DEVELOPMENT OF EFFICIENT AND INTELLIGENT FAKE NEWS DETECTION FRAMEWORK

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

DOCTOR OF PHILOSOPHY

by

Aman Jolly
(2K20/PHDCO/07)

Under the Supervision of
Prof. Shailender Kumar
Professor, Department of Computer Science and Engineering Delhi Technological University



Department of Computer Science and Engineering

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

July, 2026

ACKNOWLEDGEMENTS

"You have a right to perform your prescribed duties, but you are not entitled to the fruits of your actions." — Bhagavad Gita, 2:47

First and foremost, I bow in deepest gratitude to lord **Shri Krishna**, the Supreme Lord of the Multiverse, whose strength carried me through doubt and whose grace illuminated the path when the way was unclear. The teachings of the *Shrimad Bhagavad Gita* became the philosophical foundation of this journey. *Jai Shri Krishna.*

To my Supervisor and Guru

I am extraordinarily fortunate to have had **Prof. Shailender Kumar**, Department of Computer Science and Engineering, Delhi Technological University, as mine. His guidance was never merely technical; it was philosophical in the truest sense, teaching me not just how to build systems that detect misinformation, but how to think rigorously, question assumptions, and hold conclusions lightly. I am especially, and personally, grateful for his support during the critical juncture when I needed to convert my PhD to part-time enrolment. Many supervisors would have stepped back at that point. He did not. That act of faith changed the trajectory of this work, and I carry it with me as one of the most meaningful things anyone has done for me in my academic life.

I am also grateful to **Prof. Anil Singh Parihar**, Head of Department, Delhi Technological University, for his encouragement and institutional support throughout this doctoral journey.

I am deeply blessed to have the spiritual guidance and blessings of my Dharm Gurus, **Uddhav Mahatma Ji** and **Rupa Baiji**, who taught me that the pursuit of knowledge and the pursuit of the Divine are, at their deepest, the same road.

I am grateful to my former colleagues at **KIET Group of Institutions, Ghaziabad**, especially **Dr. Ramesh Singh**, whose early encouragement gave me the confidence to pursue this doctoral journey.

To Excelra Knowledge Solutions

For three years, they gave me the rarest professional gift: an environment where a PhD-in-progress was seen not as a distraction but as a sign of the kind of person worth keeping. I am sincerely grateful to **Dr. Chandra Sekhar** (VP, Bioinformatics), **Dr. Puneet** (Group Head), and managers **Dr. Tahseen** and **Dr. Savita** for their flexibility and encouragement. To friends and colleagues **Dr. Aatif**, **Sushant**, **Jaspreet**, **Dr. Tanusree**, and **Dr. Debamitra**: thank you for the warmth and camaraderie that made three years of parallel commitments meaningful.

To my fellow researchers at DTU **Ms. Shweta Gupta**, **Ms. Aatish Kumar**, **Mr. Nikhil Sharma**, and **Mr. Aman Kumar Pandey**, thank you for the solidarity that only fellow researchers understand.

To my family

There is a kind of support that does not announce itself; it simply persists, quietly, across years and setbacks and periods of doubt. That is the support I have received from my parents, **Mr. Pawan Jolly** and **Mrs. Santosh Jolly**, and my brother, **Naman Jolly**. They did not always understand what I was building, but they never once questioned whether I should be building it. That unconditional faith is the most durable foundation a researcher can have, and I would not have reached this point without it. And to my wife, **Dr. Neha Sharma Jolly**, you occupy a singular place in this acknowledgement that no word adequately captures. You were my first reader and my most honest critic. You absorbed my anxieties without absorbing my pessimism, and you improved this work in ways visible in every chapter while sustaining the person writing it in ways visible only to me. A thesis is ultimately a record of sustained effort over time, and you were the reason that effort was sustainable. I could not have done this without you.

Aman Jolly



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I, Aman Jolly (2K20/PHDCO/07), hereby certify that the work which is being presented in the thesis entitled **"Development of Efficient and Intelligent Fake News Detection Framework"** in partial fulfilment of the requirements for the award of the Degree of Doctor of Philosophy, submitted in the Department of Computer Science and Engineering, Delhi Technological University, is an authentic record of my own work carried out during the period from August 2020 to July 2026 under the supervision of Prof. Shailender Kumar. 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.

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor(s) **Signature of External Examiner**



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR(s)

Certified that **Aman Jolly** (enrollment no. (2K20/PHDCO/07)) has carried out their research work presented in this thesis entitled **"Development of Efficient and Intelligent Fake News Detection Framework"** for the award of **Doctor of Philosophy** from the Department of Computer Science and Engineering, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student himself 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.

Signature
Prof. Shailender Kumar
Professor
Department of Computer Science and Engineering,
Delhi Technological University, Delhi-110042

Date:

ABSTRACT

On social media a false story travels roughly six times faster than the truth and reaches about ten times as many people, carried by novelty and outrage rather than by automated bots. Platforms now generate on the order of 300 million posts a day, while the world's professional fact-checkers, working by hand, can verify only a few hundred claims a week. The gap between how fast falsehood spreads and how slowly it can be checked is where automated fake news detection becomes necessary, and its consequences are not abstract: the COVID-19 infodemic drove vaccine hesitancy and harmful self-medication, election deepfakes erode institutional trust, and fabricated reporting has been used to manipulate markets and incite violence.

Automating this task, however, runs into two problems that most detectors quietly ignore. The first is that the ground truth itself does not hold still. Scientific consensus on masks, variants, and treatments shifted from week to week during the pandemic, so a model trained on a historical snapshot learns yesterday's facts and then, with full confidence, flags today's corrected reporting as fake. As misinformation narratives keep evolving, static detectors decay, and every attempt to teach an old model a new domain tends to erase what it knew about the last, the well-known problem of catastrophic forgetting. The second is that a real claim and a fabricated one are not always cleanly separable. Conventional classifiers are forced to return a hard real-or-fake verdict, so a claim the model believes is 99% fake and one it believes is 51% fake are stamped with the same label; the model's own doubt, the very signal a human fact-checker most needs, is flattened against the decision boundary and lost. This thesis treats that lost uncertainty, together with concept drift, adversarial manipulation, unimodal bias in multimodal systems, and retrieval-evidence mismatch, as the five problems a deployable detector must confront. Eight research contributions are presented in response, spanning text-based detection, evidence-grounded retrieval, continual learning, multimodal robustness evaluation, and systematic review.

The first four contributions work on text. Three-Label CT-BERT adds a third answer, a *Can't Say* label, so that instead of forcing every claim into real or fake the model can set aside the ones it is genuinely unsure about and pull in live evidence from the Google Fact Check API; on the COVID-19 Constraint dataset this reaches 93% accuracy, 13 percentage points above a standard binary BERT. FakeNews Buster keeps a detector current as new misinformation appears: it pairs two encoders (BERT and FastFormer) with a knowledge-distillation loss, a way of letting the model learn new topics while still remembering the old ones, which cuts forgetting by 60% (forgetting rate 0.06) at an average continual-learning F1 of 0.92. RAG-for-Classification asks a simpler question than most retrieval systems do: it uses retrieved evidence to *classify* a claim with a small model (RoBERTa/DistilBERT) rather than to *generate* an answer with an expensive large language model. A controlled audit across LIAR2, FakeNewsNet, and FEVER shows that adding multi-hop graph structure, following a chain of linked facts, helps only when the task actually provides such chains (FEVER, +3.1 macro-F1 over single-hop retrieval), that the dataset and not the model architecture sets the performance ceiling, and that a compressed Fast Graph-RAG is comparable to the single-pass baseline at the lowest latency. EAKD-CFND combines per-instance

uncertainty-adaptive distillation with external fact-check verification: it gauges how uncertain it is on each example, trusts the previous model where it is confident, and queries an external fact source when its uncertainty is high. Evaluated against five baselines on three class-incremental splits, all methods, including rehearsal-based DER and plain fine-tuning, remain near 29–31% accuracy on both FakeNewsNet splits, and the exemplar-free configuration is statistically indistinguishable from them at zero storage cost, but it collapses to near-random accuracy on the five-task PHEME-Event sequence together with every other exemplar-free method, where DER's stored exemplars reach 84.8%; a rehearsal-augmented configuration (EAKD-CFND-R) recovers 85.5% there. The uncertainty signal is poorly calibrated (ECE 0.80–0.84), although accuracy is robust to its hyperparameters, and neither external verification nor instance-level weighting produced a gain outside seed noise.

The next three contributions look at systems that read text and images together. A stress test of five multimodal architectures against 17 kinds of attack uncovers *Unimodal Bias*: many models quietly lean on a single input, so removing the image makes a late-fusion model collapse to F1 = 0.00. The Synergy Gain metric $S = F1_{\text{Multi}} - \max(F1_{\text{Text}}, F1_{\text{Image}})$ measures how much a model genuinely gains from combining both inputs rather than riding one of them. The collapse is worst on the D1 dataset, whose clean scores turn out to be inflated by a hidden watermark the model had learned to exploit (a Clever-Hans shortcut), so the watermark-free ReCOVery and PolitiFact datasets are used as the more honest benchmark. Statistical tests (Wilcoxon signed-rank and paired t-test, both $p < 0.05$) show that the most damaging attack is not random noise but meaning: LLM-driven semantic inversions (DeepSeek-LLM) that flip a claim's truth while keeping it fluent, and a black-box analysis confirms this weakness carries across every architecture tested, including the state-of-the-art MFIR and CroMe.

The final contribution steps back for the wider picture. A PRISMA systematic review of 221 COVID-19 fake news detection papers (2011–May 2026) maps eleven recurring gaps in the field and argues that these lessons carry over to the next health emergency, using the 2026 Andes hantavirus outbreak as a worked example. A closing comparative evaluation places all eight contributions side by side and turns them into a practical guide for choosing a detector.

List of Publications

Journal Publications (Published):

1. Aman Jolly and Prof. Shailender Kumar "Beyond the Surface: Adversarial Stress Testing of Multimodal Fake News Detection Models with LLM-Driven Perturbations" in *Computers and Electrical Engineering*, vol. 135, art. 111203, 2026, ISSN 0045-7906 (SCIE Indexed, IF: 5.5). Elsevier. https://doi.org/10.1016/j.compeleceng.2026.111203

2. Aman Jolly and Prof. Shailender Kumar "A Rigorous Interrogation of Multimodal Fake News Detection Paradigms: Adversarial Robustness and Modality Dependence" in *Cluster Computing*, vol. 29, art. 307, 2026 (SCIE Indexed, IF: 5.5). Springer. https://doi.org/10.1007/s10586-026-06022-y

3. Aman Jolly and Prof. Shailender Kumar "A Comprehensive Evaluation of Multimodal Fake News Detection: Assessing Robustness and Efficacy" in *Arabian Journal for Science and Engineering*, 2026 (SCIE Indexed, IF: 3.1). Springer. https://doi.org/10.1007/s13369-026-11410-y

Journal Publications (Accepted):

4. Aman Jolly, Prof. Shailender Kumar and Sankalp Paliwal "EAKD-CFND: Continual Fake News Detection through Adaptive Knowledge Distillation with External Verification" accepted in *Scientific Reports* (SCIE Indexed, IF: 4.9). Nature Portfolio. Accepted on 29 August 2026.

Journal Publications (Under Review):

5. Aman Jolly and Prof. Shailender Kumar "Mitigating the Infodemic: Lessons from COVID-19 Fake News Detection for Epidemic Preparedness in the Age of Generative AI" submitted to *Information Processing & Management* (SCIE Indexed, IF: 6.9). Elsevier.

Conference Publications:

6. Aman Jolly and Prof. Shailender Kumar "Navigating the Gray Area: A Three-Label Framework for Uncovering Uncertainty in Fake News" in *International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, IEEE, Sharda University, November 2023.

7. Aman Jolly and Prof. Shailender Kumar "FakeNews Buster: A Continual Learning Approach with BERT and FastFormer for Adaptive and Explainable Fake News Detection" in *International Conference on Intelligent and Secure Engineering Solutions (CISES)*, IEEE, GL Bajaj, August 2025.

