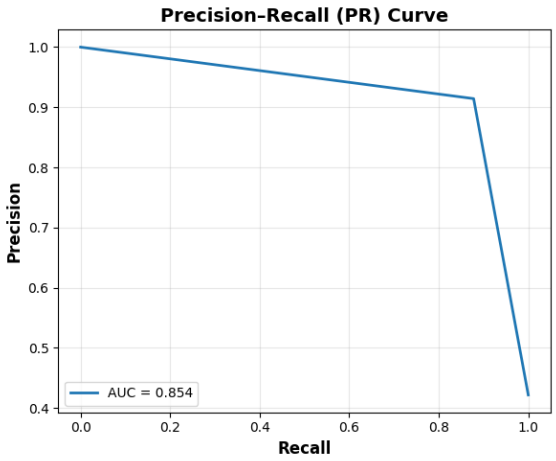

Figure 6.7: Precision–Recall (PR) curve of the proposed toxicity detection framework.

As observed in Figure 6.6, the proposed framework achieves a ROC AUC value of 0.909, indicating strong separability between toxic and non-toxic tweet classes across varying decision thresholds. The curve demonstrates that the proposed lexical toxicity modeling strategy preserves reliable discriminatory capability even under noisy social media conditions.

However, since ROC analysis may occasionally provide optimistic interpretation under class imbalance, the Precision–Recall curve shown in Figure 6.7 provides a more informative eval-

uation of minority toxicity detection behavior. The obtained PR AUC value of 0.854 demonstrates that the proposed framework maintains strong Precision while simultaneously achieving high Recall. This behavior is particularly desirable in toxicity detection systems where both false positive moderation and missed harmful content carry practical consequences.

Collectively, these results confirm that the proposed QuToX framework achieves reliable, balanced, and robust toxicity detection performance while preserving the transparency and interpretability advantages of lexicon-driven classification.

The experimental results demonstrate that the proposed QuToX framework achieves strong and balanced toxicity detection performance across multiple evaluation measures. The high Accuracy, Precision, Recall, and F1-Score indicate reliable identification of both toxic and non-toxic tweets while maintaining an effective balance between false positive and false negative predictions. Similarly, the strong MCC and Cohen's Kappa values confirm the statistical reliability and robustness of the framework under imbalanced social media conditions. The ROC AUC and PR AUC analyses further validate the model's discriminative capability across varying thresholds, whereas the low Brier Score indicates well-calibrated and reliable prediction confidence. Overall, these findings demonstrate that the proposed quantum-inspired semantic keyword extraction and lexicon-based classification strategy provides an effective, interpretable, and computationally efficient framework for toxicity detection in noisy social media environments.

6.1.8 Comparative Analysis

To evaluate the effectiveness of the proposed QuToX framework, a comparative analysis is conducted against several widely adopted toxicity detection approaches representing diverse methodological paradigms, including transformer-based language models, graph-enhanced semantic learning frameworks, multilingual transfer learning systems, unsupervised toxicity detection approaches, and traditional machine learning classifiers. All baseline models are implemented and evaluated on the same dataset under identical experimental settings to ensure fair and unbiased comparison.

The evaluated transformer-based baselines include BERT [163], BERTweet [164], DeBERTa [165], and BioBERT [166]. Additional comparisons are performed against the multilingual offensive language detection framework proposed in [167], the graph-based GraphBERT model [168], and the unsupervised HateCircle framework [169]. These models were selected due to

their previously reported effectiveness in toxicity and offensive language detection tasks.

The comparative performance results obtained using standard evaluation metrics are presented in Table 6.2.

Table 6.2: Evaluation Metrics for Toxicity Classification Model

Method	**Accuracy**	**Precision**	**Recall**	**F1 Score**
[163]	0.5407	0.5081	0.5935	0.5475
[164]	0.8231	0.8147	0.7792	0.7965
[165]	0.873	0.8411	0.868	0.8543
[166]	0.8968	0.8476	0.751	0.7962
[167]	0.7821	0.8	0.6506	0.7176
[168]	0.8855	0.8998	0.8225	0.8594
[169]	0.5187	0.4476	0.5584	0.4969
QuTox	0.9139	0.9143	0.878	0.8958

The results demonstrate that the proposed QuToX framework consistently outperforms all comparative baselines across all major evaluation metrics. In particular, the proposed methodology achieves the highest Accuracy of 91.39%, Precision of 91.43%, Recall of 87.80%, and F1-Score of 89.58%, indicating balanced and reliable toxicity detection performance.

Among the baseline methods, DeBERTa and GraphBERT achieve comparatively strong results due to their enhanced contextual and structural representation learning capabilities. BERTweet additionally performs better than standard BERT because of its adaptation to Twitter-specific linguistic patterns. In contrast, the multilingual and unsupervised baselines demonstrate comparatively lower performance, primarily due to domain mismatch and reduced adaptation to noisy English social media discourse.

To further examine the effectiveness of the proposed framework, an additional comparison is conducted against traditional machine learning classifiers, including Support Vector Machine (SVM), Naïve Bayes, Logistic Regression, Decision Tree, and Random Forest.

As illustrated in Figure 6.8, the proposed QuToX framework substantially outperforms conventional machine learning approaches across all evaluated metrics. Traditional classifiers primarily rely on surface-level lexical features such as Bag-of-Words or TF-IDF representations, which often fail to preserve contextual semantics, implicit toxicity, sarcasm, and discourse-level linguistic relationships commonly observed in social media text.

In contrast, the proposed QuToX framework combines quantum-inspired semantic interaction modeling with contextual keyword extraction, enabling improved preservation of toxicity-oriented linguistic structure while maintaining interpretability and computational efficiency.

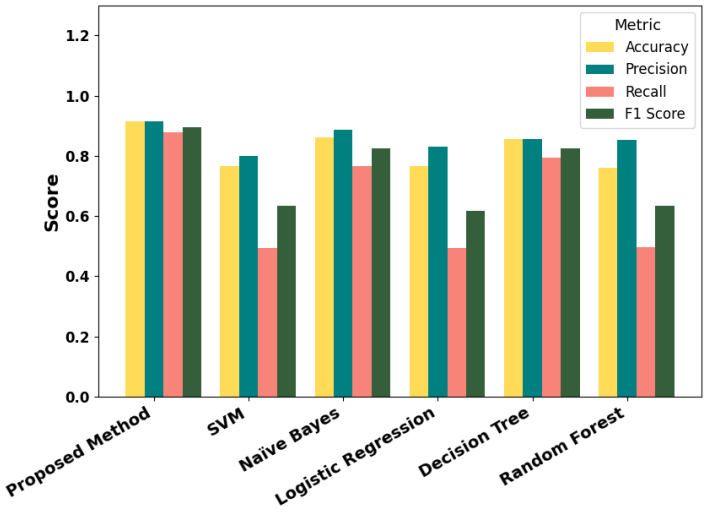

Figure 6.8: Comparative evaluation of the proposed QuToX framework against traditional machine learning classifiers.

Overall, the comparative analysis demonstrates that the proposed methodology provides a robust and effective approach for toxicity detection in noisy social media environments.

6.1.9 Ten-Fold Cross-Validation

To further evaluate the robustness and generalization capability of the proposed QuToX framework, a stratified ten-fold cross-validation experiment is conducted. Cross-validation is widely employed in machine learning evaluation because it provides a reliable estimate of model performance by systematically varying the training and validation partitions across multiple experimental iterations.

In the proposed evaluation setup, the complete dataset D is partitioned into $k = 10$ approximately equal subsets $D = \{D_1, D_2, \ldots, D_{10}\}$.

For each iteration $i \in \{1, 2, \ldots, 10\}$, the model is trained using the union of nine subsets while the remaining subset D_i is used for validation. This process is repeated ten times such that every subset serves exactly once as the validation fold. The evaluation metrics obtained from all folds are subsequently averaged to estimate the overall predictive stability and generalization performance of the framework.

Table 6.3 presents the per-fold evaluation results using Accuracy, Precision, Recall, and F1-Score.

The results demonstrate that the proposed framework maintains highly consistent performance

Table 6.3: Per-fold evaluation metrics obtained using ten-fold cross-validation

Fold	Accuracy	Precision	Recall	F1
1	0.9157	0.9155	0.8836	0.8992
2	0.9155	0.9150	0.8836	0.8990
3	0.9206	0.9189	0.8923	0.9054
4	0.9157	0.9158	0.8832	0.8992
5	0.9156	0.9172	0.8811	0.8988
6	0.9141	0.9148	0.8803	0.8972
7	0.9208	0.9274	0.8832	0.9047
8	0.9145	0.9135	0.8829	0.8979
9	0.9102	0.9080	0.8783	0.8929
10	0.9173	0.9227	0.8795	0.9006

across all validation folds. Accuracy values vary only marginally between 0.9102 and 0.9208, indicating stable classification behavior under different training-validation partitions. Similarly, Precision and Recall remain consistently high across all folds, reflecting balanced toxicity detection capability with controlled false positive and false negative behavior. The F1-Scores additionally remain concentrated near the 0.90 range, confirming stable overall predictive performance.

The low variation observed across the evaluation folds indicates that the proposed QuToX framework generalizes effectively to unseen tweet data and does not exhibit significant sensitivity to dataset partitioning. Overall, the cross-validation analysis demonstrates the robustness, reliability, and scalability of the proposed toxicity detection methodology for real-world social media environments.

6.1.10 Statistical Performance Analysis

To further validate the effectiveness and reliability of the proposed QuToX framework, statistical significance analysis is performed using two complementary strategies: the Two-Tailed Z-Test and Bootstrapping-based evaluation. These analyses are conducted to determine whether the observed improvement in toxicity detection performance is statistically significant and consistently reproducible across varying experimental conditions.

6.1.10.1 Two-Tailed Z-Test Statistical Analysis

The Two-Tailed Z-Test is employed to examine whether the F1-Score achieved by the proposed QuToX framework differs significantly from the average performance of the comparative baseline methods. The null and alternative hypotheses are defined as in Eq. 6.61.

$$\begin{aligned} H_0 &: \mu_p = \mu_b \quad \text{(no significant difference)} \\ H_1 &: \mu_p \neq \mu_b \quad \text{(significant difference exists)} \end{aligned} \tag{6.18}$$

where μ_p denotes the F1-Score of the proposed framework and μ_b represents the mean F1-Score of the baseline methods.

The Z-statistic is computed using Eq. 6.62.

$$Z = \frac{\mu_p - \mu_b}{\sigma_b / \sqrt{n}} \tag{6.19}$$

where σ_b denotes the standard deviation of the baseline F1-Scores and n represents the number of baseline methods. A significance level of $\alpha = 0.05$ is adopted, corresponding to critical Z-values of ± 1.96.

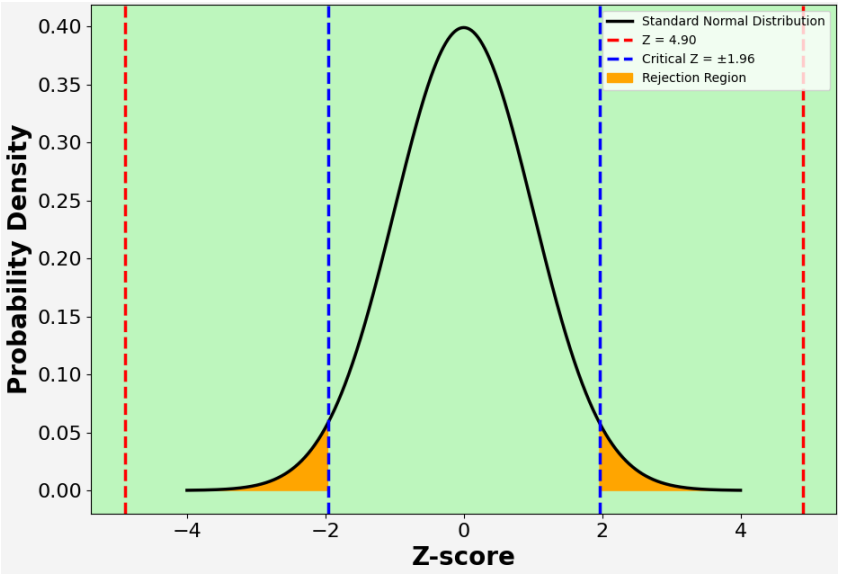

Figure 6.9: Standard normal distribution illustrating the Two-Tailed Z-Test and rejection regions.

Using the obtained results, the proposed framework achieves an F1-Score of $\mu_p = 0.8958$, whereas the mean baseline F1-Score is $\mu_b = 0.7338$ with a standard deviation of $\sigma_b = 0.1207$. This yields a Z-score of 4.8982 and a p-value of 9.67×10^{-7}.

As illustrated in Figure 6.9, the computed Z-score lies well beyond the critical threshold of ± 1.96, placing it within the rejection region. Consequently, the null hypothesis is rejected, indicating that the performance improvement achieved by the proposed QuToX framework is statistically significant and unlikely to have occurred due to random variation.

6.1.10.2 Bootstrapping-Based Statistical Evaluation

To further evaluate the stability and robustness of the proposed framework, bootstrapping-based statistical analysis is performed. Bootstrapping is a non-parametric resampling technique that estimates the sampling distribution of evaluation metrics without assuming any predefined data distribution.

Let $D = \{x_1, x_2, \ldots, x_n\}$ denote the set of F1-Scores obtained from multiple randomized evaluation splits. The bootstrapping process generates B resampled datasets $D_1^*, D_2^*, \ldots, D_B^*$ by sampling with replacement from the original score distribution.

For each resampled dataset D_b^*, the corresponding F1-Score is computed as Eq. 6.20.

$$\theta_b^* = \mathrm{F1}(D_b^*) \quad \text{for } b = 1, 2, \ldots, B \tag{6.20}$$

The 95% confidence interval is then estimated using percentile-based bootstrapping using Eq. 6.21.

$$\mathrm{CI}_{95\%} = \left[\theta^*_{(0.025B)}, \theta^*_{(0.975B)}\right] \tag{6.21}$$

In the proposed evaluation, $B = 1000$ bootstrapping iterations are performed.

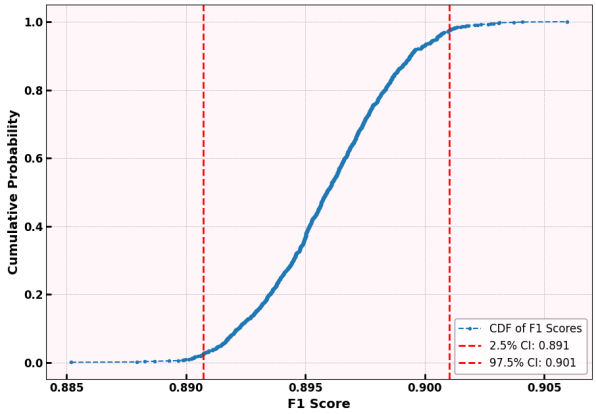

Figure 6.10: CDF of bootstrapped F1-Scores with estimated 95% confidence interval.

As observed in Figure 6.10, the bootstrapped F1-Score distribution exhibits approximately symmetric behavior, indicating stable performance across repeated resampling iterations. The estimated confidence interval is obtained as $\mathrm{CI}_{95\%} = [0.891,\ 0.901]$.

The narrow confidence interval and concentrated distribution around the median F1-Score indicate low performance variance and strong predictive consistency across varying data parti-

tions. Overall, the bootstrapping analysis confirms the robustness, stability, and generalization capability of the proposed QuToX toxicity detection framework.

The experimental evaluation, comparative analysis, cross-validation study, and statistical significance analysis collectively demonstrate that the proposed QuToX framework provides an effective, interpretable, and statistically reliable approach for toxicity detection in noisy social media environments.

6.2 Humor Detection from Text

This section presents the downstream application of the proposed HumorKey framework for humor detection in short textual content. Building upon the semantically enriched keyword extraction methodology discussed in the previous chapters, the extracted keywords are utilized as interpretable semantic representations for identifying humorous and non-humorous text instances. The objective of this application is to examine whether contextual keyword representations are capable of preserving the semantic incongruity, linguistic ambiguity, and contextual shifts commonly associated with humor in natural language.

Humor detection in short textual content is a challenging task due to the highly implicit and context-dependent nature of humorous expressions. Unlike conventional text classification problems, humor frequently emerges through semantic incongruity, double meanings, sarcasm, exaggeration, or unexpected contextual transitions. These characteristics make humor difficult to model using surface-level lexical features or frequency-based representations alone. Furthermore, short social media text often contains informal grammar, abbreviations, ambiguous phrasing, and limited contextual information, thereby increasing the complexity of computational humor understanding.

To address these challenges, the proposed HumorKey framework utilizes semantically informative keyword extraction to preserve contextual and linguistically relevant cues associated with humorous discourse. Rather than relying solely on dense black-box embeddings, the framework focuses on extracting compact semantic representations that maintain interpretability while capturing important humor-oriented linguistic structures.