8. Aman Jolly and Prof. Shailender Kumar "Retrieval Without Generation: Evidence-Grounded Fake News Classification" in *International Conference on Communication and Intelligent Systems (ICCIS)*, Springer, BITS Pilani K. K. Birla Goa Campus, September 26–27, 2026. *(Presented)*

TABLE OF CONTENTS

Acknowledgements **ii**

Candidate's Declaration **iv**

Certificate by the Supervisor **v**

Abstract **vi**

List of Publications **viii**

List of Tables **xv**

List of Figures **xvii**

List of Abbreviations **xviii**

1 Introduction **1**
1.1 Background and Motivation 1
1.2 Problem Statement 2
1.3 Research Objectives 3
1.4 Scope of the Thesis 3
1.4.1 Ethical Considerations 4
1.5 Thesis Contributions 4
1.6 Thesis Organisation 6

2 Literature Review **10**
2.1 Introduction 10
2.2 Fake News: Definition, Taxonomy, and Challenges 10
2.3 Text-Based Fake News Detection 12
2.3.1 Classical Machine Learning Approaches 12
2.3.2 Deep Learning Approaches 13
2.3.3 Transformer-Based Methods 14
2.3.4 Continual and Incremental Learning for FND 15
2.3.5 Explainability in Fake News Detection 15
2.4 Multimodal Fake News Detection 16
2.4.1 Taxonomy of Multimodal FND Approaches 16
2.4.2 Fusion Architectures 17
2.4.3 Cross-Modal Attention Mechanisms 18
2.5 Adversarial Robustness in Fake News Detection 19
2.5.1 Adversarial Text Attacks 19

2.5.2 LLM-Generated Adversarial Content 21
2.5.3 Image-Based Adversarial Attacks 22
2.6 Knowledge Distillation and Efficient Models 22
2.7 Graph-Based and RAG-Based Approaches 23
2.7.1 Graph Neural Networks for Fake News Detection 23
2.7.2 Retrieval-Augmented Generation for Fact Verification 23
2.8 Benchmark Datasets 24
2.9 Research Gaps 25
2.10 Summary 27

3 Efficient and Intelligent Fake News Detection Framework 29
3.1 Introduction 29
3.2 Related Work 30
3.2.1 BERT-Based Text Classification for Fake News 30
3.2.2 Uncertainty and Multi-Label Classification 30
3.2.3 Continual and Incremental Learning for NLP 30
3.2.4 Graph Neural Networks and Retrieval-Augmented Generation 30
3.2.5 Uncertainty as a Training Signal 31
3.2.6 Explainability in Fake News Detection 31
3.3 Three-Label Framework for Uncertainty-Aware Fake News Detection . 31
3.3.1 Motivation: The Gray Area in Fake News 31
3.3.2 Proposed Framework Architecture 32
3.3.3 Theoretical Justification of the Probability Threshold 32
3.3.4 Dataset and Exploratory Analysis 32
3.3.5 Results and Discussion 34
3.3.6 Limitations 34
3.4 FakeNews Buster: Continual Learning with BERT and FastFormer . . 35
3.4.1 Motivation: Concept Drift in Social Media Fake News 35
3.4.2 Architecture: Dual Transformer Models 35
3.4.3 Data Processing Schemes 35
3.4.4 Continual Learning via Knowledge Distillation 36
3.4.5 Custom Adaptive Loss Function 36
3.4.6 Adaptive Explainability Module 37
3.4.7 Vocabulary Enrichment 37
3.4.8 Experimental Setup 37
3.4.9 Results 37
3.4.9.1 Single-Task Performance 37
3.4.9.2 Continual Learning Performance 37
3.4.9.3 Ablation Studies 38
3.4.10 Limitations 40
3.5 Graph-RAG Classification Framework for Misinformation Mitigation 40
3.5.1 Motivation: The Retrieval-Relevance Mismatch 40
3.5.2 The RAG-for-Classification Paradigm 41
3.5.3 Architecture Design: Three Models Along the Accuracy-Efficiency Frontier 42
3.5.3.1 Architecture 1: Standard RAG 42

3.5.3.2 Architecture 2: Graph-RAG 42
3.5.3.3 Architecture 3: Fast Graph-RAG 42
3.5.4 Knowledge Base Construction and Experimental Setup 43
3.5.5 Results and Discussion 43
3.5.6 Limitations 44
3.6 EAKD-CFND: Continual Fake News Detection via Adaptive Knowledge Distillation 45
3.6.1 Motivation: Uncertainty-Aware Distillation Without Stored Exemplars 45
3.6.2 Problem Formulation 47
3.6.3 Framework Architecture 47
3.6.4 Knowledge Distillation Module 47
3.6.5 Uncertainty Signal Generation and Adaptive Weighting 47
3.6.6 Uncertainty-Driven External Verification 49
3.6.7 Combined Loss Function 51
3.6.8 EAKD-CFND-R: A Rehearsal-Augmented Configuration . . . 51
3.6.9 Experimental Setup 51
3.6.10 Results and Discussion 52
3.6.10.1 Quantitative Performance 52
3.6.10.2 Ablation Studies 54
3.6.10.3 EAKD-CFND-R Evaluation 55
3.6.10.4 Why the Benchmarks Diverge 55
3.6.10.5 Calibration, Sensitivity, and Cost 57
3.6.10.6 Feature Comparison 57
3.6.11 Significance, Limitations, and Deployment Guidance 58
3.7 Consolidated Comparison of All Four Contributions 59
3.7.1 A Unified Framework, Not a Collection of Models 60
3.7.2 Efficiency of the Proposed Framework 60
3.8 Summary 61

4 Multimodal Fake News Detection: Robustness, Adversarial Evaluation, and Bias Analysis 62
4.1 Introduction 62
4.2 Related Work and Unified Attack Taxonomy 63
4.2.1 Multimodal Fusion Architectures 63
4.2.2 Adversarial Robustness in Multimodal Systems 63
4.2.3 The Unified Adversarial Attack Taxonomy 64
4.2.3.1 Image-Based Attacks (10) 64
4.2.3.2 Text-Based Attacks (7) 64
4.2.3.3 Modality Ablation Protocol 67
4.2.3.4 The Synergy Gain Metric 67
4.3 Comprehensive Evaluation of Multimodal FND 67
4.3.1 Motivation and Scope 67
4.3.2 Models and Datasets 67
4.3.3 Baseline Performance and Adversarial Impact 68
4.3.4 Modality Removal and Unimodal Bias 68

4.3.5 XAI Diagnosis: The Clever Hans Effect 70
4.4 Adversarial Stress Testing with LLM-Driven Perturbations 70
4.4.1 Motivation: SOTA Architectures and DeepSeek Attacks . . . 70
4.4.2 SOTA Architectures 70
4.4.3 PolitiFact: A Third Dataset 71
4.4.4 LLM Attack Examples 71
4.4.5 Results: Modality Reversal on PolitiFact and DeepSeek Superiority 72
4.4.6 Proposed Defences 73
4.5 Modality Dependence and Vision-Language Model Robustness 74
4.5.1 Motivation: Visual Backbone vs Fusion Strategy 74
4.5.2 Architecture Configurations 74
4.5.3 Modality Ablation: Backbone versus Fusion 75
4.5.4 Black-Box Transfer Attacks and Statistical Confirmation . . . 75
4.6 Unified Analysis Across the Multimodal FND Contributions 76
4.6.1 The Research Programme in Full 76
4.6.2 Conclusions 76
4.6.3 Deployment Decision Framework 77
4.7 Summary 77

5 Comparative Evaluation of Fake News Detection Frameworks and the COVID-19 Infodemic Case Study 79
5.1 Introduction 79
5.2 Mitigating the Infodemic: COVID-19 Fake News Detection and Epidemic Preparedness 80
5.2.1 The COVID-19 Infodemic: Scale, Spread, and Documented Harm 80
5.2.2 Systematic Review Methodology 81
5.2.3 Datasets and Knowledge Resources 84
5.2.4 Review of COVID-19 Fake News Detection Approaches . . . 85
5.2.4.1 Taxonomy of Detection Strategies 85
5.2.4.2 Content-Based Approaches 86
5.2.4.3 Social Context-Based Approaches 86
5.2.4.4 Hybrid Approaches 87
5.2.4.5 Emerging Paradigms (2024–2026) 87
5.2.5 Generative AI as a New Threat Vector for Epidemic Misinformation 88
5.2.6 Proposed Framework for Epidemic-Aware Fake News Detection 89
5.2.7 Results and Key Findings 90
5.2.8 Eleven Critical Research Gaps 90
5.2.9 Lessons for Future Epidemic Preparedness 91
5.2.10 Explainability in Pandemic Fake News Detection 91
5.2.11 Limitations of the Review 92
5.2.12 Cross-Domain Generalisability of the Review Findings 92
5.3 Comprehensive Comparative Evaluation of FND Frameworks 95
5.3.1 Motivation and Evaluation Protocol 95

5.3.2 Method Descriptions 95
5.3.3 Performance Comparison 96
5.3.4 Analysis: Uncertainty Handling 98
5.3.5 Analysis: Concept Drift Resilience 99
5.3.6 Analysis: Explainability and Trust 99
5.3.7 Synthesis: Deployment Decision Guide 100
5.4 Discussion 102
5.5 Summary 102

6 Conclusion, Future Scope and Social Impact **103**
6.1 Conclusion 103
6.2 Future Directions 105
6.3 Social Impact 106

References **109**

Appendix-I Hyperparameter Configurations **137**

Appendix-II Dataset Statistics **140**
II.1 Text-Only Datasets 140
II.2 Multimodal Datasets 140

Appendix-III Additional Experimental Results **143**
III.1 Complete FakeNews Buster Single-Task Results 143
III.2 EAKD-CFND Complete Baseline Comparison 144
III.3 EAKD-CFND Statistical Significance 145
III.4 Per-Attack Adversarial Robustness Summary 145

List of Publications and Proofs **148**

Plagiarism Report **156**

Curriculum Vitae **163**

LIST OF TABLES

1.1 Research Objective Mapping with Publications 7

2.1 Comparison of Fake News Categories 11
2.2 Summary of Multimodal Fake News Detection Methods 20
2.3 Major Fake News Detection Benchmark Datasets 25
2.4 Research gaps and status in this thesis 26

3.1 Classification performance: BERT-base and CT-BERT 34
3.2 Single-task F1 scores of model configurations 39
3.3 Continual Learning Performance: Average F1 and Forgetting Rate . . 39
3.4 Ablation of adaptive loss weighting 39
3.5 Ablation of vocabulary enrichment 40
3.6 Consolidated ablation of FakeNews Buster components 41
3.7 Model Configuration and Hyperparameters for RAG Architectures . . 43
3.8 Performance across RAG architectures and datasets 43
3.9 EAKD-CFND datasets and task configuration 52
3.10 Continual learning performance comparison 53
3.11 Ablation of EAKD-CFND components on FNN-Poli-Time 54
3.12 EAKD-CFND-R variants across three benchmarks (5 seeds) 56
3.13 Task-to-task topical similarity by dataset 56
3.14 Training and inference cost of EAKD-CFND 58
3.15 Feature comparison of EAKD-CFND with baseline methods 58
3.16 Consolidated comparison of the four text-based contributions 59

4.1 Taxonomy of the 17 adversarial attacks 65
4.2 Baseline F1 vs SOTA and Adversarial Attack Impact Summary 68
4.3 Modality removal results on D1 69
4.4 Screening of architectures with the Synergy Gain rule 70
4.5 LLM Adversarial Text Examples from RECOVERY Dataset 72
4.6 Modality ablation of SOTA models on PolitiFact 73
4.7 Modality ablation across five architectures (D1) 76
4.8 Scope and findings of the multimodal contributions 77
4.9 Architecture selection guide 78

5.1 Thematic Breakdown of the 221-Paper Corpus 83
5.2 Bibliometric summary of the 221-paper corpus 84
5.3 Principal COVID-19 fake news datasets 85
5.4 Representative results on COVID-19 benchmarks 87
5.5 Emerging Detection Paradigms in the 2024–2026 Sub-corpus 88

5.6 Required Detection Paradigm Shifts for the Generative AI Era 90
5.7 Transferability of the PRISMA findings 93
5.8 Eleven critical research gaps 93
5.9 Master performance comparison 96
5.10 Unified cross-method summary of the eight contributions 97
5.11 Deployment decision guide 101

6.1 Prioritisation of the future research directions 107
6.2 Illustrative social impact scenarios 108

I.1 Hyperparameter configurations of all models 138
I.2 Hyperparameter configurations (continued) 139

II.1 Text-Only Benchmark Datasets Used in This Thesis 141
II.2 Multimodal Benchmark Datasets Used in This Thesis 142

III.1 Single-task F1 scores of FakeNews Buster configurations 143
III.2 Full continual learning comparison 144
III.3 EAKD-CFND versus the five baselines 145
III.4 Per-attack robustness summary 146

LIST OF FIGURES

1.1 Visual roadmap of the thesis 7

2.1 Taxonomy of fake news detection strategies 12
2.2 General workflow of content-based fake news detection 13
2.3 General XAI workflow for fake news detection 16
2.4 Examples of textual adversarial perturbations 20

3.1 Three-label uncertainty-aware framework 33
3.2 Class distribution and text length in the Constraint dataset 33
3.3 LIME explanation of a Can't Say prediction 34
3.4 Fixlength and Recurrence data processing schemes 36
3.5 FakeNews Buster architecture 38
3.6 F1 per sequential task for continual learning methods 40
3.7 The three evaluated RAG architectures 44
3.8 F1 by architecture across the three datasets 45
3.9 Accuracy against average latency on FEVER 45
3.10 Catastrophic forgetting versus continual learning objective 46
3.11 EAKD-CFND training flow diagram 48
3.12 Uncertainty signal generation and adaptive weighting 49
3.13 Adaptive weighting function 50
3.14 Accuracy and BWT across methods and datasets 54
3.15 Ablation results on FNN-Poli-Time 55

4.1 Multimodal threat model 68
4.2 Per-architecture performance under the attack suite 69
4.3 Grad-CAM attention before and after attack 71
4.4 Adversarial perturbation on a PolitiFact sample 72
4.5 Modality ablation across architectures and datasets 73
4.6 Heatmap of F1-scores across the attack suite 74
4.7 Experimental design for backbone versus fusion 75

5.1 Google Trends interest in "fake news" 81
5.2 Review methodology and corpus distribution 82
5.3 Feature space of machine- and human-generated news 89
5.4 Three dimensions of explainability 91
5.5 Deployment decision flowchart 100

LIST OF SYMBOLS, ABBREVIATIONS AND NOMENCLATURE

Abbreviation	Full Form
FND	Fake News Detection
NLP	Natural Language Processing
ML	Machine Learning
DL	Deep Learning
LLM	Large Language Model
BERT	Bidirectional Encoder Representations from Transformers
CT-BERT	COVID-Twitter BERT
RoBERTa	Robustly Optimised BERT Pre-training Approach
DistilBERT	Distilled BERT
XLM-R	Cross-lingual Language Model – RoBERTa
mBERT	Multilingual BERT
FastFormer	Additive-Attention Efficient Transformer
CNN	Convolutional Neural Network
LSTM	Long Short-Term Memory
GNN	Graph Neural Network
VGG	Visual Geometry Group (CNN)
CLIP	Contrastive Language–Image Pre-training
ViLBERT	Vision-and-Language BERT
ViT	Vision Transformer
RAG	Retrieval-Augmented Generation
Graph-RAG	Graph-based Retrieval-Augmented Generation
KD	Knowledge Distillation
EAKD-CFND	Continual Fake News Detection through Adaptive Knowledge Distillation with External Verification (exemplar-free configuration)
EAKD-CFND-R	EAKD-CFND with a rehearsal (reservoir) buffer
ECE	Expected Calibration Error
CL	Continual Learning
EWC	Elastic Weight Consolidation
LwF	Learning without Forgetting
DER	Dark Experience Replay
BWT	Backward Transfer
FWT	Forward Transfer

Abbreviation	Full Form
F1	F1-Score (harmonic mean of precision and recall)
SD	Standard Deviation
LR	Learning Rate
BS	Batch Size
MSL	Maximum Sequence Length
LoRA	Low-Rank Adaptation
QLoRA	Quantised Low-Rank Adaptation
FGSM	Fast Gradient Sign Method
PGD	Projected Gradient Descent
MIFGSM	Momentum Iterative FGSM
LIME	Local Interpretable Model-agnostic Explanations
FEVER	Fact Extraction and VERification (dataset)
LIAR	LIAR benchmark dataset
PHEME	PHEME rumour dataset
FNN	FakeNewsNet
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
WHO	World Health Organization
ε	Perturbation budget (adversarial attack)
θ_{unc}	Uncertainty threshold (EAKD-CFND)
β	Weighting-sensitivity coefficient (attention weighting in Graph-RAG; distillation weighting in EAKD-CFND)
p	Statistical significance value

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

False news stories spread significantly faster, deeper, and more broadly than true stories on social media, a disparity driven by the novelty and emotional content of misinformation rather than by bot amplification [1]. Social media platforms such as Twitter, Facebook, YouTube, and WhatsApp permit any individual to publish content to a potentially global audience within seconds, removing the editorial gatekeeping that traditional media used to filter misinformation before publication. False, misleading, and manipulated content now competes with authentic reporting on equal distributional terms and, as the evidence shows, frequently outperforms it [1].

The societal consequences of large-scale misinformation have been extensively documented [2]. In the political domain, fabricated stories about electoral candidates, falsified vote-count reports, and state-sponsored disinformation campaigns have been shown to influence electoral outcomes, erode institutional trust, and deepen partisan polarisation. In the public health domain, the COVID-19 pandemic demonstrated that health misinformation is not merely a nuisance but a measurable public health threat: false claims about vaccine safety, denial of the pandemic's severity, and promotion of ineffective or dangerous treatments circulated at scale and risked harmful health behaviour [3]. The World Health Organization coined the term *infodemic* to describe this parallel crisis, recognising that health misinformation at scale constitutes a direct threat to public safety.

Manual fact-checking, while essential, cannot operate at the scale required to address this challenge. Professional fact-checking organisations (FactCheck.org, AFP Fact Check, PolitiFact, and Boomlive among others) collectively produce hundreds of verified assessments per week. Social media platforms generate hundreds of millions of posts per day. The gap between verification capacity and publication volume is not bridgeable through human labour alone. Automated detection systems are therefore necessary if misinformation is to be addressed at the speed and scale at which it spreads.

Early automated detection systems treated fake news detection as a text classification problem, applying classical machine learning approaches, including support vector machines, logistic regression, and naïve Bayes classifiers, to hand-crafted linguistic features such as word frequency distributions, syntactic patterns, and readability indices [4]. These approaches achieved acceptable performance on curated datasets

but failed in deployment conditions due to their reliance on surface-form features easily manipulated by adversaries. BERT [5] changed the state of the art by enabling models to learn deep contextual representations of text, raising accuracy on benchmark datasets. However, as misinformation evolved in sophistication, so did its resistance to detection: adversarial text perturbations, multimodal manipulation combining false claims with authentic imagery, and the emergence of large language model (LLM)-generated content that is stylistically indistinguishable from authentic journalism have progressively eroded the effectiveness of even the most advanced transformer-based detectors.

The thesis presented here addresses these interconnected challenges through a programme of eight research contributions spanning text-based detection, advanced evidence retrieval, continual learning, multimodal robustness evaluation, and systematic review.

1.2 Problem Statement

Fake news is defined herein following Shu et al. [4] as news articles that are intentionally and verifiably false and could mislead readers. This definition distinguishes fake news from satire (which is not intended to deceive), from errors in reporting (which are unintentional), and from opinion (which is not verifiably false). It encompasses fabricated facts, false context applied to genuine events, misleading headlines, and synthetic content generated by automated systems.

Several interconnected problems constrain the effectiveness of current automated fake news detection systems:

The forced-choice problem: Existing FND systems almost universally employ binary classification, assigning every article to a *real* or *fake* category. In the latent representation space of pre-trained language models, the decision boundary between these categories is not a sharp line but a continuous region in which many articles are genuinely ambiguous, including newly published claims without fact-checker coverage, satire that may be mistaken for reporting, or evolving health guidance that is accurate today and contradicted by evidence tomorrow. Forcing classification of such articles into one of two hard labels produces systematic misclassification that cannot be corrected by improved modelling alone, because the error is architectural rather than parametric.

Concept drift: Social media misinformation evolves continuously: topics shift, language adapts, adversaries modify their tactics in response to deployed detectors. Models trained on static datasets degrade predictably as the data distribution diverges from the training distribution [6]. Retraining from scratch is computationally prohibitive; naïve fine-tuning on new data causes catastrophic forgetting of previously learned patterns.

Adversarial manipulation: All current content-based detectors are vulnerable to adversarial text perturbations that preserve the semantic content of a claim for human readers while flipping the model's prediction. Character-level noise, synonym substitutions, and grammar distortions have all been shown to degrade state-of-the-art transformer models [7], extending to text the adversarial-example phenomenon first

demonstrated for images [8]. More critically, large language models such as DeepSeek-LLM and Llama 3 enable adversaries to generate semantically coherent inversions of factual claims that are imperceptible to human reviewers and yet highly effective against automated detectors.

Multimodal manipulation and unimodal bias: Contemporary misinformation increasingly exploits the synergy between manipulated images and false text [9]: a photograph of a genuine event paired with a fabricated caption, or a politician's image paired with a false quote. Multimodal detection models that jointly encode textual and visual features are designed to address this threat. However, empirical evaluation reveals a systematic architectural failure: these models disproportionately learn to rely on whichever single modality carries more discriminative signal in the training dataset, producing near-total performance collapse when that modality is removed or perturbed, a phenomenon termed Unimodal Bias.

Retrieval and evidence grounding: Pattern-based classifiers that learn statistical regularities of fake versus real content from training data fail when confronted with novel misinformation that does not match patterns seen during training, particularly AI-generated content that mimics the stylistic features of authentic journalism. Grounding detection in external, dynamically updated knowledge bases (the retrieval-augmented paradigm) offers a principled alternative, but standard retrieval architectures suffer from a retrieval-relevance mismatch that limits their effectiveness on multi-hop verification tasks.

The COVID-19 infodemic and epidemic preparedness: The COVID-19 pandemic demonstrated that none of the above problems can be considered theoretical: they manifested at scale, with measurable public health consequences, in the largest misinformation event in recorded history. Understanding what automated detection systems can and cannot do in this domain, and identifying the specific gaps that prevented effective automated response, is both a scientific obligation and a prerequisite for preparedness against future epidemic infodemics.

1.3 Research Objectives

This thesis pursues four formally stated research objectives:

RO 1: To review different Fake News Detection techniques.

RO 2: To propose an efficient and intelligent Fake News Detection framework.

RO 3: To investigate the effectiveness of multi modal fake news detection by inspecting the biasness of multimodal fake news detection system

RO 4: To perform a comparative result analysis of the developed model with the other existing techniques.

1.4 Scope of the Thesis

This thesis covers text-based fake news detection on social media content; multimodal fake news detection combining textual and visual modalities; adversarial robustness evaluation of multimodal detection systems against image-based and text-based attacks including LLM-generated perturbations; continual and incremental learning for

fake news detection; retrieval-augmented generation as a detection paradigm; and the COVID-19 infodemic as a systematic domain-specific survey case study. Evaluation is conducted on established benchmark datasets: Constraint 2021, LIAR, FakeNewsNet (PolitiFact and GossipCop), FEVER, PHEME, D1, ReCOVery, and the MediaEval propagation dataset.

1.4.1 Ethical Considerations

Automated fake news detection acts on public speech, so the ethical framing of this work is stated explicitly. First, no new data were collected from human participants; all experiments use established public benchmarks (Constraint 2021, LIAR, FakeNewsNet, FEVER, PHEME, D1, ReCOVery, PolitiFact-based multimodal data, and the MediaEval propagation dataset), which is why formal ethics approval was not required. Second, external services such as the Google Fact Check Tools API are queried only for verified fact-check records under the provider's terms of use. Third, the adversarial studies (Chapter 4) are dual-use: LLM-generated perturbations can also be used to evade detectors. They were therefore executed only in a laboratory setting on benchmark data, no attack was deployed against a live platform, and every attack study is paired with a discussion of defences. Fourth, the thesis documents failure modes that can produce unfair outcomes, namely Unimodal Bias, dataset-artefact shortcuts, English-only evaluation, and miscalibrated uncertainty, so that deployment decisions can account for them. Finally, the framework is designed to support, not replace, human fact-checkers: the *Can't Say* class, evidence-grounded explanations, and the deployment guide of Chapter 5 keep consequential moderation decisions with human reviewers.

1.5 Thesis Contributions

The research presented in this thesis addresses critical challenges in automated fake news detection through comprehensive design, implementation, and evaluation. The key contributions of this work are outlined below, each reflecting a significant advancement toward achieving the research objectives. These contributions collectively highlight the originality, technical depth, and practical relevance of the proposed efficient and intelligent fake news detection framework.