Figure 6.11 illustrates the downstream humor detection framework, demonstrating how the extracted keyword representations are utilized for semantic modeling and final humor classification.

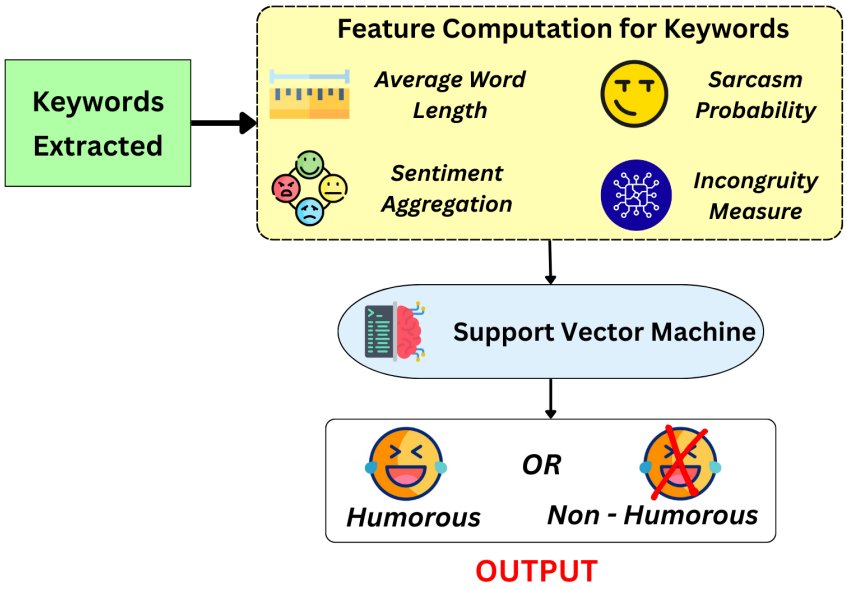

Figure 6.11: Downstream humor detection framework using semantically enriched keyword representations extracted through the proposed HumorKey methodology.

6.2.1 Feature Representation

After extracting the most semantically informative keywords using the proposed HumorKey framework, the next stage focuses on transforming these keywords into a structured feature representation suitable for downstream humor classification. Since humorous expressions frequently arise from complex interactions between lexical choice, emotional polarity, semantic incongruity, exaggeration, and pragmatic intent, the proposed methodology constructs a multi-dimensional feature space capable of capturing these diverse linguistic characteristics.

Let the complete token set corresponding to a text instance be represented as Eq. 6.22.

$$T = \{w_1, w_2, \ldots, w_n\} \tag{6.22}$$

From this set, the top five semantically relevant keywords are extracted and represented as Eq. 6.23.

$$K = \{k_1, k_2, k_3, k_4, k_5\} \tag{6.23}$$

where each keyword $k_i \in T$. These selected keywords are subsequently transformed into numerical features encoding lexical, affective, semantic, and pragmatic information. The resulting representation serves as the primary input for downstream humor classification.

6.2.1.1 Lexical Complexity Representation

Humorous text frequently contains exaggerated wording, unusual lexical constructions, or intentionally simplified expressions designed to enhance comedic effect. To capture such lexical variation, the average word length of the selected keywords is computed using Eq. 6.24.

$$L_{\text{avg}} = \frac{1}{|K|} \sum_{i=1}^{|K|} |k_i| \tag{6.24}$$

where $|K| = 5$ denotes the number of extracted keywords and $|k_i|$ represents the number of characters in keyword k_i.

This feature provides an estimate of the lexical complexity associated with the selected keyword set. Short and exaggerated expressions are frequently observed in informal humorous discourse, whereas unusually long or stylized words may indicate expressive emphasis, parody, or exaggerated narrative construction. Consequently, lexical length variation provides useful discriminatory information for humor-oriented language modeling.

6.2.1.2 Sentiment-Based Representation

Humorous expressions often exhibit strong emotional polarity, exaggerated affective tone, or sudden sentiment shifts. To capture these emotional characteristics, sentiment analysis is performed on each extracted keyword using the VADER sentiment analyzer. Each keyword k_i is assigned a normalized sentiment score represented as Eq. 6.25.

$$s(k_i) \in [0, 1] \tag{6.25}$$

Two complementary sentiment-oriented features are subsequently computed.

The mean sentiment score is calculated using Eq. 6.26.

$$S_{\text{mean}} = \frac{1}{|K|} \sum_{i=1}^{|K|} s(k_i) \tag{6.26}$$

which captures the overall affective tendency of the extracted keyword set.

In addition, the maximum sentiment score is computed using Eq. 6.27.

$$S_{\text{max}} = \max_{1 \le i \le |K|} s(k_i) \tag{6.27}$$

This feature captures the strongest emotional intensity present within the keyword representation. Highly polarized emotional expressions are frequently associated with sarcasm, exaggeration, irony, and emotionally amplified humor. Therefore, combining both average and maximum sentiment provides a richer affective representation of humorous text.

6.2.1.3 Semantic Incongruity Representation

Semantic incongruity is widely regarded as one of the fundamental characteristics of humor. Many humorous statements derive their effect from the unexpected combination of semantically distant concepts or contextually incompatible ideas. To model this behavior, semantic embeddings are generated for each extracted keyword using a transformer-based embedding model.

Each keyword k_i is mapped to a dense embedding vector represented as Eq. 6.28.

$$\mathbf{e}_i \in \mathbb{R}^d \tag{6.28}$$

The mean embedding of the complete keyword set is computed using Eq. 6.29.

$$\bar{\mathbf{e}} = \frac{1}{|K|} \sum_{i=1}^{|K|} \mathbf{e}_i \tag{6.29}$$

To estimate semantic similarity, the cosine similarity between each keyword embedding and the mean embedding is calculated using Eq. 6.30.

$$\cos_\text{sim}(k_i) = \frac{\mathbf{e}_i \cdot \bar{\mathbf{e}}}{\|\mathbf{e}_i\| \|\bar{\mathbf{e}}\|} \tag{6.30}$$

The semantic incongruity score is subsequently computed using Eq. 6.31.

$$I_{\text{semantic}} = 1 - \frac{1}{|K|} \sum_{i=1}^{|K|} \cos_\text{sim}(k_i) \tag{6.31}$$

Higher values of I_{semantic} indicate greater semantic divergence among the extracted keywords, thereby reflecting stronger contextual incongruity. Since incongruity forms a central mechanism underlying many humorous expressions, this feature provides important discriminatory information for humor classification.

6.2.1.4 Sarcasm-Oriented Pragmatic Representation

Humorous language frequently incorporates sarcasm, irony, or pragmatic contradiction that may not be directly observable through lexical or semantic analysis alone. To capture such pragmatic cues, the extracted keywords are concatenated into a short phrase τ, which is subsequently processed using a pretrained sarcasm detection model.

The model estimates a probability distribution represented as Eq. 6.32.

$$p(c|\tau) \tag{6.32}$$

over the classes represented as Eq. 6.33.

$$c \in \{\text{non-sarcastic}, \text{sarcastic}\} \tag{6.33}$$

The sarcasm-oriented feature is then computed using Eq. 6.47.

$$P_{\text{sarcasm}} = p(c = \text{sarcastic}|\tau) \tag{6.34}$$

This feature captures pragmatic contradiction and implicit ironic intent, both of which frequently contribute to humor generation in short textual content.

Finally, the complete feature representation corresponding to a text instance is expressed using Eq. 6.35.

$$\mathbf{x} = [L_{\text{avg}}, S_{\text{mean}}, S_{\text{max}}, I_{\text{semantic}}, P_{\text{sarcasm}}] \tag{6.35}$$

The resulting feature vector integrates lexical, affective, semantic, and pragmatic information into a unified representation space. By combining these complementary linguistic dimensions, the proposed HumorKey framework is able to capture multiple characteristics associated with humorous discourse while maintaining interpretability and computational efficiency.

6.2.2 Humor Classification

After constructing the feature representation for each text instance, the next stage focuses on classifying the input sentence as humorous $(y = 1)$ or non-humorous $(y = 0)$. The generated

feature vector integrates lexical complexity, affective polarity, semantic incongruity, and pragmatic sarcasm-related information, thereby providing a compact yet semantically informative representation of humorous discourse.

Since the extracted features exhibit different numerical ranges and distributions, feature normalization is performed prior to classification to ensure balanced contribution from all feature dimensions. Min-max normalization is applied using Eq. 6.36.

$$x_i' = \frac{x_i - \min(X_i)}{\max(X_i) - \min(X_i)}, \quad i = 1, \ldots, 5 \tag{6.36}$$

where X_i denotes the set of all values corresponding to feature i within the dataset, and Eq. 6.37.

$$\mathbf{x}' = [x_1', x_2', \ldots, x_5'] \tag{6.37}$$

represents the normalized feature vector used as input to the classifier.

To model the complex and often non-linear relationships between humor-oriented linguistic features, a Support Vector Machine (SVM) classifier with a Radial Basis Function (RBF) kernel is employed [170]. SVMs are particularly effective for high-dimensional classification problems because they maximize the decision margin between classes while maintaining strong generalization capability.

Given a training dataset $\{(\mathbf{x}_i', y_i)\}_{i=1}^{N}$, where $y_i \in \{0, 1\}$ denotes the class label, the SVM learns an optimal separating hyperplane by solving the soft-margin optimization problem defined in Eq. 6.38.

$$\min_{\mathbf{w}, b, \{\xi_i\}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i \tag{6.38}$$

subject to:

$$y_i \left(\mathbf{w}^\top \phi(\mathbf{x}_i') + b \right) \geq 1 - \xi_i \tag{6.39}$$

where:

- $\mathbf{w}$ and b define the separating hyperplane,
- $\xi_i \geq 0$ represents slack variables allowing limited misclassification,

- $\phi(\cdot)$ denotes the non-linear feature mapping function,
- C controls the trade-off between margin maximization and classification error minimization.

To align with the standard SVM formulation, binary labels are transformed from $\{0,1\}$ to $\{-1,+1\}$ using Eq. 6.40.

$$2y_i - 1 \tag{6.40}$$

The RBF kernel used in this work is computed using Eq. 6.41.

$$K(\mathbf{x}'_i, \mathbf{x}'_j) = \exp\left(-\gamma \|\mathbf{x}'_i - \mathbf{x}'_j\|^2\right) \tag{6.41}$$

where γ controls the influence of neighboring training instances within the transformed feature space. The RBF kernel enables the classifier to capture complex non-linear decision boundaries commonly associated with humor-oriented linguistic behavior.

For an unseen text instance $\mathbf{x}'$, the predicted humor label is computed using Eq. 6.42.

$$\hat{y} = \begin{cases} 1, & \text{sign}\left(\sum_{i \in \mathscr{S}} \alpha_i (2y_i - 1) K(\mathbf{x}'_i, \mathbf{x}') + b\right) = +1 \\ 0, & \text{sign}\left(\sum_{i \in \mathscr{S}} \alpha_i (2y_i - 1) K(\mathbf{x}'_i, \mathbf{x}') + b\right) = -1 \end{cases} \tag{6.42}$$

where:

- $\mathscr{S}$ denotes the set of support vectors,
- α_i represents the learned support vector coefficients.

This formulation enables the classifier to model subtle interactions among lexical, semantic, affective, and pragmatic humor related features. Consequently, the proposed HumorKey framework effectively captures complex humor patterns that are often implicit, context dependent, and difficult to identify using surface level textual representations alone.

Overall, the proposed humor detection framework integrates semantic keyword extraction, feature representation, and non-linear classification into a unified and interpretable pipeline for humor identification in short textual content.

6.2.3 Evaluation

This section presents the evaluation of the proposed humor detection framework using two publicly available datasets. The datasets are introduced first, followed by a comparative analysis against multiple machine learning baselines. Performance is assessed with standard measures, including accuracy, precision, recall, and F1-score, to illustrate the effectiveness of the framework across diverse text types.

6.2.3.1 Datasets

The proposed framework is evaluated on two publicly available datasets representing distinct forms of humorous and non-humorous text.

1. Dataset 1- Short Jokes Dataset [1]: This dataset contains 200,000 short text samples, evenly balanced between 100,000 humorous and 100,000 non-humorous instances (50% each). The texts are brief and conversational, with an average length of approximately 12 words. Humorous examples typically consist of direct one-liners (e.g., "Why don't scientists trust atoms? Because they make up everything."), while non-humorous entries include simple factual or neutral statements.

2. Dataset 2 – Yelp Reviews Dataset [2]: This dataset contains 50,000 user-generated Yelp reviews, also balanced between 25,000 humorous and 25,000 non-humorous samples (50% each). The reviews are substantially longer with an average length of approximately 125 words. Humorous entries frequently contain subtle sarcasm or playful exaggeration (e.g., "This salad was so small I needed binoculars to find it."), while non-humorous entries represent typical customer opinions and experiences.

These datasets together allow assessment of the model on both explicit humor in short texts and implicit humor in longer, real-world reviews.

6.2.3.2 Comparative Results and Discussion

To comprehensively evaluate the effectiveness of the proposed HumorKey framework, comparative experiments are conducted against multiple traditional machine learning classifiers as well as transformer-based language models. The objective of this analysis is to examine whether semantically enriched keyword-driven feature engineering remains competitive

[1] https://www.kaggle.com/datasets/deepcontractor/200k-short-texts-for-humor-detection
[2] https://www.kaggle.com/datasets/ilhamfp31/yelp-review-dataset

relative to both conventional classification approaches and modern deep contextual language models.

The proposed SVM-based humor classification framework is first compared against several widely used machine learning classifiers, including Random Forest [171], Logistic Regression [172], k-Nearest Neighbors (KNN) [173], XGBoost [174], Decision Tree [175], and Naive Bayes [176]. All comparative models are trained using the same feature representation generated through the proposed keyword extraction and feature engineering pipeline to ensure fair and unbiased evaluation.

The performance of the models is evaluated using standard classification metrics, namely Accuracy, Precision, Recall, and F1-Score. The comparative experimental results obtained on both datasets are presented in Table 6.4.

Table 6.4: Comparative Performance Evaluation of Humor Detection Models

Model	**Accuracy**	**Precision**	**Recall**	**F1 score**
	Dataset 1			
SVM	0.9070	0.9097	0.9202	0.9149
Random Forest	0.8632	0.8647	0.8590	0.8619
Logistic Regression	0.8629	0.8513	0.9011	0.8755
KNN	0.8305	0.8295	0.8343	0.8319
XGBoost	0.8815	0.8779	0.8917	0.8847
Decision Tree	0.8536	0.8434	0.8546	0.8489
Naive Bayes	0.8595	0.8334	0.8697	0.8512
	Dataset 2			
SVM	0.8737	0.8783	0.9154	0.8965
Random Forest	0.8279	0.8286	0.8152	0.8219
Logistic Regression	0.8572	0.8522	0.9050	0.8778
KNN	0.8205	0.8208	0.8150	0.8179
XGBoost	0.8407	0.8375	0.8830	0.8597
Decision Tree	0.8089	0.8228	0.7632	0.7918
Naive Bayes	0.8335	0.8477	0.7943	0.8201

The results demonstrate that the proposed HumorKey framework consistently achieves the best overall performance across both datasets when combined with the SVM classifier. On Dataset 1, the framework achieves an Accuracy of 90.70% and an F1-Score of 91.49%, outperforming all comparative baseline models. Similarly, on Dataset 2, which contains longer and contextually noisier user-generated reviews, the proposed framework maintains strong predictive performance with an F1-Score of 89.65%. The consistently high Recall values obtained across both datasets indicate the strong capability of the proposed methodology to correctly

identify humorous text instances, including implicit and context-dependent humor patterns.

Among the comparative baselines, XGBoost and Logistic Regression achieve comparatively competitive results due to their ability to model weighted feature interactions and non-linear decision boundaries. However, their performance remains consistently lower than that of the proposed SVM-based framework. Traditional classifiers such as KNN, Decision Tree, and Naive Bayes exhibit comparatively weaker performance, particularly in terms of Recall and F1-Score, indicating reduced capability in capturing semantic incongruity, sarcasm, and contextual humor-oriented linguistic patterns.

To further evaluate the effectiveness of the proposed methodology, an additional comparison is conducted against several transformer-based language models, including BERT [56], RoBERTa [57], DistilBERT [177], and ColBERT [178]. Unlike the proposed HumorKey framework, which performs humor detection using semantically enriched keyword extraction and feature-driven classification, the transformer baselines are fine-tuned end-to-end directly on raw textual input.

The comparative F1-Score performance across both datasets is illustrated in Figure 6.12.