A. **Comprehensive Review of Fake News Detection Techniques:**

- Conducted a structured review of fake news detection techniques across text-based, multimodal, adversarial, knowledge-distillation, and retrieval-augmented paradigms, together with a PRISMA-protocol systematic review of 221 COVID-19 publications spanning 2011–2026 [10].
- Identified critical research gaps, including binary-classification rigidity, concept drift, adversarial vulnerability, unimodal bias, the retrieval-relevance mismatch, and the epidemic-preparedness shortfall.

B. **Uncertainty-Aware Text Classification Framework:**

- Proposed a three-label framework extending CT-BERT with a probability-threshold *Can't Say* class and dynamic dataset enrichment through the Google Fact Check API, replacing forced binary verdicts with honest uncertainty representation [11].
- Achieved 93% accuracy on the COVID-19 Constraint dataset, a 13 percentage-point improvement over binary BERT baselines.

C. **Continual Learning Framework with Adaptive Knowledge Distillation:**

- Developed a dual-encoder (BERT + FastFormer) continual learning architecture with an adaptive knowledge-distillation loss to counter concept drift and catastrophic forgetting, achieving an average continual-learning F1 of 0.92 and a forgetting rate of 0.06 (a 60% reduction over fine-tuning) [12].
- Proposed per-instance uncertainty-weighted distillation with uncertainty-triggered external verification in an exemplar-free (EAKD-CFND) and a rehearsal-augmented (EAKD-CFND-R) configuration. Exemplar-free, it ties DER and fine-tuning near 30% accuracy on both FakeNewsNet splits and collapses on PHEME-Event (16.8% versus 84.8% for DER); EAKD-CFND-R reaches 85.5%, equal to DER within noise. Verification and instance weighting yielded no gain beyond noise, and the uncertainty signal is poorly calibrated (ECE 0.80–0.84) [13].

D. **Evidence-Grounded RAG-for-Classification Framework:**

- Introduced the *RAG-for-Classification* paradigm, which brings retrieval-based evidence grounding to compact discriminative classifiers (RoBERTa/DistilBERT) rather than costly generative LLMs, keeping the conditioning evidence human-readable [14].
- Through a controlled audit of three architectures across LIAR2, FakeNewsNet, and FEVER, showed that multi-hop graph structure helps specifically on tasks with genuine evidence chains (FEVER, +3.1 macro-F1) while task structure, not architecture, sets the ceiling; the compressed Fast Graph-RAG matches the best accuracy at the lowest latency, making it the deployable choice.

E. **Adversarial Robustness Evaluation and Unimodal Bias Formalisation:**

- Performed the first systematic black-box transfer attack analysis across multimodal FND architectures (including BERT-CLIP, ViLBERT, MFIR, and CroMe) under 17 adversarial attacks, statistically confirming ($p < 0.05$) that LLM-driven DeepSeek perturbations outperform Llama3 and that text-based attacks dominate image-based ones [15, 16].
- Formalised *Unimodal Bias* and introduced the Synergy Gain metric ($S = F1_{\text{Multi}} - \max(F1_{\text{Text}}, F1_{\text{Image}})$), documenting catastrophic collapse (F1 = 0.00) of fusion models when a dominant modality is removed [17].

F. **Comparative Evaluation and Deployment Decision Guide:**

- Conducted a unified cross-method comparison of all proposed frameworks across text, multimodal, adversarial, and continual learning dimensions on common benchmarks.
- Synthesised the findings into a deployment decision guide mapping operational scenarios to the most suitable proposed framework.

1.6 Thesis Organisation

This thesis is organised into six chapters, each addressing a distinct facet of building an efficient and intelligent fake news detection framework. The organisation follows the research objectives and provides a step-by-step development from problem formulation, through text-based and multimodal detection methods, to adversarial robustness analysis and a comprehensive comparative evaluation.

Chapter 1: Introduction begins with the motivation behind automated fake news detection. It outlines the problem statement, research objectives, scope, contributions, and the organisation of the thesis.

Chapter 2: Literature Review provides a comprehensive review of existing fake news detection techniques, spanning a structured FND taxonomy, text-based methods, multimodal methods, adversarial robustness, knowledge distillation, and retrieval-augmented approaches. It identifies the critical research gaps that motivate the contributions of this thesis and addresses Research Objective 1.

Chapter 3: Efficient and Intelligent FND Framework develops the core text-based detection framework. It presents the Three-Label CT-BERT model [11] (93% accuracy), FakeNews Buster [12] (average continual-learning F1 = 0.92, forgetting rate = 0.06), the RAG-for-Classification (Graph-RAG) audit [14] (graph structure helps on evidence-chain tasks such as FEVER), and EAKD-CFND [13] (exemplar-free ties DER and fine-tuning near chance on FakeNewsNet and collapses on PHEME-Event; 85.5% there with the rehearsal-augmented EAKD-CFND-R). It addresses Research Objective 2.

Chapter 4: Multimodal FND — Robustness, Adversarial Evaluation, and Bias investigates multimodal fake news detection and its biases. It formalises Unimodal Bias and the Synergy Gain metric [17], conducts an LLM-driven adversarial stress test of multiple architectures [16], and presents a black-box transfer attack analysis across vision-language models [15]. It addresses Research Objective 3.

Chapter 5: Comparative Evaluation and the COVID-19 Infodemic Case Study presents a comprehensive cross-method comparison of all eight contributions across text, multimodal, adversarial, and continual learning dimensions, together with a deployment decision guide, grounded in a COVID-19 infodemic case study (a PRISMA systematic review of 221 papers, eleven gaps, and the 2026 hantavirus case) [10]. It addresses Research Objective 4, with the COVID-19 infodemic serving as an applied case study.

Chapter 6: Conclusion, Future Directions & Social Impact audits the fulfilment of the research objectives, summarises the key contributions, discusses limitations and future research directions, and reflects on the social impact of the work.

Supplementary material is given in three appendices: Appendix-I lists the hyperparameter configurations of all models, Appendix-II gives dataset statistics, and Appendix-III reports complete experimental results, including the full EAKD-CFND comparison and its significance tests. Fig. 1.1 summarises the organisation visually: each research objective is traced to the chapter that fulfils it and to the main content of that chapter.

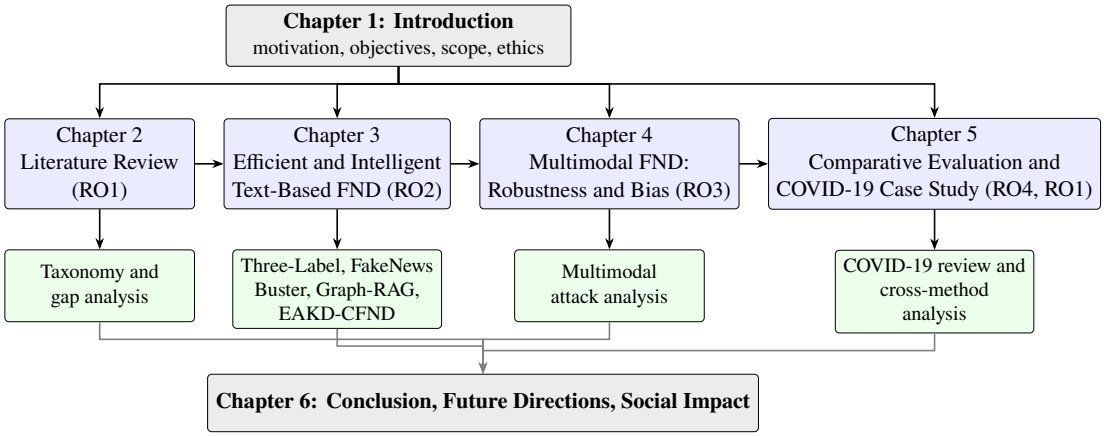


Figure 1.1: Visual roadmap of the thesis: research objectives (RO1–RO4), the chapters that fulfil them, and the main content of each chapter.

Table 1.1 maps each research objective to the publications that fulfil it, together with their current status.

Table 1.1: Research Objective Mapping with Publications

Research Objectives	**Publication(s) Remarks**
RO 1. To review different Fake News Detection techniques.	✓ "A Comprehensive Evaluation of Multimodal Fake News Detection: Assessing Robustness and Efficacy," Arabian Journal for Science and Engineering, Springer [17]. *(Published)* ✓ "Mitigating the Infodemic: Lessons from COVID-19 Fake News Detection for Epidemic Preparedness in the Age of Generative AI," Information Processing & Management, Elsevier [10]. *(Under Review)*
RO 2. To propose an efficient and intelligent Fake News Detection framework.	✓ "Navigating the Gray Area: A Three-Label Framework for Uncovering Uncertainty in Fake News," IEEE ICCCIS [11]. *(Presented)* ✓ "FakeNews Buster: A Continual Learning Approach with BERT and FastFormer for Adaptive and Explainable Fake News Detection," IEEE CISES [12]. *(Presented)* ✓ "Retrieval Without Generation: Evidence-Grounded Fake News Classification," Springer ICCIS 2026 [14]. *(Presented)*

continued on next page

Table 1.1 (continued)

Research Objectives	Publication(s) Remarks
	✓ "EAKD-CFND: Continual Fake News Detection through Adaptive Knowledge Distillation with External Verification," Scientific Reports, Nature Portfolio [13]. *(Accepted)*
RO 3. To investigate the effectiveness of multimodal fake news detection by inspecting the biasness of multimodal fake news detection systems.	✓ "A Rigorous Interrogation of Multimodal Fake News Detection Paradigms: Adversarial Robustness and Modality Dependence," Cluster Computing, Springer [15]. *(Published)*
	✓ "Beyond the Surface: Adversarial Stress Testing of Multimodal Fake News Detection Models with LLM-Driven Perturbations," Computers and Electrical Engineering, Elsevier [16]. *(Published)*
	✓ "A Comprehensive Evaluation of Multimodal Fake News Detection: Assessing Robustness and Efficacy," Arabian Journal for Science and Engineering, Springer [17]. *(Published)*
RO 4. To perform a comparative result analysis of the developed model with the other existing techniques.	✓ "Navigating the Gray Area: A Three-Label Framework for Uncovering Uncertainty in Fake News," IEEE ICCCIS [11]. *(Presented)*
	✓ "FakeNews Buster: A Continual Learning Approach with BERT and FastFormer for Adaptive and Explainable Fake News Detection," IEEE CISES [12]. *(Presented)*
	✓ "Retrieval Without Generation: Evidence-Grounded Fake News Classification," Springer ICCIS 2026 [14]. *(Presented)*
	✓ "A Rigorous Interrogation of Multimodal Fake News Detection Paradigms: Adversarial Robustness and Modality Dependence," Cluster Computing, Springer [15]. *(Published)*

continued on next page

Table 1.1 (continued)

Research Objectives	Publication(s) Remarks
	✓ "Beyond the Surface: Adversarial Stress Testing of Multimodal Fake News Detection Models with LLM-Driven Perturbations," Computers and Electrical Engineering, Elsevier [16]. *(Published)*
	✓ "A Comprehensive Evaluation of Multimodal Fake News Detection: Assessing Robustness and Efficacy," Arabian Journal for Science and Engineering, Springer [17]. *(Published)*
	✓ "EAKD-CFND: Continual Fake News Detection through Adaptive Knowledge Distillation with External Verification," Scientific Reports, Nature Portfolio [13]. *(Accepted)*
	✓ "Mitigating the Infodemic: Lessons from COVID-19 Fake News Detection for Epidemic Preparedness in the Age of Generative AI," Information Processing & Management, Elsevier [10]. *(Under Review)*

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Fake news detection now draws on natural language processing, computer vision, graph neural networks, knowledge representation, and continual learning. This chapter provides a structured review of the literature that directly informs the contributions of this thesis. The review is organised along the principal technical paradigms through which fake news detection has been approached: definitional and taxonomic foundations (Section 2.2); text-based detection methods spanning classical machine learning through modern transformers (Section 2.3); multimodal detection methods (Section 2.4); adversarial robustness (Section 2.5); knowledge distillation and efficient modelling (Section 2.6); graph-based and retrieval-augmented approaches (Section 2.7); benchmark datasets (Section 2.8); and a structured summary of research gaps (Section 2.9). This chapter directly addresses RO 1 of the thesis: to review different Fake News Detection techniques. The scope of the review is grounded in the PRISMA-protocol systematic review conducted in the COVID-19 Infodemic Review [10], which screened 850 records across six academic databases and included 221 publications spanning 2011–May 2026; the PRISMA selection process and publication distribution are presented in Chapter 5 (Fig. 5.2).

2.2 Fake News: Definition, Taxonomy, and Challenges

Shu et al. [4] defined fake news as news articles that are intentionally and verifiably false, capable of misleading readers. This definition distinguishes deliberate deception from honest error and from satire, placing the emphasis on intent and verifiability. Zhou and Zafarani [9] extended this to a broader taxonomy of information disorders, differentiating misinformation (false content shared without intent to harm), disinformation (false content shared with intent to harm), and malinformation (true content shared with intent to harm). Within the fake news category specifically, the following subtypes are conventionally recognised:

- *Fabricated content:* wholly invented stories with no grounding in fact, typically designed to generate engagement or manipulate opinion.
- *Misleading content:* genuine information used in a misleading way, such as accurate statistics presented with false context.

- *False context:* genuine content that is re-published in a different context with a false framing (e.g., photographs from one event labelled as depicting another).
- *Imposter content:* authentic sources impersonated to lend credibility to false claims.
- *Satire and parody:* content not intended to deceive but which may be misread or re-shared without context.

Table 2.1 compares these fake news categories along their defining characteristics, intent, and representative examples.

Table 2.1: Comparison of Fake News Categories

Category	**Intent**	**Verifiability**	**Detection Challenge**
Fabricated	Deceptive	Falsifiable	High: no anchor in real events
Misleading	Deceptive	Partial	High: accurate data, false narrative
False context	Deceptive	Verifiable	Medium: image/source verification needed
Imposter	Deceptive	Verifiable	Medium: source authentication
Satire/Parody	Non-deceptive	N/A	Low intent but may be misread
AI-generated (LLM)	Varies	Often falsifiable	Very high: stylometric features identical to authentic content

The distinction between content types matters practically: a system designed to detect fabricated stories by learning statistical regularities of fake writing style will fail on false-context manipulation, which may consist entirely of authentic text paired with an out-of-context image. These category-level distinctions directly motivate the evidence-grounded approaches of Chapter 3 and the multimodal approaches of Chapter 4.

Automated detection methods are likewise taxonomically organised into three primary paradigms, as illustrated in Fig. 2.1 [10]: content-based strategies (knowledge-based, style-based, and latent feature methods that operate on the news item itself); social-context-based strategies (source-credibility and propagation-based methods that exploit network structure); and hybrid approaches that combine both signals. The subsequent sections of this chapter follow this taxonomy.

Comprehensive survey papers covering the breadth of fake news detection methods include Zhou and Zafarani [9], Meel and Shrihari [18], Nasir et al. [19], Duma et al. [20], Athira et al. [21], Jadhav et al. [22], and Arega et al. [23]. AI-based fake news detection is surveyed in Mohseni et al. [24] and Ali et al. [25]. The intersection of misinformation

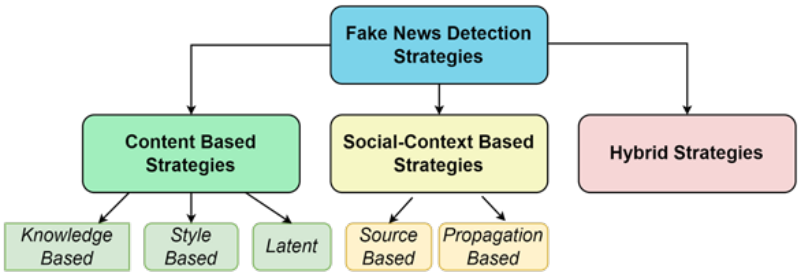


Figure 2.1: Taxonomy of fake news detection strategies reviewed in this chapter, organised into content-based (knowledge-based, style-based, latent feature), social-context-based (source-based, propagation-based), and hybrid paradigms [10].

with emerging AI technologies including generative models is examined in Xie et al. [26]. Federated learning approaches to privacy-preserving fake news detection are surveyed in Zhang et al. [27].

The social and cognitive dimensions of misinformation spread have been studied extensively. Allcott and Gentzkow [28] documented the scale and partisan asymmetry of fake news consumption in the 2016 US election. Hergovich et al. [29] established that confirmation bias and prior belief affect the evaluation of evidence, explaining why correction efforts are often ineffective. Blockchain-based approaches for establishing content provenance and combating source impersonation are explored in Yang et al. [30]. Fact-checking organisations including FactCheck.org [31], Agence France-Presse Fact Check [32], and Poynter's International Fact-Checking Network [33] provide the human verification infrastructure that automated systems must eventually complement at scale.

2.3 Text-Based Fake News Detection

Content-based fake news detection methods operate directly on the news item without reference to external propagation signals. Three sub-paradigms exist, as illustrated in Fig. 2.2: *style-based* methods extract linguistic regularities using ML/DL classifiers trained on curated labelled datasets; *latent feature* methods learn deep representations via neural networks; and *knowledge-based* methods compare Subject-Predicate-Object (SPO) triples extracted from content against a verified knowledge base [10].

2.3.1 Classical Machine Learning Approaches

Early automated fake news detection employed hand-crafted features extracted from news text, classified using support vector machines (SVM), logistic regression (LR), and naïve Bayes (NB) classifiers [4]. Feature engineering focused on lexical features (word frequency, character n-gram distributions), syntactic features (part-of-speech tag sequences, parse tree depth), psychological word category loadings from LIWC (Linguistic Inquiry and Word Count), and readability metrics such as Flesch-Kincaid scores. These approaches achieved accuracy in the range of 70–85% on available

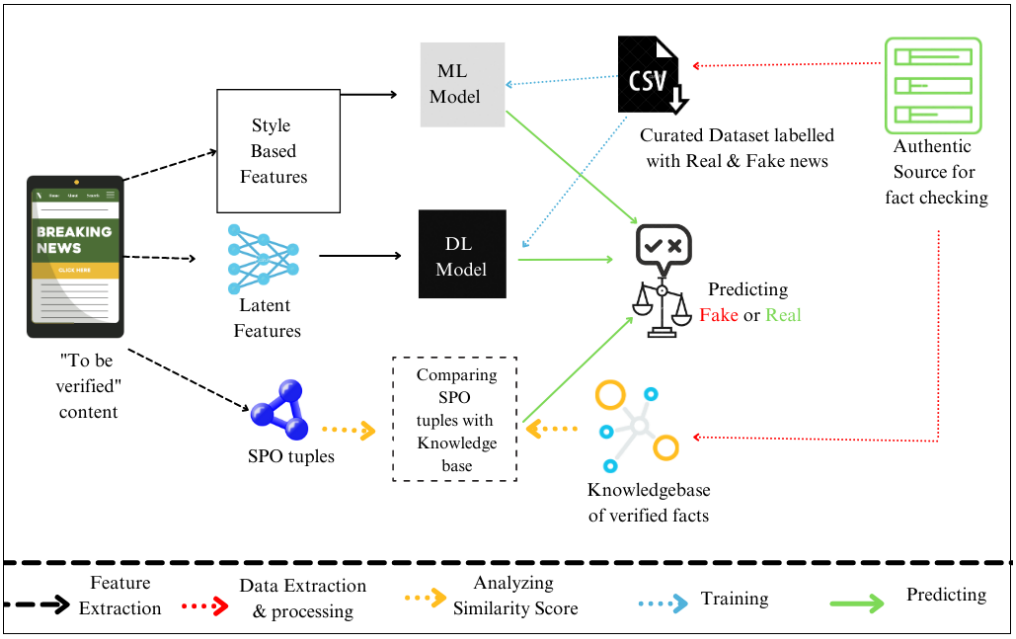


Figure 2.2: General workflow of content-based fake news detection. Style-based (top) and latent feature-based (middle) approaches train classifiers on curated labelled datasets; knowledge-based approaches (bottom) extract SPO triples and compare against a verified knowledge base to predict real or fake [10].

benchmarks but suffered from three limitations: dependence on static, hand-crafted feature sets; inability to capture semantic context across sentence boundaries; and vulnerability to simple adversarial modifications that preserve category membership while changing surface features.

SVM and related classical classifiers for FND are evaluated in Kaur et al. [34] and Khan et al. [35], with benchmark comparisons across multiple datasets. Alshuwaier and Alsulaiman [36] provides a recent comparative evaluation of classical approaches. AI-based FND using machine learning explanations to guard against overtrust is studied in Mohseni and Ragan [37] and Mohseni et al. [38]. Cross-lingual fake news detection using transfer features is explored in Li et al. [39]. SHAP-based explainability applied to classical ML classifiers in the FND context is presented in Sorokina et al. [40]. Sharma et al. [41] provides an AI-focused review of fake news detection approaches. Surveys covering both classical and deep learning methods include Mridha et al. [42], Du et al. [43], and Song et al. [44].

2.3.2 Deep Learning Approaches

The introduction of convolutional neural networks (CNNs) for text classification demonstrated that n-gram features could be learned automatically from raw text embeddings rather than hand-crafted. TextCNN [45] applied parallel convolutional filters of multiple widths to word embedding sequences, aggregating through max-pooling to produce document representations. Long Short-Term Memory networks (LSTMs) [46] and Gated Recurrent Units (GRUs) extended this capability to sequential text modelling,

capturing positional dependencies and long-range contextual relationships. Bidirectional RNNs [47] further improved contextual coverage. Attention mechanisms [48] were integrated to weight the contribution of different tokens, yielding hierarchical attention networks (HANs) that structure encoding at the word, sentence, and document levels. ELMo [49] introduced context-dependent character-level representations as a precursor to the transformer revolution. These approaches consistently outperformed classical ML methods, achieving 85–92% on benchmark datasets, while remaining dependent on training data quality and unable to capture inter-sentence semantic relationships.

CNN architectures applied specifically to fake news detection include Singh and Sabrol [50]. LSTM-based approaches for FND are explored in Padalko et al. [51]. Attention-based classification is studied in Lee et al. [52]. Contextualized detection of satire, a boundary case for deep learning methods, is examined in Kamal and Abulaish [53]. Emoji representations in social media text are studied in Naudan [54] and Pimentel [55], establishing the importance of emoji-aware tokenization for social media FND.

2.3.3 Transformer-Based Methods

The transformer architecture [56] and BERT [5] (Bidirectional Encoder Representations from Transformers) became the dominant approach for NLP-based fake news detection. BERT's bidirectional pre-training on large corpora enables it to capture subtle semantic cues that distinguish misinformation from credible reporting: hedge language, emotional register, sourcing patterns, and logical consistency. Fine-tuned BERT variants consistently achieve 95–99% accuracy on standard benchmarks [57]. Domain-adaptive variants such as CT-BERT [58], pre-trained specifically on COVID-19 tweets, show that domain specialisation improves accuracy on domain-specific tasks. ALBERT [59] reduced BERT's parameter count through cross-layer parameter sharing. RoBERTa [60] improved over BERT through optimised pretraining procedures. XLNet [61] extended the paradigm to autoregressive generation with permutation language modelling. Ensemble methods combining multiple transformer variants have achieved near-ceiling performance on major benchmarks [62], establishing a benchmark saturation problem where reported accuracy no longer differentiates system quality.

Transformer-based fake news detection systems evaluated on COVID-19 data include Chen et al. [62], Wani et al. [57], Vijjali et al. [63], Koirala [64], Al-Ahmad et al. [65], Shifath et al. [66], Gundapu and Mamidi [67], and Shishah [68]. Fake news detection using XLNet is described in Gautam et al. [69], Gautam et al. [69], and Gautam et al. [69]. The Longformer [70] extends transformer context to 4,096 tokens using sparse attention, enabling processing of full news articles beyond BERT's 512-token limit. BERT applications in fake news are examined in Fu et al. [71] and Jain [72]. Domain-specific pretrained models for financial misinformation are explored in Kamal et al. [73] and Mohankumar et al. [74]. Advances in pretrained language models for FND are surveyed in Khan [75] and Al-Zahrani and Al-Yahya [76].

Uncertainty-aware transformer approaches for FND include Arora et al. [77] and Vlachos [78]. Large language models applied directly to fake news detection, leveraging in-context learning and instruction following, include Wu et al. [79], Ma et al. [80], Wang et al. [81], Cao et al. [82], Wang et al. [83], Wang et al. [84], and Yin et al. [85].

Synergistic combinations of LLMs and retrieval for FND are explored in Hu et al. [86]. AI-generated content detection using language models is addressed in Chen and Shu [87]. Dynamic representation learning for evolving FND is presented in Wen et al. [88]. QLoRA fine-tuning for parameter-efficient adaptation of LLMs is described in Dettmers et al. [89] and Dettmers et al. [90], enabling cost-effective domain adaptation for adversarial content generation in the FND context.