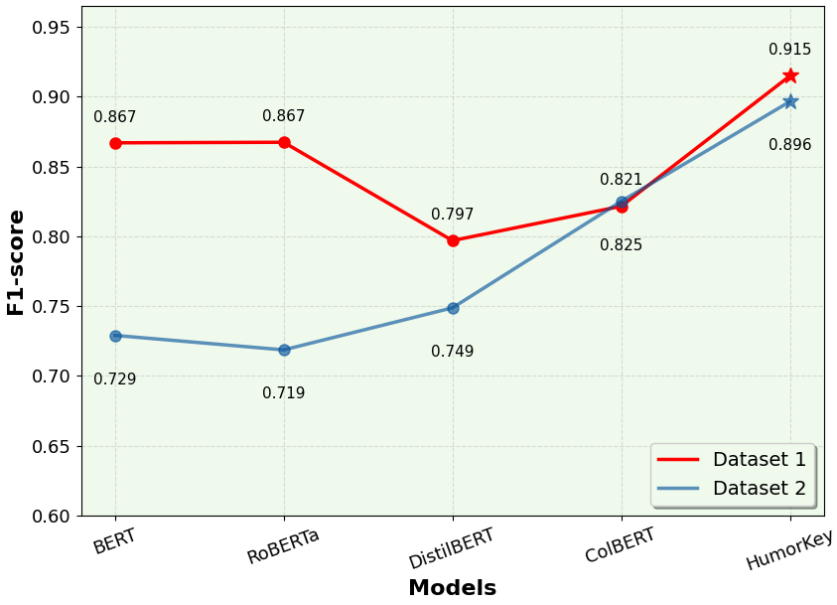

Figure 6.12: F1-Score comparison between transformer-based baseline models and the proposed HumorKey framework across both datasets.

As observed in Figure 6.12, the proposed HumorKey framework consistently achieves the highest performance across both datasets. On Dataset 1, the proposed framework outperforms the strongest transformer baseline, RoBERTa, by approximately 5% in terms of F1-Score. A similar trend is observed on Dataset 2, where HumorKey surpasses both ColBERT and

DistilBERT despite the increased contextual complexity and implicit humor characteristics present within user-generated review text.

These findings indicate that although transformer architectures provide strong contextual language modeling capabilities, they do not inherently capture humor-oriented semantic incongruity, polarity transitions, exaggeration, and pragmatic contradiction without task-specific adaptation. In contrast, the proposed HumorKey framework explicitly models these humor-related linguistic characteristics through semantically guided keyword extraction and multi-dimensional feature representation.

Furthermore, the proposed methodology maintains stronger interpretability and lower computational complexity compared to large transformer architectures requiring extensive fine-tuning and higher computational resources. Overall, the experimental results demonstrate that the proposed HumorKey framework provides an effective, robust, and semantically interpretable approach for humor detection across diverse textual environments.

6.2.3.3 Statistical Validation

To further examine the reliability and stability of the proposed HumorKey framework, statistical validation is performed using repeated experimental evaluation and paired Student's t-test analysis. The objective of this analysis is to determine whether the observed performance improvements achieved by the proposed methodology are statistically significant and consistently reproducible across multiple experimental runs.

For each classifier, experiments are repeated using different random initialization seeds, and the mean and standard deviation of the obtained F1-Scores are computed. This evaluation provides insight into both predictive performance and model stability under varying training conditions.

To assess statistical significance, a paired Student's t-test is performed between the proposed SVM-based HumorKey framework and each comparative baseline model. The t-statistic is computed using Eq. 6.43.

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \tag{6.43}$$

where:

- $d_i = x_i - y_i$ denotes the paired difference in F1-Scores for experimental run i,

- $\bar{d}$ represents the mean paired difference,
- s_d denotes the standard deviation of the paired differences,
- n corresponds to the total number of experimental runs.

A significance threshold of $p < 0.05$ is adopted to determine whether the observed performance differences are statistically significant.

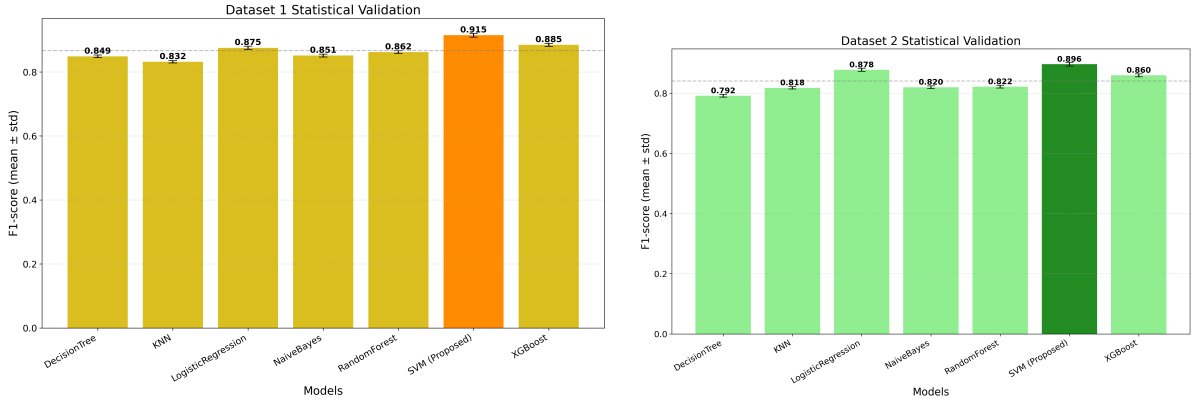

Figure 6.13: Statistical validation results showing mean F1-Scores and standard deviation across repeated experimental runs for both datasets.

Figure 6.13 illustrates the mean F1-Scores together with standard deviation error bars for all evaluated models across both datasets. The comparatively small error bars associated with the proposed HumorKey framework indicate highly stable predictive behavior and low performance variance across repeated experiments.

Across both datasets, the proposed SVM-based framework achieves the highest mean F1-Score while simultaneously maintaining minimal variance (Dataset 1: 0.9149 ± 0.006; Dataset 2: 0.8965 ± 0.006). In contrast, several comparative baseline models exhibit larger performance fluctuations, indicating comparatively lower stability and reduced generalization consistency.

Furthermore, the paired t-test analysis confirms that the improvements achieved by the proposed framework are statistically significant relative to all evaluated baselines, with all obtained p-values remaining below the significance threshold of 0.05. These findings indicate that the observed performance gains are not attributable to random variation.

Overall, the statistical validation analysis demonstrates that the proposed HumorKey framework provides stable, reliable, and statistically significant humor detection performance across diverse textual environments.

This section presented the downstream application of the proposed HumorKey framework for humor detection in textual content. By utilizing semantically enriched keyword extraction together with lexical, affective, semantic, and pragmatic feature representation, the framework effectively captured important linguistic characteristics associated with humorous discourse, including semantic incongruity, emotional polarity, and sarcasm. Experimental evaluation conducted on both short-form jokes and long-form review datasets demonstrated that the proposed methodology consistently outperformed multiple traditional machine learning classifiers as well as transformer-based language models. The statistical validation further confirmed the robustness, stability, and significance of the obtained performance improvements. Overall, the results demonstrate that the proposed HumorKey framework provides an interpretable, computationally efficient, and semantically effective approach for humor detection across diverse textual environments.

6.3 Offensive Meme Detection

This section presents the downstream offensive meme detection framework developed using the multimodal offensive-aware keywords extracted in the previous chapter. While the earlier methodology focused on multimodal semantic acquisition, offensive lexicon expansion, and NSGA-II-based offensive keyword optimization, the present section utilizes those extracted semantic representations for downstream offensiveness prediction and meme classification.

The primary objective of this framework is to determine whether the extracted multimodal keywords preserve sufficient semantic, contextual, and toxicity oriented information for reliable identification of offensive meme content. Unlike conventional offensive language detection tasks that rely primarily on textual analysis, offensive memes frequently communicate harmful intent through complex interactions among visual symbolism, embedded text, sarcasm, demographic cues, emotional context, and implicit social references. Consequently, effective offensiveness detection requires semantic modeling capable of jointly capturing both explicit and implicit multimodal associations.

To address these challenges, the proposed framework transforms the extracted offensive-aware keywords into a set of explicit and implicit semantic features representing lexical overlap, semantic similarity, sentiment polarity, contextual sarcasm, and toxicity-oriented semantic alignment. These features are subsequently incorporated into a thermodynamic offense modeling framework that estimates the overall offensiveness of a meme using entropy- and energy-based

semantic modeling principles.

Figure 6.14 illustrates the complete downstream offensiveness detection pipeline, including explicit and implicit feature extraction, thermodynamic offense modeling, and final meme classification.

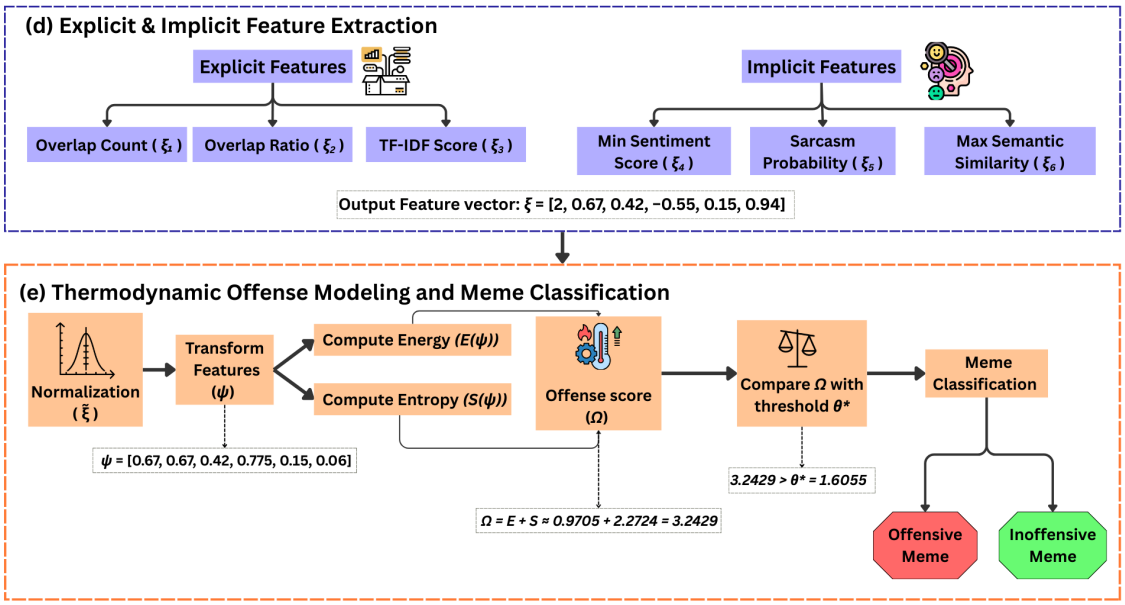

Figure 6.14: Downstream offensive meme detection framework utilizing multimodal offensive-aware keywords for explicit and implicit feature extraction, thermodynamic offense modeling, and final meme classification.

6.3.1 Extraction of Explicit and Implicit Offense-Indicative Features

Following the extraction of offensive-aware multimodal keywords, the next stage focuses on transforming these semantic representations into a structured feature space suitable for downstream offensiveness modeling. Since offensive memes frequently communicate harmful intent through both direct lexical expressions and subtle contextual cues, the proposed framework extracts a combination of explicit and implicit semantic features capable of capturing multiple dimensions of offensiveness.

Let $\mathscr{K}_{\text{off}} = \{w_1, w_2, \ldots, w_k\}$ denote the optimized offensive-aware keyword set extracted from a meme instance. The extracted features are designed to model both surface-level toxicity indicators and deeper contextual semantic behavior commonly associated with offensive meme content.

To facilitate systematic analysis, the extracted features are categorized into two principal

groups:

- **Explicit Features:** Features derived from observable lexical and statistical properties.
- **Implicit Features:** Features capturing contextual, affective, semantic, and pragmatic characteristics using pretrained neural representations.

This distinction enables the framework to jointly model overt toxicity cues together with latent social and contextual signals that frequently occur within multimodal meme discourse.

6.3.1.1 Explicit Offense-Oriented Features

Explicit features primarily capture direct lexical alignment between the extracted keywords and the offensive semantic space represented by the expanded offensive lexicon.

The first explicit feature, denoted as ξ_1, measures the total number of extracted keywords present within the offensive lexicon $\mathscr{L}_{\text{off}}$. The second feature, ξ_2, represents the normalized offensive keyword proportion relative to the total number of extracted keywords. These quantities are computed using Eq. 6.44.

$$\xi_1 = |\mathscr{K}_{\text{off}} \cap \mathscr{L}_{\text{off}}|, \quad \xi_2 = \frac{\xi_1}{|\mathscr{K}_{\text{off}}|} \tag{6.44}$$

The feature ξ_1 captures the absolute presence of offensive lexical indicators, whereas ξ_2 estimates the relative density of offensive semantic content within the extracted keyword representation. Together, these features provide direct evidence regarding the degree of lexical offensiveness present in the meme.

To further estimate the contextual importance of the extracted keywords, a TF-IDF-based semantic salience score is additionally computed. Let $f_{\text{tfidf}}(\cdot)$ denote the TF-IDF weighting function learned over the training corpus. The semantic salience feature ξ_3 is computed using Eq. 6.45.

$$\xi_3 = \frac{1}{k} \sum_{i=1}^{k} f_{\text{tfidf}}(w_i) \tag{6.45}$$

This feature estimates the informational importance of the extracted offensive-aware keywords within the meme corpus. Keywords exhibiting higher TF-IDF values generally correspond to semantically distinctive or contextually informative linguistic units, thereby improving discriminatory capability during downstream offensiveness modeling.

6.3.1.2 Implicit Offense-Oriented Features

Although explicit lexical indicators provide important offensiveness cues, offensive memes frequently rely on implicit semantics, sarcasm, emotional polarity, contextual contradiction, and subtle semantic associations that cannot be adequately captured through lexical overlap alone. To model such latent semantic behavior, the proposed framework extracts multiple implicit offense-oriented features.

The first implicit feature, denoted as ξ_4, represents the minimum sentiment polarity observed among the extracted keywords. Sentiment polarity is computed using the VADER sentiment analyzer, and the feature is evaluated using Eq. 6.46.

$$\xi_4 = \min_{w_i \in \mathscr{K}_{\text{off}}} \text{VADER}(w_i) \tag{6.46}$$

Lower sentiment values indicate stronger negative affective orientation, which frequently correlates with abusive, hostile, or offensive meme content. Unlike sentence-level sentiment analysis, the proposed methodology applies VADER directly to optimized offense-aware keywords in order to focus specifically on semantically relevant toxicity-oriented lexical units while minimizing contextual noise introduced by descriptive or stylistic content.

To capture pragmatic contradiction and implicit irony commonly observed in offensive memes, a sarcasm-oriented feature is additionally extracted using a transformer-based sarcasm classifier $f_{\text{sar}}(\cdot)$. The extracted keywords are concatenated into a unified semantic sequence, and the classifier output is transformed into a sarcasm probability score using Eq. 6.47.

$$\xi_5 = \text{Softmax}\left(f_{\text{sar}}\left(\text{concat}(\mathscr{K}_{\text{off}})\right)\right)_{[1]} \tag{6.47}$$

where the index $[1]$ corresponds to the probability associated with the sarcastic class.

This feature enables the framework to capture implicit offensive behavior communicated through irony, mockery, exaggeration, or contextually contradictory expressions that are frequently embedded within meme content.

Finally, semantic alignment with the offensive semantic space is modeled using the maximum semantic similarity feature ξ_6. Let $\boldsymbol{\mu}_{\text{off}}$ denote the centroid representation of the offensive lexicon in the semantic embedding space. The feature is computed using cosine similarity as defined in Eq. 6.48.

$$\xi_6 = \max_{w_i \in \mathcal{K}_{\text{off}}} \cos\left(f_{\text{emb}}(w_i), \boldsymbol{\mu}_{\text{off}}\right) \tag{6.48}$$

This feature measures the strongest semantic alignment between any extracted keyword and the learned offensive semantic representation. Higher similarity values indicate stronger proximity to offensive semantic regions within the embedding space, thereby improving the framework's ability to identify implicitly offensive content even in the absence of direct toxic vocabulary.

The final feature representation corresponding to a meme instance is therefore expressed as:

$$\boldsymbol{\xi} = [\xi_1, \xi_2, \xi_3, \xi_4, \xi_5, \xi_6] \tag{6.49}$$

The resulting feature vector provides a compact yet semantically informative representation integrating lexical, affective, contextual, pragmatic, and semantic characteristics associated with offensive meme content. By jointly modeling explicit and implicit offensiveness cues, the proposed framework effectively bridges interpretable surface-level toxicity indicators with deeper contextual semantic behavior necessary for reliable offensive meme detection.