2.3.4 Continual and Incremental Learning for FND

A critical limitation of transformer-based FND systems is their inability to maintain performance in dynamically evolving misinformation environments [6]. The phenomenon of *catastrophic forgetting*, wherein neural networks trained on new data forget patterns learned from prior data, is well-documented in the continual learning literature [91]. Elastic Weight Consolidation (EWC) [91] addresses this through a regularisation penalty on parameter updates for weights deemed important to past tasks. Learning without Forgetting (LwF) employs knowledge distillation from the previous model as a soft-label regulariser. Dark Experience Replay (DER) [92] combines replay buffers with distillation. Hasan et al. [93] showed that incremental deep learning can adapt fake news detectors to concept drift, but at the cost of forgetting on prior-task performance. The FakeNews Buster [12] and EAKD-CFND [13] contributions of this thesis (Chapter 3) investigate this problem through adaptive loss scheduling and instance-level uncertainty-guided distillation, respectively; the latter is evaluated against rehearsal-based and adaptive-KD baselines and indicates that exemplar-free continual learning fails under severe drift (Chapter 3).

Concept drift analysis specific to fake news detection and related text stream mining is provided in Alrowili and Roy [94], Alrowili and Roy [94], and Saha et al. [95]. Class-incremental learning strategies are examined in Kang et al. [96]. Knowledge distillation for neural network optimisation in the continual learning context is explored in Ratna [97] and Song et al. [98]. Continual learning for fake news is further studied in Rahaman et al. [99]. Graph-based continual learning is presented in Han et al. [100]. Densely connected networks for continual representation learning are proposed in Tian et al. [101]. Adaptive loss weighting for knowledge distillation, studied for speech-recognition models, is described in Ganguly et al. [102]. Dynamic representation learning via optimal-weighting frameworks relevant to streaming FND is explored in Feng et al. [103].

2.3.5 Explainability in Fake News Detection

Explainability is a system-level requirement that is orthogonal to detection accuracy: a classifier that achieves 99% F1 but provides no interpretable rationale for its decisions cannot be operationally deployed in editorial, legal, or public health contexts where human accountability is mandated [10]. Fig. 2.3 illustrates the general XAI pipeline applicable to FND systems: the input is processed by a black-box classifier to produce a prediction; an interpretability technique (LIME, SHAP, Grad-CAM, or attention attribution) runs in parallel to produce an explanation; the explanation supports human

intervention that validates or overrides the classifier decision; and the combined output builds trust and transparency in the system.

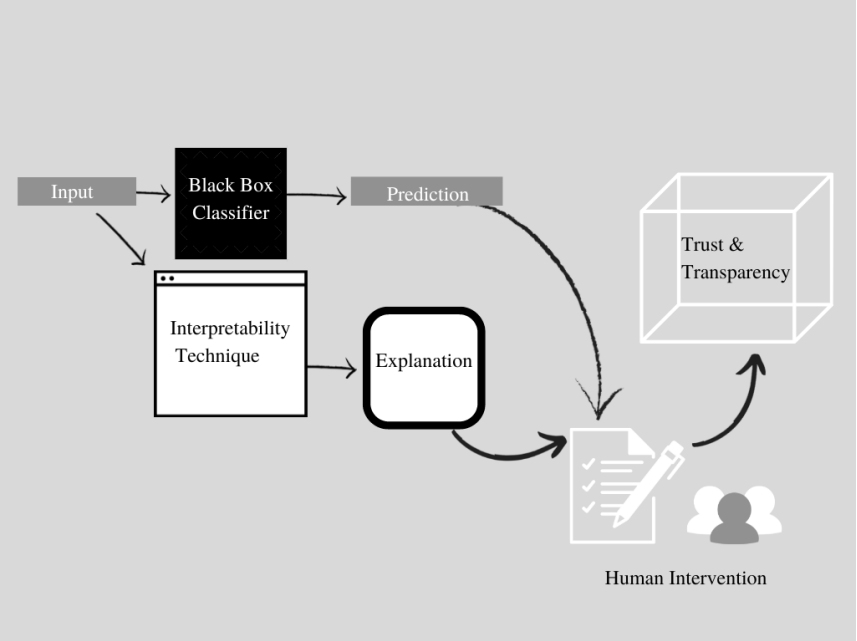


Figure 2.3: General XAI workflow for fake news detection systems. An interpretability technique runs alongside the black-box classifier to generate an explanation; human intervention mediated by the explanation produces trust and transparency in the system output [10].

LIME [104] approximates the local decision boundary of a classifier with an interpretable linear model on perturbed input samples, attributing prediction weight to individual tokens. SHAP [105] computes Shapley values from cooperative game theory to assign each feature a contribution score consistent across the feature set. Grad-CAM [106] projects gradient-weighted class activation maps onto input image regions, revealing which spatial areas drive the classification decision. Attention attribution methods extract the internal attention weight distributions of transformer encoders as a proxy for token-level salience. The thesis contributions Three-Label CT-BERT [11] (LIME over CT-BERT), FakeNews Buster [12] (adaptive explainability combining Fact Check API retrieval with vector similarity), Graph-RAG [14] (interpretable evidence chains), and the multimodal studies [15, 16, 17] (Grad-CAM) together constitute a multi-paradigm explainability suite spanning all detection modalities addressed in this thesis.

2.4 Multimodal Fake News Detection

2.4.1 Taxonomy of Multimodal FND Approaches

Jolly and Kumar [17] provide a structured taxonomy of deep learning-based multimodal FND, organising 60+ published methods along four analytical dimensions. *Input*

modalities range from unimodal text or image to multimodal combinations incorporating video, audio, and social context propagation signals [107]. *Feature extraction* employs RNNs, CNNs, and Transformers (BERT, LongBERT) for text, and CNNs (VGG, ResNet) or Vision Transformers (CLIP ViT-B/32) for visual content. *Data fusion mechanisms* span simple concatenation, attention-based fusion, transformer co-attentional fusion, and dynamic modality weighting; effective fusion must model cross-modal *complementarity* rather than merely aggregating features, since simple concatenation can introduce noise and lose contextual relationships [108, 109]. *Detection paradigms* divide into two families: pattern-based detection (classifying from intrinsic content features using learned semantic inconsistencies and visual artefacts, enabling real-time low-latency deployment) and evidence-based detection (verifying claims against external knowledge bases, document retrieval, or multimodal evidence packages). This taxonomy provides the organising framework for the remainder of this section and directly informs the choice of architectures evaluated in Chapter 4.

2.4.2 Fusion Architectures

Multimodal fake news detection models jointly encode textual and visual (and sometimes metadata) inputs to detect cross-modal inconsistencies not detectable from either modality alone [107, 110]. Three canonical fusion paradigms have been explored. *Early fusion* concatenates raw or low-level feature representations from each modality before any modality-specific processing, allowing the model to learn joint low-level representations. *Late fusion* processes each modality through dedicated encoders and combines the resulting representations at the classification head, typically through concatenation, element-wise multiplication, or attention weighting. *Hybrid fusion* applies modality-specific encoding at multiple levels and combines at each level, enabling both low-level and semantic cross-modal interaction.

The visual encoders used across multimodal FND systems have changed considerably. VGG-19 [111] was among the first deep CNNs widely adopted for visual feature extraction in FND. ResNet [112] introduced residual connections enabling much deeper networks. Vision Transformers (ViT) [113] replaced convolutional operations entirely with patch-based self-attention, while CLIP [114, 115] introduced contrastive language-image pretraining on 400 million image-text pairs, enabling zero-shot visual understanding.

EANN [116] employed event adversarial training to learn representations invariant to event-specific features, improving generalisation across topics. MVAE [117] applied a multimodal variational autoencoder to jointly model the text and image distribution of news articles. SpotFake+ [118] demonstrated cross-modal transfer learning on limited labelled data. SAFE [119, 120] computed semantic alignment scores between text and image representations, exploiting cross-modal inconsistency as a direct veracity signal. dEFEND [121] incorporated user comments alongside news content in a co-attention explainable framework. BDANN [122] applied bi-directional domain adversarial neural networks combining text and visual features.

Further multimodal FND architectures include Jin et al. [123] and Jin et al. [124], which represent some of the earliest systematic multimodal FND evaluations on social media data. Cross-modal consistency and attention mechanisms are explored in Wu

et al. [125], Tan et al. [126], and Bejan [127]. Qian et al. [128] studied bidirectional cross-modal inference for veracity prediction. Wang et al. [129] employed hierarchical cross-modal attention. Guo et al. [108] proposed a two-branch architecture separating semantic and stylistic signals. Hangloo and Arora [130] surveyed methods for combating online misinformation through multimodal approaches. Meel and Vishwakarma [131] provided a comprehensive comparative evaluation of multimodal methods. Prakash et al. [132] examined multimodal FND in social media contexts. Suzuki et al. [133] analysed early fusion effectiveness across modalities. Cao et al. [134] studied the behaviour of multimodal models behind the scene. Agarwal et al. [135] described method and apparatus for collecting and detecting fake news in multimodal contexts.

Emoji-enhanced multimodal representations for social media analysis are studied in Han et al. [136], which incorporates non-textual sentiment signals into FND. Financial domain multimodal misinformation is addressed in Mohankumar et al. [137]. Straive [138] examined the state of multimodal FND solution design.

2.4.3 Cross-Modal Attention Mechanisms

ViLBERT [139] introduced co-attentional transformer layers in which text and visual streams attend to each other at each transformer block, enabling deep iterative cross-modal fusion beyond the concatenation applied in late-fusion architectures. LXMERT [140] extended this with cross-modal self-attention across three transformer encoders for question answering and visual reasoning. VL-BERT [141] and UNITER [142] provided further vision-language pretraining frameworks widely adopted in downstream multimodal tasks. Evaluations of cross-modal representations include Shekhar et al. [143], Caglayan et al. [144], and Parcalabescu et al. [145], which probe the extent to which joint embeddings capture genuine cross-modal relationships. More recent architectures such as MFIR [146] incorporate explicit inconsistency reasoning modules targeting semantic contradictions between text-image pairs, while CroMe [147] employs cross-modal metric learning with Proxy Anchor Loss to tighten intra-modal semantic clusters.

Recent advances in vision-language models have opened new directions for multimodal FND. The 2024–2026 literature clusters into four broad strands, reviewed in turn below: (i) vision-language and LLM-driven detectors; (ii) cross-modal reasoning and consistency methods; (iii) graph- and temporal-propagation models; and (iv) approaches targeting modality imbalance and robustness. Within the first strand, CLIP-based FND approaches are explored in Rocamora et al. [148]. Retrieval-augmented multimodal FND is presented in Li et al. [149]. Zhang et al. [150], Chen et al. [151], and Chen et al. [152] proposed LLM-driven multimodal FND systems. Niu et al. [153] and Yuan et al. [154] developed cross-modal reasoning frameworks for veracity prediction. Li et al. [155] proposed multi-domain adversarial pre-training for robust multimodal FND. Domain-transfer challenges in multimodal FND are studied in Mosallanezhad et al. [156]. Shi et al. [157] explored multimodal misinformation detection under distribution shift. Liu et al. [158] examined multi-modal fact verification. Temporal dynamics in multimodal misinformation are addressed in Qiu et al. [159]. Sun et al.

[160] introduced feature pyramid networks for multimodal FND. Wei et al. [161] proposed dual-channel cross-modal attention. Li et al. [162] developed uncertainty-aware multimodal approaches. Meng et al. [163] proposed dynamic multimodal FND for evolving misinformation. Han et al. [164] explored enhanced multimodal representations. Wang et al. [165] and Liu et al. [166] applied LLMs to dual-stream multimodal FND. Wu and Wang [167] addressed modality imbalance in multimodal FND. Qian et al. [168] proposed multi-modal FND with graph augmentation. Jing et al. [169] introduced dynamic propagation-semantic graphs. Wu et al. [170] and Wu et al. [171] addressed data imbalance in multimodal FND. Kheddache et al. [172] evaluated effectiveness of multimodal approaches. Jadhav et al. [173] assessed AI-driven multimodal FND systems. Qiao et al. [174] proposed cross-modal consistency improvements. Liu et al. [175] studied modality-specific feature learning for FND. Shen et al. [176], Zhu et al. [177], and Yan et al. [109, 178] contributed further multimodal architectures.