6.3.2 Thermodynamic Offense Modeling and Meme Classification

After constructing the explicit and implicit offense-indicative feature representation, the final stage of the proposed framework focuses on estimating the overall offensiveness of a meme using a thermodynamically inspired semantic modeling strategy. The motivation behind this formulation is that offensive meme content is not determined solely by the presence of isolated toxic cues, but rather by the combined interaction, intensity, and distribution of multiple semantic indicators such as lexical toxicity, sarcasm, negative affect, and semantic alignment with offensive concepts.

To model this behavior, the proposed framework introduces a thermodynamic offense formulation inspired by concepts from statistical mechanics and information theory [179, 120]. Within this formulation:

- **Energy** represents the overall magnitude or intensity of offensive semantic evidence.
- **Entropy** captures the diversity and distributional spread of offense-indicative semantic cues.

This combined formulation enables the framework to model not only the strength of offensive signals but also their contextual distribution across multiple semantic dimensions.

Let the extracted feature representation corresponding to a meme instance be denoted as Eq. 6.50.

$$\boldsymbol{\xi} = [\xi_1, \xi_2, \xi_3, \xi_4, \xi_5, \xi_6] \tag{6.50}$$

Since the extracted features exhibit different numerical scales and distributions, min-max normalization is first applied to transform all feature components into the range $[0, 1]$. The normalized representation is denoted as:

$$\tilde{\boldsymbol{\xi}} = [\tilde{\xi}_1, \tilde{\xi}_2, \ldots, \tilde{\xi}_6] \tag{6.51}$$

To further align the features with offensiveness-oriented semantic interpretation, a linear transformation function is introduced. The transformed representation is computed using Eq. 6.52.

$$\boldsymbol{\psi} = [\tilde{\xi}_1, \tilde{\xi}_2, \tilde{\xi}_3, (1 - \tilde{\xi}_4), \tilde{\xi}_5, (1 - \tilde{\xi}_6)] \tag{6.52}$$

The inversion applied to ξ_4 and ξ_6 reflects the observation that stronger negative sentiment polarity and greater semantic proximity to offensive semantic regions correspond to increased offensive potential. Importantly, only linear transformations are utilized in order to preserve feature interpretability and avoid introducing dataset-specific nonlinear distortions.

To estimate the relative importance of each semantic component, the empirical standard deviation of each normalized feature across the dataset is computed as Eq. 6.53.

$$\sigma_j = \text{std}(\tilde{\xi}_j) \tag{6.53}$$

These variance-based coefficients are subsequently used as feature importance weights during offense energy estimation. Features exhibiting higher variance across the dataset are assumed to contribute stronger discriminatory information toward offensiveness prediction.

The thermodynamic offense energy corresponding to a meme instance is then computed using Eq. 6.54.

$$\mathscr{E}(\psi) = \sum_{j=1}^{6} \sigma_j \cdot \psi_j \tag{6.54}$$

This formulation models offensive intensity through weighted semantic aggregation, where higher feature magnitudes correspond to stronger offensiveness-oriented evidence.

Although offensive intensity is important, memes frequently communicate harmful intent through combinations of multiple subtle semantic signals rather than through a single dominant cue. Therefore, the proposed framework additionally incorporates an entropy-based regularization component that captures the distributional diversity of offense-indicative evidence.

First, the normalized contribution proportion of each transformed feature is computed as Eq. 6.55.

$$p_j = \frac{\psi_j}{\sum_{i=1}^{6} \psi_i + \varepsilon} \tag{6.55}$$

where ε denotes a small constant introduced for numerical stability.

Using these normalized proportions, the thermodynamic entropy term is computed using Eq. 6.56.

$$S(\psi) = -\sum_{j=1}^{6} p_j \log_2(p_j + \varepsilon) \tag{6.56}$$

The entropy component measures how evenly offensive evidence is distributed across the feature space. Higher entropy values indicate that multiple semantic dimensions collectively contribute toward offensiveness prediction, whereas lower entropy values imply excessive reliance on isolated cues. Consequently, the entropy formulation improves robustness by reducing dependence on any single modality or semantic signal.

The final thermodynamic offense score is subsequently computed using Eq. 6.57.

$$\Omega = \mathscr{E}(\psi) + S(\psi) \tag{6.57}$$

The scalar offense score Ω therefore jointly captures:

- the magnitude of offensiveness-oriented semantic evidence,
- the contextual diversity of semantic cues,
- and the distributional balance among multiple offense-indicative features.

To perform final meme classification, the computed offense score is compared against a decision threshold θ such that:

$$\hat{y} = \mathbb{I}(\Omega > \theta) \tag{6.58}$$

where $\hat{y} = 1$ denotes an offensive meme and $\hat{y} = 0$ denotes a non-offensive meme.

The optimal threshold value θ^* is determined empirically using validation data by maximizing the F1-Score as represented in Eq. 6.59.

$$\theta^* = \arg\max_{\theta} F1(\theta) \tag{6.59}$$

Overall, the proposed thermodynamic offense modeling framework provides a unified and interpretable mechanism for offensive meme detection by jointly integrating semantic intensity and contextual diversity into a single offensiveness estimation process. Unlike conventional methods relying solely on isolated visual or textual signals, the proposed framework effectively captures both explicit toxicity and implicit contextual behavior, including sarcasm, semantic proximity, negative affect, and multimodal semantic interactions. Consequently, the methodology offers a robust, scalable, and semantically grounded approach for offensive meme classification in complex social media environments.

6.3.3 Experimental Evaluation

This section presents the experimental evaluation of the proposed offensive meme detection framework. The objective of the evaluation is to examine the effectiveness of the multimodal offensive-aware keyword extraction methodology and the proposed thermodynamic offense modeling strategy for identifying offensive meme content under diverse semantic and contextual conditions.

The evaluation is conducted using multiple publicly available benchmark meme datasets containing offensive and non-offensive meme instances exhibiting varying degrees of multimodal complexity, implicit semantics, sarcasm, and contextual ambiguity. The proposed framework is assessed using standard classification metrics in order to evaluate predictive performance, robustness, and discriminatory capability across different meme distributions.

6.3.3.1 Datasets

To comprehensively evaluate the proposed framework, experiments are conducted on three publicly available benchmark datasets widely used for offensive meme detection research, namely MultiOFF [108], Memotion [114], and Facebook Hateful Memes [107]. These datasets differ substantially in terms of meme composition, contextual complexity, semantic ambiguity, and offensiveness characteristics, thereby enabling robust evaluation under diverse multimodal conditions.

- **MultiOFF Dataset**

  The MultiOFF dataset contains approximately 5,000 meme instances annotated using binary offensive and non-offensive labels. The dataset primarily focuses on memes containing implicit offensiveness communicated through complex visual-textual interactions, symbolic imagery, and contextual semantic associations. Offensive intent within MultiOFF memes is often subtle and context-dependent, making the dataset particularly suitable for evaluating semantic reasoning capability in multimodal offensiveness detection.

- **Memotion Dataset**

  The Memotion dataset consists of approximately 8,000 memes annotated across multiple affective and semantic dimensions, including sentiment, emotion, and offensiveness. In this work, only the binary offensiveness labels are utilized for evaluation. The dataset contains memes exhibiting sarcasm, irony, emotional exaggeration, and contextual humor, thereby introducing additional complexity for automated offensiveness prediction.

- **Facebook Hateful Memes Dataset**

  The Facebook Hateful Memes dataset contains approximately 10,000 carefully curated meme instances designed specifically for multimodal hateful content detection. In many cases, neither the image nor the textual content alone is sufficient to infer offensiveness. Instead, harmful intent emerges only through joint multimodal interpretation, making this dataset a particularly challenging benchmark for evaluating multimodal semantic understanding and contextual reasoning capability.

Collectively, these datasets enable comprehensive assessment of the proposed framework across multiple forms of offensive meme content, including explicit toxicity, implicit semantic offensiveness, sarcasm-driven abuse, and multimodal contextual hostility.

6.3.3.2 Evaluation Metrics

The effectiveness of the proposed offensive meme detection framework is evaluated using standard classification metrics, including Accuracy, Precision, Recall, F1-Score, and the Area Under the Receiver Operating Characteristic Curve (ROC AUC). These metrics collectively provide insight into classification reliability, sensitivity toward offensive content, and overall discriminatory capability.

Accuracy measures the proportion of correctly classified meme instances across both offensive and non-offensive categories. Precision evaluates the proportion of correctly identified offensive memes among all memes predicted as offensive, thereby measuring the reliability of offensive predictions. Recall quantifies the framework's ability to successfully identify actual offensive meme instances, which is particularly important for harmful content moderation tasks. The F1-Score provides a balanced evaluation by jointly considering both Precision and Recall.

To further evaluate threshold-independent discriminatory capability, the ROC AUC metric is additionally computed. The Receiver Operating Characteristic (ROC) curve represents the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) across varying decision thresholds. The ROC AUC is computed using Eq. 6.60.

$$\text{ROC AUC} = \int_0^1 \text{TPR}(x)\,dx \tag{6.60}$$

This metric measures the capability of the framework to distinguish offensive meme instances from non-offensive instances independent of a fixed classification threshold. Higher ROC AUC values indicate stronger separability between the two classes and more reliable offensiveness prediction across varying operating conditions.

6.3.3.3 Performance of the Proposed Method

The performance of the proposed offensive meme detection framework is evaluated across the Memotion, MultiOFF, and Facebook Hateful Memes datasets. The obtained results are summarized in Table 6.5, while the corresponding ROC curves are illustrated in Figure 6.15.

The proposed framework achieves consistently strong performance across all benchmark datasets, with ROC AUC values exceeding 0.95, indicating strong separability between offensive and non-offensive meme instances. Among the evaluated datasets, the Facebook Hateful Memes

Table 6.5: Performance Evaluation of the Proposed Offensive Meme Detection Framework Across Benchmark Datasets

Metric	Memotion	MultiOFF	FB Dataset
Accuracy	0.8812	0.8926	0.9023
Precision	0.8090	0.8972	0.9550
Recall	0.9919	0.9505	0.7865
F1 Score	0.8911	0.9231	0.8663
ROC AUC	0.9539	0.9573	0.9634
PR AUC	0.9545	0.9817	0.9531

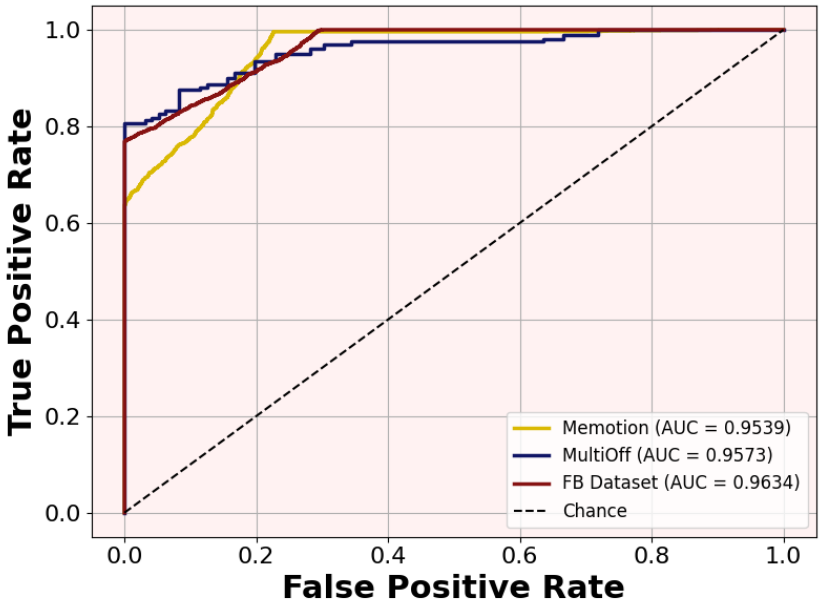

Figure 6.15: ROC curves of the proposed framework across benchmark offensive meme datasets.

dataset achieves the highest Accuracy and Precision, suggesting that the framework effectively captures explicit multimodal offensive patterns. In contrast, the Memotion dataset yields the highest Recall, demonstrating strong sensitivity toward identifying subtle and emotionally expressive offensive content. The MultiOFF dataset exhibits the most balanced performance across all evaluation metrics, indicating strong generalization capability under linguistically diverse meme environments. The ROC curves further confirm stable discriminatory performance across varying decision thresholds, highlighting the robustness and effectiveness of the proposed multimodal offensiveness modeling framework.

6.3.3.4 KDE-Based Threshold Analysis

To determine the optimal decision boundary for offensive meme classification, the threshold parameter θ^* is selected independently for each dataset by maximizing the F1-Score over the

validation set. The resulting optimal thresholds are obtained as $\theta^* = 1.6134$ for Memotion, $\theta^* = 0.7986$ for MultiOFF, and $\theta^* = 1.6055$ for the Facebook Hateful Memes dataset.

Figure 6.16 presents the Kernel Density Estimation (KDE) distributions of the thermodynamic offense scores corresponding to offensive and non-offensive meme instances across all datasets.

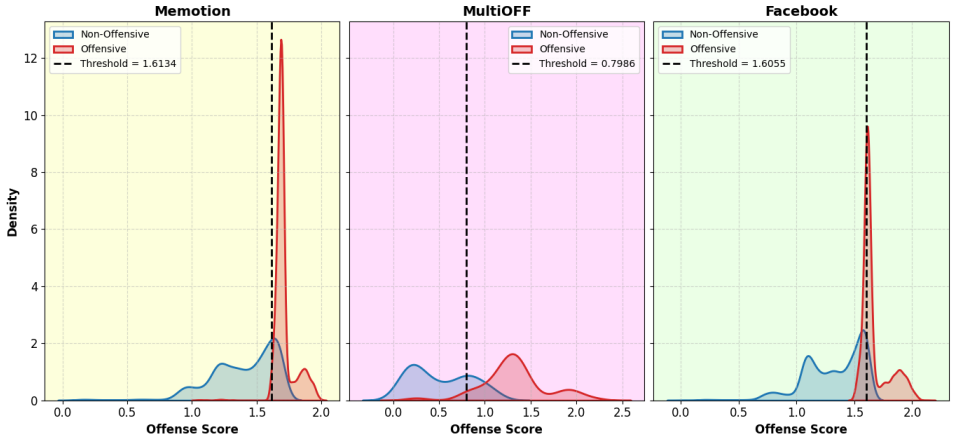

Figure 6.16: KDE distributions of thermodynamic offense scores for offensive and non-offensive meme instances across benchmark datasets.

As illustrated in Figure 6.16, the selected thresholds are positioned near the regions exhibiting maximal separation between offensive and non-offensive score distributions. This separation minimizes distributional overlap and consequently reduces classification ambiguity between the two classes. The comparatively distinct density distributions further indicate that the proposed thermodynamic offense modeling framework generates discriminative offense scores capable of effectively separating offensive meme content from non-offensive instances across diverse multimodal environments.

Overall, the KDE analysis provides empirical evidence supporting the suitability and stability of the selected threshold values for downstream offensiveness prediction.

6.3.3.5 Comparative Analysis

To evaluate the effectiveness of the proposed framework, a comparative analysis is conducted against multiple state-of-the-art multimodal baselines across the MultiOFF, Memotion, and Facebook Hateful Memes datasets. For fair evaluation, identical preprocessing strategies, data splits, and evaluation metrics are maintained across all methods. The comparison includes transformer-based multimodal architectures, fusion-based models, and multimodal sentiment-aware frameworks commonly used for offensive meme detection.

The evaluated baselines include CLIP-Art [180], MMBT [107], MMDL [181], FewShot [182], mBERT+VGG-19 [183], mBERT+ViT [184], VisualBERT [185], LXMERT [186], MSCLIP [187], Sentiment [188], EmoffMeme [189], and CLIP+UNITER+BERT [190]. The comparative performance results are summarized in Table 6.6.