Adversarial robustness of multimodal FND systems has been examined in Sen and Dasgupta [179] (FGSM attacks on multimodal systems), Li et al. [180] (robustness under missing modalities), Yang et al. [181] (defence strategies against adversarial multimodal attacks), Zhang et al. [182] (provable robustness bounds), Schlarmann et al. [183] (adversarial vulnerabilities of vision-language models), and Rosebrock [184] (defensive techniques for multimodal classifiers). Recent adversarial multimodal FND analyses include Lu et al. [185], Liao et al. [186], Jolly and Kumar [187], and Li et al. [188]. Interpretable multimodal FND is studied in Chen et al. [189]. Zhong et al. [190] proposed ECA-RNet for enhanced cross-modal attention. Raja et al. [191] explored harnessing cross-modal knowledge for FND. Saliency and connectionist methods for explainability in multimodal FND include Lapuschkin et al. [192] (spectral relevance analysis).

Table 2.2 summarises the principal multimodal fake news detection methods, their fusion strategies, and reported characteristics.

2.5 Adversarial Robustness in Fake News Detection

2.5.1 Adversarial Text Attacks

Adversarial attacks on text classifiers generate minimal perturbations to input text that preserve human-readable semantics while causing misclassification. HotFlip [193] employs gradient-based beam search over the character and word embedding space to identify optimal token substitutions. TextBugger [194] applies five perturbation operations (insert, delete, swap, character substitution, and synonym substitution) ranked by word importance scores. TextFooler replaces important words with semantically similar alternatives from WordNet filtered by a Universal Sentence Encoder similarity threshold. These approaches achieve attack success rates of 50–80% against BERT-based classifiers, demonstrating that even deep contextual encoders rely heavily on surface-form features susceptible to synonym-level manipulation. Fig. 2.4 illustrates representative word-level substitutions generated by these methods: highlighted tokens are replaced by semantically plausible alternatives that flip classifier prediction while remaining fluent to human readers [10].

Table 2.2: Summary of Multimodal Fake News Detection Methods

Method	Architecture	Dataset	F1 / Acc.
EANN [116]	Adversarial event-invariant encoder	Weibo, Twitter	82.2% (F1)
MVAE [117]	Multimodal VAE + KL divergence	MediaEval	74.2% (F1)
SpotFake+ [118]	Transfer learning + cross-modal fusion	Weibo	0.97 (F1)
SAFE [119]	Cross-modal semantic alignment	ReCOVery	0.836 (F1)
dEFEND [121]	Co-attention content + comments	MM-COVID	91% (Acc.)
MFIR [146]	BERT + ResNet-50 + inconsistency module	D1, RECOVERY	0.989 (F1)
CroMe [147]	BLIP-2 + ViT + Proxy Anchor Loss	D1, RECOVERY	0.992 (F1)
ViLBERT [139]	Co-attentional dual-stream transformer	Multiple	0.976 (F1)

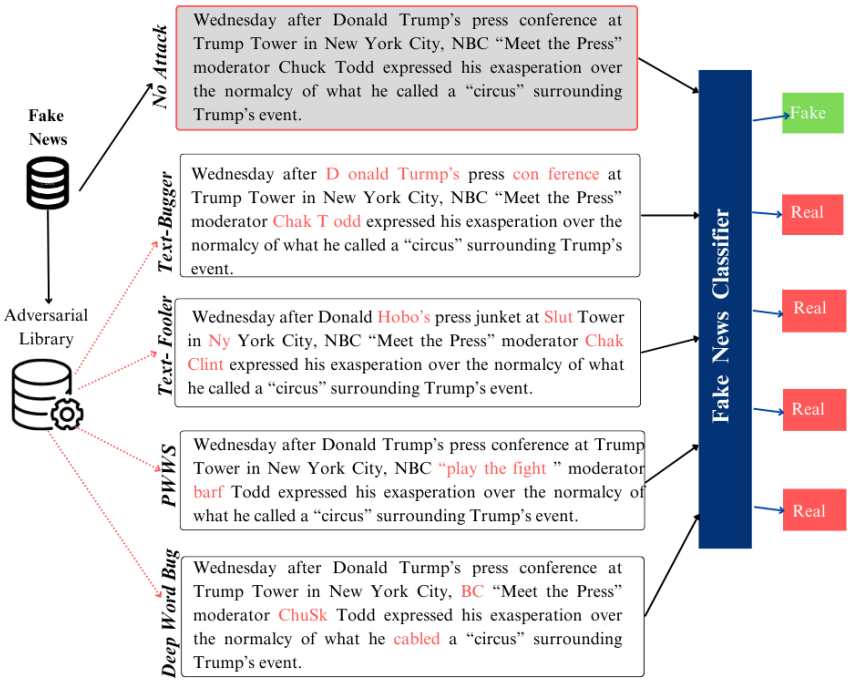


Figure 2.4: Examples of textual perturbations generated by adversarial attack methods from the TextAttack library [195], illustrating how minor word-level substitutions deceive state-of-the-art fake news detectors while preserving surface fluency [10].

A broader taxonomy of NLP adversarial attack methods is surveyed in Morris et al. [195]. The adversarial-FND literature divides into two complementary strands: *attack generation and vulnerability analysis*, which exposes how perturbations defeat detectors, and *defence and explainability*, which seeks to harden or interpret models against such attacks. Within the attack-generation strand, Jin et al. [7] demonstrated that BERT is not robustly resistant to synonym-based natural language attacks. Ali et al. [196] evaluated adversarial vulnerabilities of language models in practical settings. Koenders et al. [197] analysed real-world adversarial vulnerability of deployed FND systems. Semantics-preserving text attack methods are studied in Lee et al. [198]. Bui et al. [199] proposed improvements to adversarial text generation for FND evaluation. Turning to the defence and explainability strand, feature squeezing as a defence against input perturbations is explored in Xu et al. [200]. Adversarial training with curriculum strategies is introduced in Cai et al. [201]. Group-adaptive adversarial training for FND robustness is presented in Ren et al. [202]. Koren et al. [203] evaluated LLM-based adversarial robustness comprehensively. Li et al. [204] studied constructive adversarial examples for FND. Fang et al. [205] analysed BERT-based adversarial generation for fake news. AdvFakeBench [206] provides a systematic adversarial evaluation benchmark for FND. Explainability and attribution methods applied in the adversarial FND context include Lapuschkin et al. [192] (spectral relevance) and Luo et al. [207] (explainable adversarial FND). Al-Sayed et al. [208] provides explainability analysis for adversarially trained FND models. Gupta et al. [209] and Kumar et al. [210] build explainable adversarial FND frameworks. Das et al. [211] and Das et al. [212] propose heuristic-driven adversarial identification.

2.5.2 LLM-Generated Adversarial Content

The emergence of capable LLMs has introduced a qualitatively new attack surface: the generation of semantically coherent inversions of factual claims that are perceptually indistinguishable from authentic journalism [213]. Unlike character-level or synonym-substitution attacks that introduce detectable artefacts, LLM-generated semantic inversions produce fluent, well-structured prose that completely reverses the veracity of the underlying claim. The foundational LLM families used in adversarial FND research include Llama [214], Llama 3 [215], and DeepSeek [216]. Instruction-tuning strategies relevant to adversarial FND include Tang et al. [217]. Park et al. [218] demonstrated that LLM-generated style augmentation constitutes a particularly effective black-box attack. Ren et al. [219] proposed retrieval-augmented adversarial refinement as a defence. LLM-based explainability for adversarial FND is explored in Wu et al. [79]. AI-generated content detection in the era of advanced generative models is addressed in Gehrmann et al. [220], Zellers et al. [221], and OpenAI [222]. The contributions of this thesis [15, 16] (Chapter 4) provide the first systematic multi-architecture evaluation of DeepSeek-LLM and Llama3-generated attacks, confirming via independent statistical tests ($p < 0.05$) that DeepSeek attacks outperform Llama3 equivalents.

2.5.3 Image-Based Adversarial Attacks

For the visual modality, adversarial perturbations computed via gradient-based methods add imperceptible pixel-level noise that redirects model attention from semantically relevant image regions to irrelevant artefacts. FGSM [8] applies a single-step gradient sign perturbation. PGD [223] iterates with random restarts under an L_∞ norm constraint. Momentum variants MIFGSM [224] and NIFGSM [225] improve transferability across architectures via gradient momentum. Input diversity preprocessing is introduced in DIFGSM [226]. Boosting-based transferable adversarial attacks are described in Dong et al. [227]. Fast adversarial training methods are proposed in Wong et al. [228]. Pixle [229] provides black-box image attacks through pixel rearrangement without gradient access. Provably adversarially robust certification for vision models is presented in Salman et al. [230]. Defence strategies including feature denoising are introduced in Xie et al. [231]. Dropout-induced robustness improvements are explored in Dai et al. [232]. Grad-CAM [106] visualisations reveal that under adversarial conditions, model attention shifts from semantically informative image regions to background noise or dataset-specific artefacts such as watermarks.

2.6 Knowledge Distillation and Efficient Models

Knowledge distillation (KD), introduced by Hinton et al. [233], transfers knowledge from a large, accurate *teacher* model to a smaller, more efficient *student* model by training the student to match the teacher's soft output distribution (softened logits) rather than hard labels. The soft targets convey richer information about inter-class relationships than one-hot labels, enabling the student to achieve higher accuracy than training on hard labels alone would provide. DistilBERT [234] applies this paradigm to produce a BERT model 40% smaller and 60% faster while retaining approximately 97% of BERT's performance through a triple distillation loss combining language modelling, soft-label, and cosine embedding objectives.

In the continual learning setting, KD serves a dual purpose: it transfers knowledge from the current task's teacher to the next-task student, and simultaneously prevents catastrophic forgetting by penalising deviations from the teacher's output on past-task data. Standard KD applies a fixed distillation weight throughout training. EAKD-CFND [13] (Chapter 3) computes per-instance distillation weights from teacher uncertainty and, above a threshold, queries an external fact-checking source for a substitute label; its evaluation finds no benefit of either component beyond seed noise. To the best of the author's knowledge, few prior studies pose fake news detection as a class-incremental problem in this sense, the closest related work being concept-drift adaptation in streaming text classification [235]. The exemplar-free versus rehearsal trade-off is established in vision, where exemplar-free methods are competitive under moderate drift and degrade under severe drift [236]; the EAKD-CFND evaluation is consistent with this, as exemplar-free methods fail on PHEME-Event.

Efficient transformer architectures relevant to long-document FND include Informer [237], Autoformer [238], FEDformer [239], BigBird [240], FNet [241], MLP-Mixer [242], and gMLP [243]. These efficient architectures reduce the quadratic

attention complexity of standard transformers, enabling processing of longer news documents within practical computational budgets. AdamW optimisation, widely used in transformer fine-tuning for FND, is described in Loshchilov and Hutter [244].

2.7 Graph-Based and RAG-Based Approaches

2.7.1 Graph Neural Networks for Fake News Detection

Graph neural networks (GNNs) model the structural properties of news propagation and social interaction networks to capture veracity signals unavailable from content alone [245]. GCAN [246] employed graph-aware co-attention combining news content with user retweet propagation graphs. MVAN [247] applied multi-view attention to text content, propagation trees, and source credibility simultaneously. Dynamic GNNs [248] incorporate temporal evolution of the propagation graph, capturing the characteristic spreading dynamics of fake news, including faster initial propagation and larger initial diffusion trees, as time-evolving structural features. On the MediaEval 2020 COVID-19 dataset, GNNs trained only on structural propagation features achieved F1 of 95%, compared to 65.33% for content-based ensembles [10]. Structural and content signals are complementary.

Foundational GNN architectures underlying propagation-based FND include spectral graph convolutions [249], GraphSAGE for inductive representation learning [250], graph convolutional networks (GCN) for semi-supervised classification [251], and Graph Attention Networks (GAT) [252] which assign learned importance weights to edges during message passing. Fake news propagation dynamics are studied in Zhao et al. [253]. Knowledge graph approaches for FND include DEAP-FAKED [254] and discriminative predicate path mining [255]. Attribution and verifiability in summarisation using knowledge graphs is explored in Narayan et al. [256]. Cooperative deep learning for FND in online social networks is proposed in Mallick et al. [257]. Recent graph-based FND approaches include neighbourhood-order GNNs [258], structure-aware propagation learning [259], and propagation-based detection [260]. SEAN [261] combines social context with entity-aware networks. FaceKEG [262] integrates knowledge graphs with social graphs for combined evidence. Infosurgeon [263] applies graph-based cross-media representation for FND. DTN [264] employs dual transformer networks for propagation modelling. Network-based approaches to fake news are surveyed in Zhou and Zafarani [265]. CFFN [266] and RDCM [267] present recent graph-based content and context fusion methods.