Table 6.6: Comparative Performance Analysis Across Benchmark Offensive Meme Datasets

Method	**MultiOff**				**Memotion**				**FB**			
	A	P	R	F1	A	P	R	F1	A	P	R	F1
Proposed	**0.8812**	**0.809**	**0.9919**	**0.8911**	**0.8926**	**0.8972**	**0.9505**	**0.9231**	**0.9023**	**0.955**	**0.7865**	**0.8663**
CLIP-Art	0.5000	0.3012	0.2155	0.2513	0.4537	0.6073	0.3048	0.4059	0.5407	0.5972	0.4852	0.5354
MMBT	0.4000	0.2660	0.1379	0.1818	0.5034	0.5920	0.5676	0.5796	0.6700	0.5493	0.5259	0.5373
MMDL	0.8200	0.7100	0.8700	0.7800	0.7400	0.6300	0.8600	0.7200	0.6854	0.6600	0.7200	0.6886
FewShot	0.8115	0.8050	0.8150	0.8099	0.7672	0.7435	0.7568	0.7501	0.7146	0.7023	0.7261	0.7140
mBERT+VGG-19	0.5860	0.5630	0.5400	0.5513	0.6500	0.6235	0.7800	0.6930	0.6356	0.6135	0.7800	0.6868
MBERT+ViT	0.7246	0.6800	0.7000	0.6899	0.7660	0.7333	0.8200	0.7740	0.6956	0.6802	0.8836	0.7686
VisualBERT	0.6667	0.6316	0.6667	0.6486	0.6300	0.6276	0.7859	0.6979	0.7450	0.7200	0.8012	0.7584
LXMERT	0.7923	0.7767	0.7500	0.7631	0.8080	0.7860	0.9100	0.8434	0.8230	0.7900	0.9766	0.8730
MSCLIP	0.8173	0.8032	0.8261	0.8144	0.7438	0.7236	0.7389	0.7311	0.6934	0.6815	0.7022	0.6917
Sentiment	0.4430	0.4430	0.3017	0.3590	0.4771	0.5964	0.4517	0.5141	0.5408	0.4902	0.7702	0.5990
EmoffMeme	0.7800	0.7540	0.8300	0.7900	0.7500	0.7000	0.7300	0.7146	0.6800	0.6500	0.6600	0.6540
CLIP+UNITER+BERT	0.8040	0.7970	0.8120	0.8040	0.7940	0.7750	0.8900	0.8290	0.7070	0.7150	0.6980	0.7060

The proposed framework consistently achieves the best overall performance across all three benchmark datasets. In particular, the framework attains the highest F1-Scores on MultiOFF (0.8911), Memotion (0.9231), and Facebook Hateful Memes (0.8663), demonstrating strong generalization capability under diverse multimodal offensiveness conditions.

On the MultiOFF dataset, the proposed method achieves exceptionally high Recall, indicating strong capability in identifying offensive meme instances with minimal false negatives. On the Memotion dataset, the framework maintains a balanced trade-off between Precision and Recall despite the presence of emotionally expressive and sarcasm-oriented content. Similarly, on the Facebook Hateful Memes dataset, the framework achieves the highest Precision, indicating strong discriminative capability for subtle and context-dependent offensive content.

Transformer-based multimodal architectures such as LXMERT, VisualBERT, and MSCLIP generally outperform conventional fusion-based approaches due to their stronger cross-modal contextual modeling capability. However, the proposed framework consistently surpasses all comparative baselines across Accuracy, Recall, and F1-Score. These results indicate that the integration of multimodal semantic acquisition, offensive-aware keyword optimization, and thermodynamic offense modeling provides more effective semantic discrimination than conventional end-to-end multimodal fusion strategies.

Overall, the comparative analysis demonstrates the robustness, effectiveness, and generalization capability of the proposed framework for offensive meme detection across diverse multimodal environments.

6.3.3.6 Statistical Significance Analysis

To further validate the reliability of the proposed framework, statistical significance analysis is performed using a Two-Tailed Z-Test and Bootstrapping. These analyses examine whether the observed improvements over baseline methods are statistically significant and stable across different datasets.

Two-Tailed Z-Test Analysis A two-tailed Z-test is performed to evaluate whether the F1-Score of the proposed framework differs significantly from the mean F1-Score of the comparative baseline models. The null and alternative hypotheses are defined as Eq. 6.61.

$$\begin{aligned} H_0 &: \mu_p = \mu_b \\ H_1 &: \mu_p \neq \mu_b \end{aligned} \tag{6.61}$$

The Z-statistic is computed using Eq. 6.62.

$$Z = \frac{\mu_p - \mu_b}{\sigma_b / \sqrt{n}} \tag{6.62}$$

where μ_p denotes the F1-Score of the proposed framework, μ_b represents the mean F1-Score of the baseline models, σ_b is the standard deviation of the baseline scores, and n denotes the number of baseline methods.

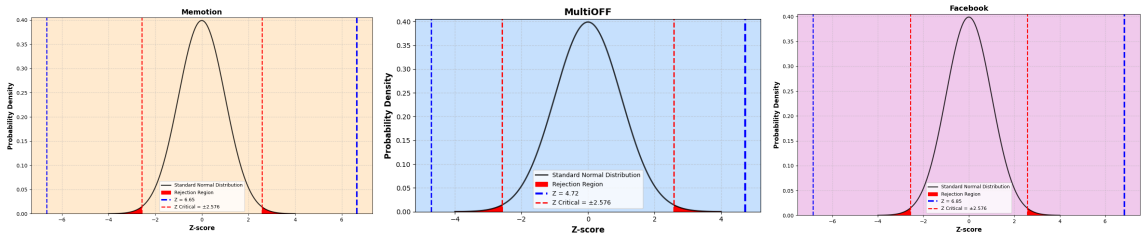

Figure 6.17: Two-tailed Z-test results across benchmark datasets at a 99% confidence level.

As illustrated in Figure 6.17, the obtained Z-scores for Facebook (6.85), Memotion (6.65), and MultiOFF (4.72) substantially exceed the critical threshold of ± 2.576. These results indicate that the improvements achieved by the proposed framework are statistically significant and unlikely to occur due to random variation.

Bootstrapping-Based Statistical Evaluation To further examine robustness and performance stability, bootstrap sampling is performed using $B = 1000$ bootstrap replicates. For each replicate, the F1-Score is recomputed by sampling instances with replacement from the

original dataset. The resulting bootstrap distribution is represented as Eq. 6.63

$$\mathscr{F} = \{F1_1, F1_2, \ldots, F1_B\} \tag{6.63}$$

The corresponding 95% confidence interval is estimated using the percentile method defined in Eq. 6.64.

$$\mathrm{CI}_{95\%} = [F1_{\lfloor 0.025B \rfloor}, F1_{\lfloor 0.975B \rfloor}] \tag{6.64}$$

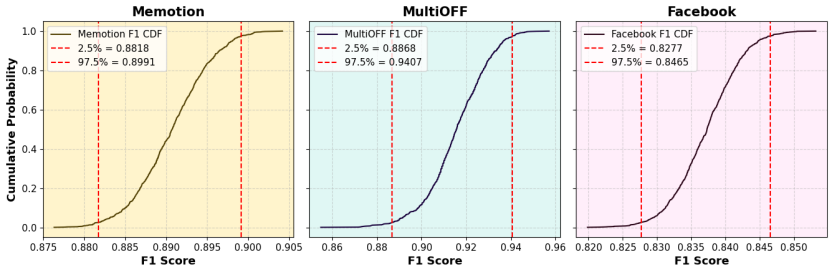

Figure 6.18: Bootstrapped CDF distributions of F1-Scores with 95% confidence intervals across benchmark datasets.

Figure 6.18 illustrates the bootstrapped cumulative distribution functions of the F1-Scores across all datasets. The obtained confidence intervals remain narrow for all datasets, indicating stable and consistent predictive performance with low variance across repeated sampling experiments. The results further confirm that the proposed framework maintains robust offensiveness prediction capability across diverse multimodal meme distributions.

Overall, both the Z-test and bootstrap analyses provide strong statistical evidence supporting the reliability, stability, and effectiveness of the proposed offensive meme detection framework.

This section demonstrated the downstream application of the proposed multimodal offensive-aware keyword extraction framework for offensive meme detection. The extracted keywords were transformed into explicit and implicit semantic representations capturing lexical toxicity, semantic alignment, sentiment polarity, and sarcasm-oriented contextual behavior. These features were subsequently integrated within a thermodynamic offense modeling framework for final offensiveness prediction. Through this application, the study illustrates how semantically optimized multimodal keyword representations can effectively bridge multimodal understanding and interpretable offensive content detection in complex meme environments.

6.4 Conclusion

This chapter demonstrated the effectiveness of the proposed keyword extraction methodologies through their application to multiple downstream social media analysis tasks, including toxicity detection in tweets, humor detection from text, and offensive meme detection. The extracted keyword representations successfully preserved contextual, semantic, affective, and task-specific information necessary for reliable downstream prediction across diverse and noisy social media environments. Through the QuToX framework, toxicity-oriented semantic interactions were effectively utilized for interpretable tweet toxicity detection. Similarly, the HumorKey framework demonstrated that semantically enriched keywords can capture humor-oriented linguistic properties such as incongruity, sentiment variation, and sarcasm for robust humor classification. Furthermore, the multimodal offensive-aware keyword extraction framework effectively integrated explicit and implicit semantic cues for offensive meme detection through thermodynamic offense modeling. The experimental and statistical analyses conducted across all applications consistently validated the robustness, generalization capability, interpretability, and practical applicability of the proposed methodologies. Overall, the findings of this chapter establish that semantically optimized keyword extraction can serve not only as an information retrieval mechanism but also as a powerful semantic representation framework for complex downstream social media understanding tasks.

Chapter 7

Conclusion, Future Scope, and Social Impact

This chapter concludes the thesis by summarizing the major contributions, findings, limitations, and broader implications of the proposed research. The study focused on the development of semantically enriched keyword extraction methodologies for noisy and context-dependent social media environments, with particular emphasis on toxicity detection, humor understanding, and offensive meme analysis. Throughout the thesis, multiple keyword extraction frameworks were proposed by integrating contextual semantics, quantum-inspired interaction modeling, multimodal representation learning, optimization strategies, and interpretable semantic analysis. The effectiveness of these methodologies was further demonstrated through diverse downstream applications and extensive experimental evaluation across multiple benchmark datasets. In addition to summarizing the major findings of the research, this chapter discusses the limitations of the proposed approaches, outlines potential future research directions, and highlights the broader social impact and relevance of the work in relation to sustainable and responsible AI-driven social media analysis.

7.1 Summary of Findings

The major findings of the research conducted in this thesis are summarized as follows:

1. **Semantically optimized keyword extraction functions as a stable and transferable semantic representation mechanism.**

   The experimental findings demonstrated that the proposed keyword extraction methodologies preserve meaningful contextual and semantic information beyond simple lexical relevance. Unlike conventional extraction approaches that primarily rely on frequency statistics, the proposed frameworks generated semantically enriched keyword subsets capable of supporting multiple downstream language understanding tasks. The extracted keywords consistently retained sufficient contextual information for toxicity detection, humor understanding, and offensive meme classification without requiring complete dependence on large end-to-end supervised architectures. This establishes keyword extraction as a transferable semantic representation layer rather than merely an auxiliary preprocessing operation.

2. **Context-aware and knowledge-enriched semantic modeling substantially improves understanding of implicit social media language.**

   The research showed that integrating contextual semantics, semantic interaction modeling, and external knowledge expansion significantly improves performance in implicit and context-dependent language scenarios. The proposed methodologies effectively captured subtle semantic behaviors commonly observed in social media discourse, including sarcasm, semantic incongruity, emotional exaggeration, contextual toxicity, and implicit offensive intent. In particular, the incorporation of semantic similarity modeling, contextual interaction learning, and multimodal semantic acquisition enabled the frameworks to identify deeper semantic relationships that are often overlooked by surface-level lexical approaches.

3. **Optimization-driven keyword selection provides principled control over semantic relevance, diversity, and task alignment.**

   The thesis established that keyword extraction formulated as an optimization problem produces more informative and discriminative semantic representations than heuristic or purely statistical approaches. Through the integration of optimization objectives related

to semantic relevance, contextual coverage, redundancy reduction, diversity preservation, and offense alignment, the proposed methodologies generated keyword subsets that were simultaneously semantically meaningful and highly effective for downstream prediction tasks. Experimental comparisons consistently demonstrated that optimization-driven keyword extraction outperformed conventional frequency-based methods in terms of both semantic quality and downstream classification performance.

4. **Keyword-centric frameworks achieve strong predictive performance while maintaining semantic interpretability.**

   A major contribution of the research lies in demonstrating that interpretable keyword-centric frameworks can achieve competitive predictive performance relative to complex transformer-centric models. The proposed approaches preserved explicit semantic traceability through interpretable keyword representations, lexical clustering strategies, semantic feature modeling, and rule-driven decision mechanisms. This interpretability is particularly important in socially sensitive applications such as toxicity detection and offensive content moderation, where transparent and explainable predictions are necessary for trustworthy deployment. The findings therefore highlight that high predictive capability and semantic interpretability are not mutually exclusive objectives.

5. **The proposed methodologies exhibit robustness across heterogeneous data modalities and social media environments.**

   The research demonstrated that the proposed frameworks generalize effectively across diverse forms of social media content, including short textual discourse, conversational tweets, humorous textual corpora, and multimodal image-text memes. The ability of the methodologies to operate successfully across both textual and multimodal domains validates the flexibility and scalability of the proposed semantic modeling paradigm. Furthermore, the obtained results confirmed that semantically optimized keyword representations remain effective even under noisy, ambiguous, and contextually complex social media conditions.

6. **Keyword extraction can serve as a unifying semantic layer for downstream language understanding and multimodal reasoning.**

   Collectively, the findings of this thesis establish keyword extraction as more than a traditional information retrieval or preprocessing technique. The proposed frameworks

demonstrated that semantically enriched keywords can function as a unifying intermediate semantic representation capable of supporting downstream natural language understanding, contextual reasoning, sentiment interpretation, offensiveness prediction, and multimodal semantic analysis. This unified perspective bridges interpretable semantic modeling with practical downstream social media applications, thereby providing a scalable and computationally effective alternative to purely black-box language understanding systems.

7.2 Key Contributions

The research presented in this thesis contributes to the advancement of semantic keyword extraction and social media understanding through the development of interpretable, optimization-driven, and context-aware methodologies capable of operating across both textual and multimodal environments. The principal contributions of the thesis are summarized as follows:

1. **Development of a Unified Semantic Keyword Extraction Framework**

   This thesis introduced a unified semantic keyword extraction paradigm capable of modeling contextual relevance across heterogeneous data modalities, including textual social media content and multimodal image-text memes. Unlike traditional frequency-based extraction techniques, the proposed frameworks incorporated semantic representations, contextual interactions, and multimodal understanding mechanisms to preserve deeper linguistic and contextual information.

2. **Design of Optimization-Driven Keyword Selection Strategies**

   The research proposed optimization-oriented keyword extraction methodologies that jointly balance semantic relevance, contextual diversity, redundancy reduction, and task-specific semantic alignment. By formulating keyword extraction as a principled optimization problem, the proposed approaches generated semantically discriminative keyword subsets that consistently outperformed conventional heuristic and statistical baselines across multiple downstream tasks.

3. **Introduction of Entropy-Energy Based Semantic Modeling**

   An entropy-energy inspired semantic modeling framework was developed to capture subtle, implicit, and context-dependent linguistic cues frequently observed in noisy social media environments. The proposed thermodynamic formulation integrated semantic

intensity estimation with contextual diversity modeling, enabling effective identification of sarcasm, toxicity, implicit offensiveness, and multimodal semantic hostility.

4. **Integration of Keyword Extraction with Downstream Social Media Applications**

   The thesis demonstrated that semantically enriched keyword representations can be effectively integrated with downstream social media analysis tasks without requiring heavy end-to-end supervised architectures. The extracted keywords served as interpretable semantic representations for feature construction, contextual reasoning, and downstream prediction, thereby preserving transparency while maintaining strong predictive performance.

5. **Comprehensive Validation Across Multiple Social Media Applications**

   The robustness, scalability, and generalization capability of the proposed methodologies were validated across diverse downstream applications, including toxicity detection in tweets, humor analysis from text, and offensive meme classification. Extensive experimental evaluation, comparative analysis, cross-validation, and statistical significance testing consistently demonstrated the effectiveness and practical applicability of the proposed keyword-centric semantic modeling frameworks.

7.3 Limitations of the Study

Although the proposed methodologies demonstrated strong performance across multiple social media analysis tasks, certain limitations remain that should be acknowledged. Identifying these limitations is important not only for critically assessing the scope of the present research, but also for motivating future improvements and extensions of the proposed frameworks.

1. **Dependence on Pretrained Semantic Representations**

   Several components of the proposed frameworks utilize pretrained transformer-based embedding models, sentiment analyzers, and sarcasm detection systems for contextual semantic understanding. While these models provide powerful linguistic representations, their effectiveness is inherently influenced by the quality, diversity, and biases of the data on which they were originally trained. As a result, certain domain-specific expressions, culturally localized references, or newly emerging social media patterns may not always be represented accurately within the learned semantic space.

2. **Dynamic Nature of Social Media Language**

   Social media language evolves rapidly through continuously emerging slang, abbreviations, symbolic expressions, and meme-driven communication styles. Although the proposed keyword extraction methodologies demonstrated robustness across multiple benchmark datasets, the frameworks may require periodic semantic adaptation or retraining in order to effectively capture evolving linguistic behavior over time. Static lexical resources and pretrained representations may become less effective when confronted with newly emerging online expressions.

3. **Limited Modeling of Cultural and Contextual Knowledge**

   Certain forms of humor, sarcasm, toxicity, and offensive intent are highly dependent on cultural context, historical references, and community-specific interpretations. While the proposed methodologies incorporate contextual semantics and multimodal understanding mechanisms, some deeply implicit social or cultural meanings may remain difficult to capture using semantic embeddings and optimization-based keyword modeling alone. This limitation becomes particularly relevant in meme analysis, where meaning frequently depends on external contextual knowledge beyond the observable textual and visual content.

4. **Computational Overhead of Optimization-Based Approaches**

   The optimization-driven nature of several proposed frameworks introduces additional computational complexity when compared with traditional frequency-based keyword extraction methods. Processes involving semantic clustering, contextual interaction modeling, and multi-objective optimization may increase execution time and resource requirements, particularly when applied to large-scale real-time social media streams or continuously evolving multimodal datasets.

5. **Restricted Linguistic Diversity of Experimental Datasets**

   The experimental evaluation conducted in this thesis primarily focused on English-language datasets. Consequently, the applicability of the proposed methodologies to multilingual, low-resource, or code-mixed social media environments has not been extensively explored. Since online communication frequently involves mixed-language expressions and region-specific linguistic structures, additional adaptation strategies may be required to generalize the proposed frameworks effectively across diverse linguistic

settings.

6. **Absence of Conversational and Temporal Context Modeling**

   The proposed frameworks primarily operate at the level of individual tweets, textual instances, or meme samples. However, online social media interactions often evolve through conversational exchanges, temporal progression, and user interaction histories. The absence of explicit temporal and conversational context modeling may therefore limit the ability of the proposed methodologies to fully capture evolving semantic intent and long-range contextual dependencies in dynamic online environments.

7. **Sensitivity to Hyperparameter and Threshold Selection**

   Certain components of the proposed methodologies depend on empirically selected thresholds and hyperparameters, including clustering boundaries, toxicity thresholds, and offense score decision criteria. Although sensitivity analyses were conducted to identify stable operating ranges, variations in dataset characteristics may still influence optimal parameter selection across different application domains.

Despite these limitations, the proposed research establishes a comprehensive and interpretable foundation for semantically enriched keyword-centric social media analysis. The identified limitations additionally highlight several promising opportunities for future research involving multilingual semantic modeling, adaptive learning strategies, contextual reasoning, and scalable real-time deployment frameworks.

7.4 Future Scope

Although the proposed research demonstrated the effectiveness of semantically enriched keyword extraction across multiple downstream applications, several promising directions remain for future investigation and methodological enhancement.

1. **Multilingual and Code-Mixed Semantic Keyword Modeling**

   The methodologies developed in this thesis can be extended to multilingual and code-mixed social media environments, where linguistic diversity and mixed-language communication introduce additional semantic challenges. Future work may focus on developing language-agnostic or cross-lingual semantic representations capable of preserving contextual relevance across low-resource languages and culturally diverse online

communities [191, 192]. Such extensions would improve the applicability of keyword-centric semantic modeling in multilingual online ecosystems.

2. **Hybrid Neural-Optimization Based Architectures**

   Future research may explore deeper integration between optimization-driven keyword selection strategies and neural representation learning frameworks. Combining semantic optimization mechanisms with transformer-based encoders could enable adaptive end-to-end architectures that preserve both predictive capability and interpretability while dynamically learning task-specific semantic representations. Such hybrid frameworks may further improve semantic generalization while reducing dependence on fully opaque end-to-end pipelines.

3. **Temporal and Dynamic Semantic Keyword Evolution**

   The current research primarily focuses on static semantic extraction from individual instances. Future extensions may investigate dynamic keyword extraction frameworks capable of modeling semantic evolution over time within streaming social media environments. Such approaches could enable continuous tracking of emerging trends, evolving discourse patterns, online behavioral shifts, and temporally changing semantic relationships in real-time communication settings.

4. **Advanced Multimodal Semantic Fusion**

   The proposed keyword-centric modeling paradigm can be extended toward richer multimodal environments involving additional information sources such as audio, video, temporal interaction sequences, and user behavioral signals. Incorporating multiple complementary modalities may further improve contextual reasoning capability for complex multimodal understanding tasks [193, 194]. Such extensions could support richer semantic interpretation in multimedia social media analysis and content moderation systems.

5. **Robust and Uncertainty-Aware Semantic Selection**

   Future work may incorporate uncertainty estimation and robustness-oriented optimization objectives within the keyword selection process. Such extensions could improve the reliability of semantic extraction under noisy, ambiguous, incomplete, or adversarial social media conditions. Integrating uncertainty-aware semantic modeling may addition-

ally support the development of more trustworthy and stable AI-driven content analysis systems.

6. **Extension to Broader Safety-Critical Applications**

   The interpretable keyword-centric semantic modeling frameworks proposed in this thesis can be extended to additional safety-critical natural language processing applications, including misinformation detection, cyberbullying analysis, hate speech detection, online abuse monitoring, crisis communication analysis, and automated content moderation systems. The transparency and interpretability of keyword-driven semantic reasoning make such frameworks particularly suitable for socially sensitive deployment environments requiring explainable AI systems.

Overall, the proposed research establishes a flexible and extensible foundation for future development of semantically informed, interpretable, and optimization-driven language understanding systems capable of addressing increasingly complex real-world social media analysis challenges.

7.5 Social Impact and Relevance to Sustainable Development Goals

The research undertaken in this thesis focuses on the development of optimization-based and semantically informed computational approaches for automatic keyphrase extraction and the analysis of textual and multimodal social media content. The proposed methodologies address challenges associated with identifying meaningful information from complex, heterogeneous, and noisy digital data. The research has implications for intelligent information processing, the organization and accessibility of digital knowledge, and the computational analysis of harmful online communication.

The United Nations Sustainable Development Goals (SDGs) provide a framework for relating scientific research and technological innovation to broader societal objectives. In accordance with the guidelines prescribed by Delhi Technological University for incorporating SDGs into Ph.D. research, the mapping presented in this section identifies specific research components and their relevance to the selected goals. Only those goals having a meaningful and justifiable association with the research are considered, and the strength of contribution is assessed as Strong, Moderate, or Weak based on the relevance and supporting evidence.

Figure 7.1 presents the SDGs associated with the research undertaken in this thesis. Five goals are considered: **SDG 9: Industry, Innovation and Infrastructure**, **SDG 4: Quality Education**, **SDG 16: Peace, Justice and Strong Institutions**, **SDG 3: Good Health and Well-being**, and **SDG 17: Partnerships for the Goals**. Their relevance is assessed in relation to the specific research methodologies, experimental investigations, and potential applications developed in the thesis.



Figure 7.1: Research contributions to selected United Nations Sustainable Development Goals.

The research-specific mapping of the selected SDGs, together with the assessed strength of contribution, is presented in Table 7.1.

SDG 9: Industry, Innovation and Infrastructure

The thesis makes a **strong contribution to SDG 9** through the development and evaluation of optimization-driven methodologies for intelligent information processing. The proposed approaches employ techniques such as fuzzy Ant Colony Optimization, Bayesian belief updating, contextual Principal Component Analysis, fuzzy TOPSIS, adaptive entropy-based weighting, and NSGA-II across different research components. These techniques address challenges associated with semantic representation, keyword ranking, multi-criteria evaluation, and multimodal content analysis. The development of such computational methodologies constitutes

Table 7.1: Research-specific mapping of the thesis to selected United Nations Sustainable Development Goals.

S. No.	SDG (Number and Official Title)	Relevance of the SDG to the Research Work	Strength of Contribution
1	**SDG 9: Industry, Innovation and Infrastructure**	The thesis develops optimization-based and semantically informed methodologies for automatic keyphrase extraction and multimodal content analysis. The integration of fuzzy systems, graph-based modelling, semantic representations, and multi-criteria decision-making advances computational approaches for processing complex digital information.	**Strong**
2	**SDG 4: Quality Education**	The proposed keyphrase extraction methods have potential applications in automatic indexing, document organization, concept identification, summarization, and information retrieval. These capabilities may improve the discoverability of educational and scientific documents.	**Moderate**
3	**SDG 16: Peace, Justice and Strong Institutions**	The research on multimodal offensive content detection integrates textual, visual, and semantic information to identify potentially harmful social media content. The evaluated framework provides a computational basis for offensive content analysis and may support content moderation-related applications.	**Strong**
4	**SDG 3: Good Health and Well-being**	The identification of offensive and abusive online content may support the development of safer digital environments and help reduce users' exposure to distressing communication. However, the thesis does not directly evaluate health or psychological well-being outcomes.	**Weak**
5	**SDG 17: Partnerships for the Goals**	The research integrates techniques from natural language processing, semantic knowledge representation, graph modelling, fuzzy systems, optimization, and multimodal analysis. This integration demonstrates interdisciplinary methodological relevance, although formal partnerships or collaborative development outcomes are not directly evaluated.	**Weak**

a central research objective of the thesis and provides the primary basis for this SDG mapping.

SDG 4: Quality Education

The thesis makes a **moderate contribution to SDG 4** through the potential application of automatic keyphrase extraction to educational and scientific information resources. The proposed methodologies can facilitate the identification of important concepts and support automatic indexing, document organization, summarization, and information retrieval. These capabilities may improve access to relevant information in large digital repositories used by students, educators, and researchers. The contribution is classified as moderate because the research develops computational methods that may be incorporated into educational information systems, rather than directly designing and evaluating dedicated educational interventions or platforms.

SDG 16: Peace, Justice and Strong Institutions

The thesis makes a **strong contribution to SDG 16** through its investigation of offensive and toxic content in social media environments. The proposed multimodal offensive content detection framework combines textual, visual, and semantic information to identify potentially harmful digital content. Its evaluation on benchmark datasets demonstrates the applicability of the proposed computational approach to the analysis of offensive online communication. The research may support the development of content analysis and moderation-support systems that contribute to safer digital environments. The strength of this contribution arises from the direct alignment between a substantive research component of the thesis and the identification and analysis of harmful online content.

SDG 3: Good Health and Well-being

The thesis has a **weak contribution to SDG 3** through the potential indirect benefits of identifying offensive and abusive online content. Exposure to distressing digital communication may adversely affect users' online experiences, and computational approaches for identifying such content may assist moderation-related processes. However, the research does not directly investigate mental health, measure psychological outcomes, or evaluate interventions for improving well-being. Accordingly, the relevance to this goal is recognized as a potential societal implication rather than a demonstrated health outcome.

SDG 17: Partnerships for the Goals

The thesis has a **weak contribution to SDG 17** in terms of interdisciplinary methodological integration. The proposed research combines techniques from natural language processing, semantic knowledge representation, graph modelling, fuzzy systems, optimization, and multimodal data analysis. This integration demonstrates how complementary computational approaches can be combined to address complex information-processing problems. Nevertheless, methodological integration alone does not establish the existence of institutional partnerships, formal collaborations, or measurable partnership outcomes. The association is therefore limited to the interdisciplinary nature of the research rather than a direct contribution to partnership-building objectives.

Overall, the SDG mapping identifies **SDG 9** and **SDG 16** as the goals most directly aligned with the principal research contributions of the thesis. **SDG 4** is supported by the potential application of keyphrase extraction to knowledge organization and retrieval, whereas **SDG 3** and **SDG 17** have weaker associations through prospective societal benefits and interdisciplinary methodological integration, respectively. The mapping is grounded in the specific objectives and components of the research and distinguishes demonstrated methodological contributions from potential practical applications. It complements the broader discussion of social impact by identifying the particular SDGs to which the research has a meaningful and defensible connection.

7.6 Conclusion

This thesis, titled *"Optimization Techniques Driven Approaches to Automatic Keyphrase Extraction"*, presented a comprehensive study on semantically enriched and optimization-driven methodologies for automatic keyphrase extraction across textual, social media, and multimodal environments. The research addressed key limitations of conventional frequency-based and heuristic extraction techniques, particularly their inability to effectively preserve contextual semantics, handle noisy heterogeneous data, and support complex downstream language understanding tasks.

The proposed work formulated keyphrase extraction as a semantic optimization problem by integrating graph-based representations, contextual semantic modeling, probabilistic reasoning, fuzzy decision mechanisms, quantum-inspired interaction modeling, and multimodal semantic

analysis. Through this formulation, the proposed methodologies improved semantic relevance, contextual coverage, redundancy reduction, diversity preservation, and interpretability within extracted keyword representations.

The thesis introduced multiple frameworks for structured documents, social media text, and multimodal content, including FuzBake, TOPSKey, ForestRank, ComKey, QuToX, HumorKey, and a knowledge-enriched semantic keyword extraction framework with NSGA-II optimization for offensive meme detection. Collectively, these frameworks demonstrated the effectiveness of optimization-driven semantic keyword modeling in capturing contextual semantics, emotional polarity, semantic incongruity, toxicity-oriented interactions, and multimodal offensive behavior across diverse online communication environments.

The extracted semantic keyword representations were further utilized for downstream applications including toxic tweet detection, humor detection, and offensive meme classification. Experimental evaluation across multiple benchmark datasets, supported by comparative analysis, cross-validation, sensitivity analysis, ablation studies, and statistical significance testing, consistently demonstrated the robustness, effectiveness, scalability, and interpretability of the proposed methodologies. The results showed that optimization-driven semantic keyword extraction can achieve competitive and often superior performance compared with traditional machine learning, graph-based, and transformer-driven baseline approaches while preserving semantic transparency and explainability.

Overall, this thesis establishes optimization-driven automatic keyphrase extraction as a unified and interpretable semantic representation paradigm for downstream natural language understanding and multimodal reasoning. The proposed research contributes toward advancing semantically informed, scalable, and responsible artificial intelligence systems for complex social media analysis and language understanding applications.

References

[1] Hanh Thi Hong Tran, Matej Martinc, Jaya Caporusso, Julien Delaunay, Antoine Doucet, and Senja Pollak. Recent advances in automatic term extraction: A comprehensive survey. *ACM Computing Surveys*, 58(9):1–35, 2026.

[2] Jana Sedlakova, Paola Daniore, Andrea Horn Wintsch, Markus Wolf, Mina Stanikic, Christina Haag, Chloé Sieber, Gerold Schneider, Kaspar Staub, Dominik Alois Ettlin, et al. Challenges and best practices for digital unstructured data enrichment in health research: A systematic narrative review. *PLOS digital health*, 2(10):e0000347, 2023.

[3] Karen White. Publications output: Us trends and international comparisons. science & engineering indicators 2020. nsb-2020-6. *National Science Foundation*, 2019.

[4] Tadashi Nomoto. Keyword extraction: a modern perspective. *SN Computer Science*, 4 (1):92, 2022.

[5] M Song, Y Feng, and L Jing. A survey on recent advances in keyphrase extraction from pre-trained language models, findings of the association for computational linguistics: Eacl 2023. 2023. *DOI*, 10:v1, 2023.

[6] Varun Musunuru, Sai Pavan Krishna Nagam, Praneeth Villuri, Gopikrishna Vasantha, et al. Surveying keyword extractors: Classification, applications, and empirical analysis. In *2024 Parul International Conference on Engineering and Technology (PICET)*, pages 1–8. IEEE, 2024.

[7] Matej Martinc, Hanh Thi Hong Tran, Senja Pollak, and Boshko Koloski. Seke: specialised experts for keyword extraction. *arXiv preprint arXiv:2412.14087*, 2024.

[8] Jingyu Zhang, Xinyi Yan, Yi Xiang, Yingyi Zhang, and Chengzhi Zhang. Building a multimodal dataset of academic paper for keyword extraction. *Proceedings of the Association for Information Science and Technology*, 61(1):435–446, 2024.

[9] Xuanyu ZHOU, Lin LIU, Xiao LU, Xuan LI, and Siming ZHANG. A chinese keyphrase extraction method for multimodal information enhancement representation. *Journal of Tsinghua University (Science and Technology)*, 64(10):1785–1796, 2024.

[10] Chenghang Lai and Shoumeng Qiu. Mker: multi-modal knowledge extraction and reasoning for future event prediction. *Complex & Intelligent Systems*, 11(2):138, 2025.

[11] Peter D Turney. Learning algorithms for keyphrase extraction. *Information retrieval*, 2: 303–336, 2000.

[12] Anette Hulth. Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 216–223, 2003.

[13] Ian H Witten, Gordon W Paynter, Eibe Frank, Carl Gutwin, and Craig G Nevill-Manning. Kea: Practical automatic keyphrase extraction. In *Proceedings of the fourth ACM conference on Digital libraries*, pages 254–255, 1999.

[14] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[15] Gonenc Ercan and Ilyas Cicekli. Using lexical chains for keyword extraction. *Information Processing & Management*, 43(6):1705–1714, 2007.

[16] Narjes Nikzad-Khasmakhi, Mohammad-Reza Feizi-Derakhshi, Meysam Asgari-Chenaghlu, Mohammad-Ali Balafar, Ali-Reza Feizi-Derakhshi, Taymaz Rahkar-Farshi, Majid Ramezani, Zoleikha Jahanbakhsh-Nagadeh, Elnaz Zafarani-Moattar, and Mehrdad Ranjbar-Khadivi. Phraseformer: Multimodal key-phrase extraction using transformer and graph embedding. *arXiv preprint arXiv:2106.04939*, 2021.

[17] Subramani Pasupathi, K Senthamil Selvan, Suruchi G Dedgaonkar, Manish Vishnoi, Sourav Rampal, and Dhananjay Yadav. Unlocking the power of text mining for information understanding. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE, 2024.

[18] Antonio Usai, Marco Pironti, Monika Mital, and Chiraz Aouina Mejri. Knowledge discovery out of text data: a systematic review via text mining. *Journal of knowledge management*, 22(7):1471–1488, 2018.

[19] S Archana and E Mathiselvan. Information extraction and knowledge discovery in biomedical engineering and health informatics. In *Computational Intelligence and Blockchain in Biomedical and Health Informatics*, pages 39–54. CRC Press, 2024.

[20] Blaž Škrlj, Andraž Repar, and Senja Pollak. Rakun: Ra nk-based k eyword extraction via un supervised learning and meta vertex aggregation. In *International Conference on Statistical Language and Speech Processing*, pages 311–323. Springer, 2019.

[21] Reza Yousefi Maragheh, Chenhao Fang, Charan Chand Irugu, Parth Parikh, Jason Cho, Jianpeng Xu, Saranyan Sukumar, Malay Patel, Evren Korpeoglu, Sushant Kumar, et al. Llm-take: Theme-aware keyword extraction using large language models. In *2023 IEEE International Conference on Big Data (BigData)*, pages 4318–4324. IEEE, 2023.

[22] Laura Plaza, Mark Stevenson, and Alberto Díaz. Resolving ambiguity in biomedical text to improve summarization. *Information Processing & Management*, 48(4):755–766, 2012.

[23] Tingting Wei, Yonghe Lu, Huiyou Chang, Qiang Zhou, and Xianyu Bao. A semantic approach for text clustering using wordnet and lexical chains. *Expert Systems with applications*, 42(4):2264–2275, 2015.

[24] Jingxia Ma. [retracted] research on keyword extraction algorithm in english text based on cluster analysis. *Computational Intelligence and Neuroscience*, 2022(1):4293102, 2022.

[25] Aadeesh Bali, Aniket Bhagwat, Aditya Bhise, and Sarang Joshi. Semantic similarity detection and analysis for text documents. In *2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)*, pages 1–9. IEEE, 2024.

[26] Hanan A Asiri and Muhammad Shoaib. Arabic fake news detection on x (twitter) using bi-lstm algorithm and bert embedding. *IEEE Access*, 13:189372–189386, 2025.

[27] Amit Sheth, Valerie L Shalin, and Ugur Kursuncu. Defining and detecting toxicity on social media: context and knowledge are key. *Neurocomputing*, 490:312–318, 2022.

[28] Jason Chuang, Christopher D Manning, and Jeffrey Heer. "without the clutter of unimportant words" descriptive keyphrases for text visualization. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 19(3):1–29, 2012.

[29] Avaneesh Kumar Yadav, Ranvijay, Rama Shankar Yadav, and Ashish Kumar Maurya. State-of-the-art approach to extractive text summarization: a comprehensive review. *Multimedia Tools and Applications*, 82(19):29135–29197, 2023.

[30] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[31] Rui Wang, Yanan Wang, Ziang Li, Haitao Cheng, and Guozi Sun. Distributed keyword-guided topic model with lexical knowledge supervision. *ACM Transactions on Knowledge Discovery from Data*, 19(6):1–19, 2025.

[32] Marco Dorigo and Thomas Stützle. The ant colony optimization metaheuristic: Algorithms, applications, and advances. *Handbook of metaheuristics*, pages 250–285, 2003.

[33] Mohammad Reza Bonyadi and Zbigniew Michalewicz. Particle swarm optimization for single objective continuous space problems: a review. *Evolutionary computation*, 25(1):1–54, 2017.

[34] Alex A Freitas. A critical review of multi-objective optimization in data mining: a position paper. *ACM SIGKDD Explorations Newsletter*, 6(2):77–86, 2004.

[35] Mehdi Hosseinzadeh Aghdam, Nasser Ghasem-Aghaee, and Mohammad Ehsan Basiri. Text feature selection using ant colony optimization. *Expert systems with applications*, 36(3):6843–6853, 2009.

[36] Mohammad Syafrullah and Naomie Salim. Improving term extraction using particle swarm optimization techniques. *arXiv preprint arXiv:1002.4041*, 2010.

[37] Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer, 2003.

[38] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. Yake! collection-independent automatic keyword extractor. In *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40*, pages 806–810. Springer, 2018.

[39] Stuart J Rose, Wendy E Cowley, Vernon L Crow, and Nicholas O Cramer. Rapid automatic keyword extraction for information retrieval and analysis, March 6 2012. US Patent 8,131,735.

[40] Xingbing Liu, Zhen Zhang, Baoxia Li, and Fei Zhang. Keywords extraction method for technological demands of small and medium-sized enterprises based on lda. In *2019 Chinese Automation Congress (CAC)*, pages 2855–2860, 2019. doi: 10.1109/CAC48633.2019.8996936.

[41] Kamal Sarkar, Mita Nasipuri, and Ghose Suranjan. A new approach to keyphrase extraction using neural networks. *International Journal of Computer Science Issues*, 7, 04 2010.

[42] Erick Odhiambo Omuya, George Onyango Okeyo, and Michael Waema Kimwele. Feature selection for classification using principal component analysis and information gain. *Expert Systems with Applications*, 174:114765, 2021.

[43] Qi Liu, Wenjun Ke, Xiaoguang Yuan, Yuting Yang, Hua Zhao, and Peng Wang. Adaptiveuke: Towards adaptive unsupervised keyphrase extraction with gated topic modeling. *Expert Systems with Applications*, page 123926, 2024.

[44] Bui Quang Hung, Masanori Otsubo, Yoshinori Hijikata, and Shogo Nishida. Hits algorithm improvement using semantic text portion. *Web Intelligence and Agent Systems: An International Journal*, 8(2):149–164, 2010.

[45] Mr Anuj Joshi and Priyanka Patel. Google page rank algorithm and it's updates. In *International Conference on Emerging Trends in Science, Engineering and Management, ICETSEM-2018*, 2018.

[46] Corina Florescu and Cornelia Caragea. Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1105–1115, 2017.

[47] Adrien Bougouin, Florian Boudin, and Béatrice Daille. Topicrank: Graph-based topic ranking for keyphrase extraction. In *International joint conference on natural language processing (IJCNLP)*, pages 543–551, 2013.

[48] Marina Litvak, Mark Last, Hen Aizenman, Inbal Gobits, and Abraham Kandel. Degext—a language-independent graph-based keyphrase extractor. In *Advances in Intelligent Web Mastering–3: Proceedings of the 7th Atlantic Web Intelligence Conference, AWIC 2011, Fribourg, Switzerland, January, 2011*, pages 121–130. Springer, 2011.

[49] Saroj Kr Biswas, Monali Bordoloi, and Jacob Shreya. A graph based keyword extraction model using collective node weight. *Expert Systems with Applications*, 97:51–59, 2018.

[50] Aidin Zehtab-Salmasi, Mohammad-Reza Feizi-Derakhshi, and Mohamad-Ali Balafar. Frake: Fusional real-time automatic keyword extraction. *arXiv preprint arXiv:2104.04830*, 2021.

[51] Tuohang Li, Liang Hu, Hongtu Li, Chengyu Sun, Shuai Li, and Ling Chi. Tripleranк: An unsupervised keyphrase extraction algorithm. *Knowledge-Based Systems*, 219: 106846, 2021.

[52] Chun Xu, Xian-Ling Mao, Cheng-Xin Xin, Yu-Ming Shang, Tian-Yi Che, Hong-Li Mao, and Heyan Huang. Hcuke: A hierarchical context-aware approach for unsupervised keyphrase extraction. *Knowledge-Based Systems*, 304:112511, 2024.

[53] Chun Xu, Xian-Ling Mao, Tian-Yi Che, HongLi Mao, and Heyan Huang. Sdrank: A shallow-to-deep ranking framework for enhanced unsupervised keyphrase extraction. *Expert Systems with Applications*, page 126748, 2025.

[54] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[55] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[56] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[57] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[58] Maarten Grootendorst. Keybert: Minimal keyword extraction with bert., 2020. URL https://doi.org/10.5281/zenodo.4461265.

[59] Yifeng Sun, Huan Qiu, Yanan Zheng, Zhi Wang, and Chao Zhang. Sifrank: A new baseline for unsupervised keyphrase extraction based on pre-trained language model. *arXiv preprint arXiv:2010.01379*, 2020.

[60] David B Bracewell, Fuji Ren, and Shingo Kuriowa. Multilingual single document keyword extraction for information retrieval. In *2005 international conference on natural language processing and knowledge engineering*, pages 517–522. IEEE, 2005.

[61] Zhiyuan Liu, Peng Li, Yabin Zheng, and Maosong Sun. Clustering to find exemplar terms for keyphrase extraction. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 257–266, 2009.

[62] Ángel Hernández-Castañeda, René Arnulfo García-Hernández, Yulia Ledeneva, and Christian Eduardo Millán-Hernández. Language-independent extractive automatic text summarization based on automatic keyword extraction. *Computer Speech & Language*, 71:101267, 2022.

[63] Peng Zhou, Xin Jiang, and Shu Zhao. Unsupervised technical phrase extraction by incorporating structure and position information. *Expert Systems with Applications*, 245:123140, 2024.

[64] Eibe Frank, Gordon W Paynter, Ian H Witten, Carl Gutwin, and Craig G Nevill-Manning. Domain-specific keyphrase extraction. jcai'99: Proceedings of the sixteenth international joint conference on artificial intelligence, 668-673, 1999.

[65] Olena Medelyan, Eibe Frank, and Ian H Witten. Human-competitive automatic topic indexing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1318–1327. ACL, 2009.

[66] Kuo Zhang, Hui Xu, Jie Tang, and Juanzi Li. Keyword extraction using support vector machine. In *Advances in Web-Age Information Management: 7th International Confer-*

ence, WAIM 2006, Hong Kong, China, June 17-19, 2006. Proceedings 7, pages 85–96. Springer, 2006.

[67] Bao Hong and Deng Zhen. An extended keyword extraction method. *Physics Procedia*, 24:1120–1127, 2012.

[68] Chengzhi Zhang. Automatic keyword extraction from documents using conditional random fields. *Journal of Computational Information Systems*, 4(3):1169–1180, 2008.

[69] Jie Tang, Juan-Zi Li, Ke-Hong Wang, and Yue-Ru Cai. Loss minimization based keyword distillation. In *Asia-Pacific Web Conference*, pages 572–577. Springer, 2004.

[70] Yu Zhang, Mingxiang Tuo, Qingyu Yin, Le Qi, Xuxiang Wang, and Ting Liu. Keywords extraction with deep neural network model. *Neurocomputing*, 383:113–121, 2020.

[71] Mingyang Song, Lin Xiao, and Liping Jing. Learning to extract from multiple perspectives for neural keyphrase extraction. *Computer Speech & Language*, 81:101502, 2023.

[72] Y. Luo, Q. Zhou, and F. Zhou. Enhancing phrase representation by information bottleneck guided text diffusion process for keyphrase extraction. *Proceedings of LREC 2024*, 2024.

[73] Alex Tsvetkov and Avi Kipnis. Entropyrank: Unsupervised keyphrase extraction via side-information optimization for language model-based text compression. *arXiv preprint arXiv:2308.13399*, 2023.

[74] P. Mishra, L. Pattanaik, A. Sundar, N. Yadav, and M. Kulkarni. Clustering-based sampling for few-shot cross-domain keyphrase extraction. *Amazon Science*, 2023.

[75] Reetika Singh and Goonjan Jain. Forestrank: Automatic keyphrase extraction leveraging random forest classifier and multi-criteria decision-making. In *2025 5th Asia Conference on Information Engineering (ACIE)*, pages 62–67. IEEE, 2025.

[76] Reetika Singh and Goonjan Jain. Fubake: Bayesian gamified fuzzy ant colony optimization for semantic-driven unsupervised keyphrase extraction. *Applied Soft Computing*, page 114104, 2025.

[77] Siba Prasad Pati and Rasmita Rautray. Sentence selection for extractive text summarization using topsis approach. *Procedia Computer Science*, 235:1532–1538, 2024.

[78] Tianxiong Wang. A novel approach of integrating natural language processing techniques with fuzzy topsis for product evaluation. *Symmetry*, 14(1):120, 2022.

[79] Reetika Singh and Goonjan Jain. Tops-key: Semantic keyphrase extraction using contextual principal component analysis and fuzzy technique for order preference by similarity to ideal solution with adaptive shannon entropy weights. *The European Journal on Artificial Intelligence*, page 30504554251395957, 2026.

[80] Willyan D Abilhoa and Leandro N De Castro. A keyword extraction method from twitter messages represented as graphs. *Applied Mathematics and Computation*, 240: 308–325, 2014.

[81] Abdelghani Bellaachia and Mohammed Al-Dhelaan. Ne-rank: A novel graph-based keyphrase extraction in twitter. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 372–379. IEEE, 2012.

[82] Mena B Habib and Maurice Van Keulen. Twitterneed: A hybrid approach for named entity extraction and disambiguation for tweet. *Natural language engineering*, 22(3): 423–456, 2016.

[83] Min Yang, Yuzhi Liang, Wei Zhao, Wei Xu, Jia Zhu, and Qiang Qu. Task-oriented keyphrase extraction from social media. *Multimedia Tools and Applications*, 77(3): 3171–3187, 2018.

[84] R Devika and V Subramaniyaswamy. A semantic graph-based keyword extraction model using ranking method on big social data. *Wireless Networks*, 27(8):5447–5459, 2021.

[85] Ayush Aggarwal, Chhavi Sharma, Minni Jain, and Amita Jain. Semi supervised graph based keyword extraction using lexical chains and centrality measures. *Computación y Sistemas*, 22(4):1307–1315, 2018.

[86] Amita Jain, Kanika Mittal, and Kunwar Singh Vaisla. Flake: fuzzy graph centrality-based automatic keyword extraction. *The Computer Journal*, 65(4):926–939, 2022.

[87] Minni Jain, Grusha Bhalla, Amita Jain, and Swati Sharma. Automatic keyword extraction for localized tweets using fuzzy graph connectivity measures. *Multimedia Tools and Applications*, 81(30):42931–42956, 2022.

[88] Htet Myet Lynn, Eunji Lee, Chang Choi, and Pankoo Kim. Swiftrank: an unsupervised statistical approach of keyword and salient sentence extraction for individual documents. *Procedia computer science*, 113:472–477, 2017.

[89] KBJ Lemuel and V Subramaniyaswamy. Keyphrase extraction from twitter data—a supervised deep learning approach. In *Proceedings of Data Analytics and Management: ICDAM 2021, Volume 2*, pages 181–195. Springer, 2022.

[90] Hüma Kılıç Ünlü and Aydın Çetin. Keyword extraction as sequence labeling with classification algorithms. *Neural Computing and Applications*, 35(4):3413–3422, 2023.

[91] Qiang Liu, Yan Hui, Shangdong Liu, and Yimu Ji. Y-rank: A multi-feature-based keyphrase extraction method for short text. *Applied Sciences*, 14(6):2510, 2024.

[92] Muhammad Qasim Khan, Abdul Shahid, M Irfan Uddin, Muhammad Roman, Abdullah Alharbi, Wael Alosaimi, Jameel Almalki, and Saeed M Alshahrani. Impact analysis of keyword extraction using contextual word embedding. *PeerJ Computer Science*, 8: e967, 2022.

[93] Hyeon-Kyeong Shin, Hyewon Han, Doyeon Kim, Soo-Whan Chung, and Hong-Goo Kang. Learning audio-text agreement for open-vocabulary keyword spotting. *arXiv preprint arXiv:2206.15400*, 2022.

[94] Kumari Nishu, Minsik Cho, Paul Dixon, and Devang Naik. Flexible keyword spotting based on homogeneous audio-text embedding. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5050–5054. IEEE, 2024.

[95] Youngmoon Jung, Yong-Hyeok Lee, Myunghun Jung, Jaeyoung Roh, Chang Woo Han, and Hoon-Young Cho. Adversarial deep metric learning for cross-modal audio-text alignment in open-vocabulary keyword spotting. *arXiv preprint arXiv:2505.16735*, 2025.

[96] Youkyum Kim, Jaemin Jung, Jihwan Park, Byeong-Yeol Kim, and Joon Son Chung. Bridging the gap between audio and text using parallel-attention for user-defined keyword spotting. *IEEE Signal Processing Letters*, 2024.

[97] Yu Xi, Baochen Yang, Hao Li, Jiaqi Guo, and Kai Yu. Contrastive learning with audio discrimination for customizable keyword spotting in continuous speech. In *ICASSP*

2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 11666–11670. IEEE, 2024.

[98] Suya Zhang, Yana Zhang, Zeyu Chen, and Zhaohui Li. Vsam-based visual keyword generation for image caption. *IEEE Access*, 9:27638–27649, 2021.

[99] Fatemeh Daraee, Saeed Mozaffari, and Seyyed Mohammad Razavi. Handwritten keyword spotting using deep neural networks and certainty prediction. *Computers & Electrical Engineering*, 92:107111, 2021.

[100] Junyu Luo, Zekun Li, Jinpeng Wang, and Chin-Yew Lin. Chartocr: Data extraction from charts images via a deep hybrid framework. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1917–1925, 2021.

[101] Angelos P Giotis, Giorgos Sfikas, and Christophoros Nikou. Adversarial deep features for weakly supervised document image keyword spotting. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5. IEEE, 2022.

[102] Yifan Dong, Suhang Wu, Fandong Meng, Jie Zhou, Xiaoli Wang, Jianxin Lin, and Jinsong Su. Towards better multi-modal keyphrase generation via visual entity enhancement and multi-granularity image noise filtering. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3897–3907, 2023.

[103] Sana Khamekhem Jemni, Sourour Ammar, Mohamed Ali Souibgui, Yousri Kessentini, and Abbas Cheddad. St-keys: Self-supervised transformer for keyword spotting in historical handwritten documents. *Pattern Recognition*, 170:112036, 2026.

[104] Hamid Hassani, Mohammad Javad Ershadi, and Azadeh Mohebi. Lvtia: A new method for keyphrase extraction from scientific video lectures. *Information Processing & Management*, 59(2):102802, 2022.

[105] Da Cao, Lianhai Miao, Huigui Rong, Zheng Qin, and Liqiang Nie. Hashtag our stories: Hashtag recommendation for micro-videos via harnessing multiple modalities. *Knowledge-Based Systems*, 203:106114, 2020.

[106] Anish Gupta, Romy Blake, Camden Riley, et al. Shrse: Multimodal short video hashtag recommendation based on sentiment enhancement. *Preprint at Research Square*, May

2023. doi: 10.21203/rs.3.rs-2907733/v1. URL https://doi.org/10.21203/rs.3.rs-2907733/v1. Version 1.

[107] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33: 2611–2624, 2020.

[108] Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 32–41, 2020.

[109] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. Momenta: A multimodal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184*, 2021.

[110] Yi Zhou, Zhenhao Chen, and Huiyuan Yang. Multimodal learning for hateful memes detection. In *2021 IEEE International conference on multimedia & expo workshops (ICMEW)*, pages 1–6. IEEE, 2021.

[111] Abhishek Das, Japsimar Singh Wahi, and Siyao Li. Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2012.14891*, 2020.

[112] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021.

[113] Lanyu Shang, Yang Zhang, Yuheng Zha, Yingxi Chen, Christina Youn, and Dong Wang. Aomd: An analogy-aware approach to offensive meme detection on social media. *Information Processing & Management*, 58(5):102664, 2021.

[114] Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas Pykl, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gamback. Semeval-2020 task 8: Memotion analysis–the visuo-lingual metaphor! *arXiv preprint arXiv:2008.03781*, 2020.

[115] Parth Patwa, Sathyanarayanan Ramamoorthy, Nethra Gunti, Shreyash Mishra, S Suryavardan, Aishwarya Reganti, Amitava Das, Tanmoy Chakraborty, Amit Sheth, Asif Ekbal, et al. Findings of memotion 2: Sentiment and emotion analysis of memes. In *Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR*, 2022.

[116] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[117] Pranaydeep Singh, Nina Bauwelinck, and Els Lefever. Lt3 at semeval-2020 task 8: Multi-modal multi-task learning for memotion analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1155–1162, 2020.

[118] George-Alexandru Vlad, George-Eduard Zaharia, Dumitru-Clementin Cercel, Costin-Gabriel Chiru, and Stefan Trausan-Matu. Upb at semeval-2020 task 8: Joint textual and visual modeling in a multi-task learning architecture for memotion analysis. *arXiv preprint arXiv:2009.02779*, 2020.

[119] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[120] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[121] Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

[122] Joël Plisson, Nada Lavrac, Dunja Mladenic, et al. A rule based approach to word lemmatization. In *Proceedings of IS*, volume 3, pages 83–86, 2004.

[123] Daniel Jurafsky and James H Martin. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.

[124] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[125] Hasna Nur Hanani, Herlina Jayadianti, and Heru Cahya Rustamaji. Fuzzy string matching for semi-automatication of words with jaro winkler distance algorithm on microsoft word documents. In *Seminar nasional informatika (SEMNASIF)*, volume 1, pages 145–160, 2021.

[126] Li Yujian and Liu Bo. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095, 2007.

[127] Isabelle Augenstein, Mrinal Das, Sebastian Riedel, Lakshmi Vikraman, and Andrew McCallum. Semeval 2017 task 10: Scienceie-extracting keyphrases and relations from scientific publications. *arXiv preprint arXiv:1704.02853*, 2017.

[128] Xiaojun Wan and Jianguo Xiao. Single document keyphrase extraction using neighborhood knowledge. In *AAAI*, volume 8, pages 855–860, 2008.

[129] Florian Boudin. Unsupervised keyphrase extraction with multipartite graphs. *arXiv preprint arXiv:1803.08721*, 2018.

[130] Kamil Bennani-Smires, Claudiu Musat, Andreea Hossmann, Michael Baeriswyl, and Martin Jaggi. Simple unsupervised keyphrase extraction using sentence embeddings. *arXiv preprint arXiv:1801.04470*, 2018.

[131] Linhan Zhang, Qian Chen, Wen Wang, Chong Deng, ShiLiang Zhang, Bing Li, Wei Wang, and Xin Cao. Mderank: A masked document embedding rank approach for unsupervised keyphrase extraction. *arXiv preprint arXiv:2110.06651*, 2021.

[132] Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xiaoyan Bai. Promptrank: Unsupervised keyphrase extraction using prompt. *arXiv preprint arXiv:2305.04490*, 2023.

[133] Umair Ahmed, Charalampos Alexopoulos, Marco Piangerelli, and Andrea Polini. Bryt: Automated keyword extraction for open datasets. *Intelligent systems with applications*, 23:200421, 2024.

[134] Ronald E Walpole, Raymond H Myers, Sharon L Myers, and Keying Ye. *Probability and statistics for engineers and scientists*, volume 5. Macmillan New York, 1993.

[135] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. Springer, 1992.

[136] John W Tukey. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114, 1949.

[137] Ronald Aylmer Fisher. Statistical methods for research workers. In *Breakthroughs in statistics: Methodology and distribution*, pages 66–70. Springer, 1970.

[138] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[139] Bo Han and Timothy Baldwin. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 368–378, 2011.

[140] Jacob Eisenstein. What to do about bad language on the internet. In *Proceedings of the 2013 conference of the North American Chapter of the association for computational linguistics: Human language technologies*, pages 359–369, 2013.

[141] Kevin Gimpel, Nathan Schneider, Brendan O'connor, Dipanjan Das, Daniel P Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47, 2011.

[142] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[143] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[144] Ahmad Zareie and Rizos Sakellariou. Centrality measures in fuzzy social networks. *Information Systems*, 114:102179, 2023.

[145] Ren-Jie Hu, Qing Li, Guang-Yu Zhang, and Wen-Cong Ma. Centrality measures in directed fuzzy social networks. *Fuzzy Information and Engineering*, 7(1):115–128, 2015.

[146] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18 (1):39–43, 1953.

[147] Nan Ma, Jiancheng Guan, and Yi Zhao. Bringing pagerank to the citation analysis. *Information Processing & Management*, 44(2):800–810, 2008.

[148] Alireza Alinezhad, Javad Khalili, Alireza Alinezhad, and Javad Khalili. Waspas method. *New methods and applications in multiple attribute decision making (Madm)*, pages 93–98, 2019.

[149] George Hripcsak and Adam S Rothschild. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3): 296–298, 2005.

[150] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[151] John Von Neumann and Oskar Morgenstern. Theory of games and economic behavior: 60th anniversary commemorative edition. In *Theory of games and economic behavior*. Princeton university press, 2007.

[152] Faisal Shah Khan, Neal Solmeyer, Radhakrishnan Balu, and Travis S Humble. Quantum games: a review of the history, current state, and interpretation. *Quantum Information Processing*, 17:1–42, 2018.

[153] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[154] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, Andrew Y Ng, et al. Multimodal deep learning. In *Icml*, volume 11, pages 689–696, 2011.

[155] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019.

[156] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[157] KN Apinaya Prethi, Satheeshkumar Palanisamy, S Nithya, and Ayodeji Olalekan Salau. Edge based intelligent secured vehicle filtering and tracking system using yolo and easy-ocr. *International Journal of Intelligent Transportation Systems Research*, 23(1):330–353, 2025.

[158] Qazi Mazhar ul Haq, Shanq-Jang Ruan, Muhammad Amirul Haq, Said Karam, Jeng Lun Shieh, Peter Chondro, and De-Qin Gao. An incremental learning of yolov3 without catastrophic forgetting for smart city applications. *IEEE Consumer Electronics Magazine*, 11(5):56–63, 2021.

[159] Omkar Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association, 2015.

[160] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[161] Goonjan Jain and DK Lobiyal. Word sense disambiguation using cooperative game theory and fuzzy hindi wordnet based on conceptnet. *Transactions on Asian and Low-Resource Language Information Processing*, 21(4):1–25, 2022.

[162] Kalyanmoy Deb, Samir Agrawal, Amrit Pratap, and Tanaka Meyarivan. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: Nsga-ii. In *International conference on parallel problem solving from nature*, pages 849–858. Springer, 2000.

[163] Ghinaa Zain Nabiilah, Simeon Yuda Prasetyo, Zahra Nabila Izdihar, and Abba Suganda Girsang. Bert base model for toxic comment analysis on indonesian social media. *Procedia Computer Science*, 216:714–721, 2023.

[164] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*, 2020.

[165] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

[166] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[167] Tharindu Ranasinghe and Marcos Zampieri. Multilingual offensive language identification for low-resource languages. *Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–13, 2021.

[168] Jiele Wu, Chunhui Zhang, Zheyuan Liu, Erchi Zhang, Steven Wilson, and Chuxu Zhang. Graphbert: Bridging graph and text for malicious behavior detection on social media. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 548–557. IEEE, 2022.

[169] Sayani Ghosal and Amita Jain. Hatecircle and unsupervised hate speech detection incorporating emotion and contextual semantics. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–28, 2023.

[170] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20 (3):273–297, 1995.

[171] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[172] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

[173] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[174] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[175] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[176] George H John and Pat Langley. Estimating continuous distributions in bayesian classifiers. *arXiv preprint arXiv:1302.4964*, 2013.

[177] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[178] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48, 2020.

[179] Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. "hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, 18(2):1–36, 2024.

[180] Marcos V Conde and Kerem Turgutlu. Clip-art: Contrastive pre-training for fine-grained art classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3956–3960, 2021.

[181] Ahmad Alzu'bi, Lojin Bani Younis, Abdelrahman Abuarqoub, and Mohammad Hammoudeh. Multimodal deep learning with discriminant descriptors for offensive memes detection. *J. Data and Information Quality*, 15(3), September 2023. ISSN 1936-1955. doi: 10.1145/3597308. URL https://doi.org/10.1145/3597308.

[182] Xiuxian Wang, Lanjun Wang, Yuting Su, Hongshuo Tian, Guoqing Jin, and An-An Liu. Few-shot in-context learning for implicit semantic multimodal content detection and interpretation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[183] Google Research. Multilingual bert. https://github.com/google-research/bert/blob/master/multilingual.md, 2019. Accessed: July 28, 2025.

[184] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[185] Liunian Harold Li, Mark Yatskar, D Yin, CJ Hsieh, and KW Chang. Visualbert: A simple and performant baseline for vision and language. arxiv. *arXiv preprint arXiv:1908.03557*, 10, 2019.

[186] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[187] Pranaydeep Singh and Els Lefever. Lt3 at semeval-2021 task 6: Using multi-modal compact bilinear pooling to combine visual and textual understanding in memes. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)*, pages 1051–1055, 2021.

[188] Dushyant Singh Chauhan, SR Dhanush, Asif Ekbal, and Pushpak Bhattacharyya. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4351–4360, 2020.

[189] Gitanjali Kumari, Dibyanayan Bandyopadhyay, and Asif Ekbal. Emoffmeme: identifying offensive memes by leveraging underlying emotions. *Multimedia Tools and Applications*, 82(29):45061–45096, 2023.

[190] Prashant Kapil and Asif Ekbal. A transformer based multi task learning approach to multimodal hate speech detection. *Natural Language Processing Journal*, 11:100133, 2025. ISSN 2949-7191. doi: https://doi.org/10.1016/j.nlp.2025.100133. URL https://www.sciencedirect.com/science/article/pii/S2949719125000093.

[191] Tharun Suresh, Ayan Sengupta, Md Shad Akhtar, and Tanmoy Chakraborty. A comprehensive understanding of code-mixed language semantics using hierarchical transformer. *IEEE Transactions on Computational Social Systems*, 11(3):4139–4148, 2024.

[192] Ravindra Nayak and Raviraj Joshi. L3cube-hingcorpus and hingbert: A code mixed hindi-english dataset and bert language models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 7–12, 2022.

[193] Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau. Deep multimodal fusion for semantic image segmentation: A survey. *Image and Vision Computing*, 105: 104042, 2021.

[194] Yujian Cai, Xingguang Li, Yingyu Zhang, Jinsong Li, Fazheng Zhu, and Lin Rao. Multimodal sentiment analysis based on multi-layer feature fusion and multi-task learning. *Scientific Reports*, 15(1):2126, 2025.

List of Publications

Journal Publications

1. Reetika Singh and Goonjan Jain. "FuBAKE: Bayesian gamified fuzzy ant colony optimization for semantic-driven unsupervised keyphrase extraction." *Applied Soft Computing*, 114104, 2025
   (SCIE, Impact Factor – 6.6)

2. Reetika Singh and Goonjan Jain. "TopS-Key: Semantic Keyphrase Extraction Using Contextual Principal Component Analysis and Fuzzy Technique for Order Preference by Similarity to Ideal Solution With Adaptive Shannon Entropy Weights." *The European Journal on Artificial Intelligence*, 30504554251395957, 2026
   (SCIE, Impact Factor – 1.0)

Papers presented in Conferences

1. Reetika Singh and Goonjan Jain. "ForestRank: Automatic Keyphrase Extraction leveraging Random Forest Classifier and Multi-Criteria Decision-Making", *2025 5th Asia Conference on Information Engineering (ACIE)* at Phuket, Thailand.

2. Reetika Singh and Goonjan Jain. "HumorKey: A Novel Keyword Extraction Framework Using Bit-Level Encoding and Fisher–Yates Shuffling for Humor Detection", *2025 International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)* at Kalyani, India.

Manuscripts Communicated

1. Reetika Singh and Goonjan Jain. ‘COM-Key: Commonsense-Augmented Fuzzy Graphs with Multi-Centrality for Keyword Discovery and Knowledge Mining in Social Media Data.” (Under Revision)

2. Reetika Singh and Goonjan Jain. “QuToX: A Quantum Game-Theoretic Framework for Keyword Extraction and Lexicon-Based Toxicity Detection in Social Media Text.” (Communicated)

3. Reetika Singh and Goonjan Jain. “Knowledge-Enriched Semantic Keyword Extraction with NSGA II for Offensive Meme Detection via Thermodynamic Modeling.” (Communicated)

4. Reetika Singh and Goonjan Jain. “A Chronological Survey of Keyword Extraction Across Textual, Acoustic, and Visual Modalities (2020–Present): A Unified Perspective.” (Communicated)

Plagiarism Report

Title of the Thesis:	Optimization Techniques Driven Approaches To Automatic Keyphrase Extraction
Name of the Scholar:	Reetika Singh
Supervisor:	Dr. Goonjan Jain
Department:	Applied Mathematics

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: Turnitin, **Similarity Index:** 5%, **Total Word Count:** 71,699.

Date: 02/06/2026

turnitin Page 1 of 298 - Cover Page Submission ID trn:oid:::27535:141377145

Reetika Singh

Thesis_Draft (1)

Paper 1

Document Details

Submission ID
trn:oid:::27535:141377145

Submission Date
Jun 2, 2026, 1:46 PM GMT+5:30

Download Date
Jun 2, 2026, 3:33 PM GMT+5:30

File Name
Thesis_Draft (1).pdf

File Size
24.2 MB

283 Pages

71,699 Words

451,882 Characters

turnitin Page 1 of 298 - Cover Page Submission ID trn:oid:::27535:141377145

5% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 10 words)

Match Groups

256 Not Cited or Quoted 5%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

2% Internet sources
3% Publications
3% Submitted works (Student Papers)