2.7.2 Retrieval-Augmented Generation for Fact Verification

Retrieval-Augmented Generation (RAG) [268] augments language model inference with dynamically retrieved context from an external knowledge base, grounding model outputs in current verified information rather than static training distributions. Applied to fake news detection, RAG enables veracity assessment based on retrieved evidence chains rather than learned statistical regularities of surface form. Edge et al. [269] introduced Graph-RAG for knowledge-intensive question answering, showing

that multi-hop graph traversal of the evidence structure improves performance on complex queries requiring contradictory evidence. CRAG [270] extends this with corrective retrieval that self-evaluates retrieved documents and triggers new search when quality is insufficient.

The evidence-grounded literature separates into two groups: classical *evidence-based fact-verification systems* that match claims against curated knowledge, and the newer *RAG-for-FND pipelines* that retrieve and reason over evidence at inference time. Among the former, evidence-based fact verification systems include Li and Zhou [271] (connecting fact verification with FND), Vo and Lee [272] (searching fact-checked information to alleviate false news), Ciampaglia et al. [273] (computational fact-checking through knowledge graphs), FEVEROUS [274] (fact extraction and verification using unstructured and structured evidence), and automated claim verification frameworks [275, 276]. The Google Fact Check Tools API [277] aggregates 28 verified fact-checking sources and is used as an external verification signal in the FND contributions of this thesis [11, 12, 13]. Turning to the RAG-for-FND pipelines, retrieval-augmented generation applied to semi-supervised text classification is explored in Li et al. [278]. DRIVE [279] proposes an adjustable parallel evidence-aware architecture for efficient fact-checking. MIRAGE [280] introduces a retrieval-augmented generation framework specifically designed for FND tasks. Sniffer [281] and VeriAct [282] represent recent evidence-grounded FND systems. RASR [283] and FRAME [284] further advance the retrieval-augmented paradigm for social media fact-checking. CAD [285] and ERF [286] represent system-level approaches combining retrieval with FND pipelines. FIND [287] introduces a consortium-developed retrieval FND system. Yukta et al. [288] present AI-assisted evidence retrieval for FND in 2026.

The Graph-RAG contribution of this thesis [14] (Chapter 3) adapts these principles specifically for fake news detection as a discriminative classification task, introducing the *RAG-for-Classification* paradigm and showing, through a controlled audit, that multi-hop evidence expansion helps on tasks with genuine evidence chains (such as FEVER) while task structure, rather than architecture, dominates performance.

2.8 Benchmark Datasets

Table 2.3 summarises the major benchmark datasets used in fake news detection research, spanning their modality, size, task, domain, and year of release.

The diversity of dataset characteristics across task type, modality, domain, language, and size has important implications for benchmark evaluation. Most published FND results are reported on a single dataset under a specific train/test split, making cross-study comparison unreliable. The Constraint dataset dominates COVID-19 FND evaluation, appearing in 14 of 23 COVID-19-specific studies reviewed in this thesis [10], creating a benchmark monoculture that masks poor cross-dataset generalisation. The contributions of this thesis evaluate across multiple datasets wherever possible, and the multimodal and continual-learning results are reported as mean ± standard deviation over three to five random seeds; the Three-Label CT-BERT and Graph-RAG results are single-run and are treated as indicative.

Table 2.3: Major Fake News Detection Benchmark Datasets

Dataset	Modality	Size	Task / Domain	Year
LIAR [289]	Text	12.8K	6-class veracity / PolitiFact	2017
FakeNewsNet [290]	Text+Social	22K+	Binary / Political+Celebrity	2020
Constraint [3]	Text	10.7K	Binary / COVID-19	2021
FEVER [291]	Text	185K	Supported/Refuted/NEI + evidence	2018
PHEME	Text	–	Rumour verification / Twitter	2016
D1 [292]	Multimodal	4,843	Binary / General news	2022
ReCOVery [119]	Multimodal	1,445	Binary / COVID-19 news	2020
Fakeddit	Multimodal	1M+	6-class / Reddit	2019
NewsCLIPpings	Multimodal	71K	Cross-modal consistency	2021
MediaEval 2020	Propagation	–	Propagation graph / COVID-19	2020
GeoCoV19 [293]	Text	524M	Multilingual / Twitter	2020
MM-COVID	Multimodal	–	Binary / COVID-19	2020

Additional benchmark datasets and evaluations in FND research include the GeoCoV19 multilingual Twitter dataset [293], the PolitiFact-Oslo corpus [294] for cross-lingual FND analysis, and CLAF [295] for contrastive learning on FND benchmarks. FakeDetector [296] and MetaDetector [297] provide model-level fact detection resources. Topic-agnostic detection benchmarks are described in Castelo et al. [298]. The CIMTDETECT dataset [299] focuses on check-worthy claim detection. Domain-adaptive FND datasets including the WOA benchmark are evaluated in Sheikhi [300]. Multifake [301] provides a multimodal multilingual FND dataset. GRASS [302] introduces a graph-based dataset for social-media-aware FND. Zhang et al. [303] and Zhang et al. [304] provide recent FND evaluation benchmarks. Surveys and analyses of fake news datasets are provided in Jadhav et al. [22], Khan et al. [35], and Mridha et al. [42].

2.9 Research Gaps

The following research gaps are identified in Table 2.4 through the structured review above and directly motivate the contributions of this thesis:

Table 2.4: Identified Research Gaps, Novelty of Each Gap, and Status in This Thesis

Gap	Existing Limitation	Status in prior work	Status in this thesis
Binary classification rigidity	Forced real/fake labels misclassify borderline content; no mechanism for uncertainty expression	Partially addressed: multi-class label schemes exist, but not threshold-based abstention enriched with live evidence	Addressed (Chapter 3, Three-Label CT-BERT)
Concept drift and catastrophic forgetting	Static models degrade on evolving misinformation; retraining causes forgetting	Partially addressed: concept-drift adaptation in text streams and incremental FND studies	Addressed (FakeNews Buster, EAKD-CFND)
Retrieval-relevance mismatch	Standard RAG retrieves semantically similar rather than contradictory evidence	Entirely novel framing: no prior audit of RAG used as a classifier	Addressed on evidence-chain tasks (Graph-RAG)
Unimodal bias in multimodal systems	Fusion models collapse when dominant modality is removed or perturbed	Entirely novel: no metric separated fusion gain from single-modality reliance	Addressed (Synergy Gain, Chapter 4)
LLM-generated adversarial content	LLMs produce stylometrically authentic misinformation defeating pattern-based detection	Partially addressed: detection of machine-generated text exists; attacks on multimodal detectors do not	Addressed (attack evaluation and proposed defences)
Lack of systematic COVID-19 review	No PRISMA-protocol survey of COVID FND literature exists covering the full 2011–2026 period	Entirely novel for the 2011–2026 period	Addressed (Chapter 5)
Cross-method comparative evaluation	No unified comparison of the thesis contributions on shared dimensions exists	Entirely novel: the contributions had not been compared on shared dimensions	Addressed (Chapter 5)

2.10 Summary

This chapter has surveyed the principal technical paradigms in fake news detection: classical machine learning, deep learning, transformer-based encoders, multimodal fusion, continual learning, and retrieval-augmented generation. Classical ML methods established the detection framework but remained limited by hand-crafted features and adversarial vulnerability. Deep learning improved representational power, and transformer architectures pushed benchmark performance close to ceiling, a result that exposes benchmark saturation as a limitation in its own right rather than a solved problem. The multimodal extension introduced further open questions around cross-modal fusion quality, bias, and adversarial transferability, while continual learning and RAG-based architectures target the concept-drift and evidence-grounding problems respectively. Table 2.4 distils this survey into seven research gaps that structure the remainder of the thesis. These seven are thesis-level technical gaps spanning the full FND landscape, distinct from the eleven *domain-specific* gaps identified separately in the COVID-19 Infodemic Review [10] (Chapter 5), which specialise this agenda to the pandemic-deployment setting.

Additional efficient transformer architectures include Linformer [305] (linear-complexity self-attention) and hierarchical stochastic attention for uncertainty estimation [306]. FastFormer [307] (additive attention achieving linear complexity) is employed in the FakeNews Buster contribution of Chapter 3. The Papers With Code task page on learning with noisy labels [308] is relevant to the dynamic dataset enrichment in the Three-Label CT-BERT framework [11]. DARE [309] proposes dynamic attentive recurrent encoding for multimodal FND. MDF-FND [310] introduces a dynamic fusion model for multimodal fake news detection. ERIC-FND [311] proposes evidence-reinforced integrated classification. Fake news early detection using theory-driven modelling is studied in Zhou et al. [312]. TIB-based cross-modal detection is explored in Cheema et al. [313]. SAFE [314] computes semantic similarity-based alignment scores for multimodal FND. Open issues in combating fake news with a focus on interpretability are examined in Mohseni et al. [315]. Arabic misinformation detection with data preprocessing is studied in Alokaili et al. [316]. The Fighting an Infodemic COVID-19 fake news dataset [3] provides foundational resources for COVID-19 FND evaluation. Connecting stance and fake news detection with blockchain is explored in Salah et al. [317].

Additional related work not captured in the paradigm-level sections above includes: CLAF [295] (contrastive learning for adversarial FND), domain-transfer FND in social media [156], beyond-news-content approaches leveraging social context [318], the dEFEND framework in its original published form [121, 319], ExFake [320] (explainable FND), Lu et al. [321] (multi-task sifted learning), SÄDI-LIAR multi-modal detection [322], fact detection using knowledge graphs [296, 265], CIMTDETECT [299], topic-agnostic FND [298], MetaDetector [297], modality-interaction approaches [323], and 2026 unified and enhanced FND frameworks [324, 325, 326, 327]. Heterogeneous information network approaches are explored in Wang et al. [328]. LLM-assisted FND systems including RAG-enhanced detection [329] and dialogue-driven detection [330]

complete the methods surveyed. The PRISMA reporting protocol used in the COVID-19 Infodemic Review [10] of this thesis is described in Page et al. [331].

Chapter 3 presents the four text-based contributions addressing the binary classification rigidity, concept drift, and retrieval-evidence mismatch gaps: the Three-Label CT-BERT framework, FakeNews Buster, the Graph-RAG classification framework, and EAKD-CFND.

CHAPTER 3

EFFICIENT AND INTELLIGENT FAKE NEWS DETECTION FRAMEWORK

3.1 Introduction

Text-based fake news detection is the starting point for any misinformation detection system. The overwhelming majority of fake news originates as textual content (social media posts, news articles, and short-form claims) before secondary modalities such as images and video are introduced for amplification. Three interconnected limitations motivate the contributions of this chapter. The forced-choice problem: existing FND systems assign every article to a *real* or *fake* category with no mechanism for uncertainty, producing systematic misclassification of genuinely borderline content. Concept drift: social media misinformation evolves continuously; models trained on static datasets degrade predictably, and naïve fine-tuning causes catastrophic forgetting of previously learned patterns. The retrieval-relevance mismatch: standard retrieval-based detection ranks evidence by semantic similarity, which penalises the contradictory documents most useful for debunking, and pattern-based classifiers that rely solely on model memory fail on novel AI-generated content.

This chapter addresses these limitations through four interconnected contributions, collectively fulfilling RO 2 of this thesis: *to propose an efficient and intelligent Fake News Detection framework.*

Section 3.2 reviews the relevant literature. Section 3.3 presents the three-label CT-BERT framework addressing the forced-choice problem. Section 3.4 presents FakeNews Buster, a continual learning architecture addressing concept drift through adaptive knowledge distillation. Section 3.5 presents the Graph-RAG classification framework addressing the retrieval-relevance mismatch through multi-hop evidence expansion. Section 3.6 presents EAKD-CFND, which advances continual learning to per-instance uncertainty granularity and is evaluated as an exemplar-free default and a rehearsal-augmented configuration. Section 3.7 consolidates all four contributions in a unified comparison.

3.2 Related Work

3.2.1 BERT-Based Text Classification for Fake News

BERT [5] changed fake news detection by enabling models to capture subtle semantic cues, including hedge language, emotional register, and sourcing patterns, that distinguish misinformation from credible reporting. Domain-adaptive variants such as CT-BERT [58], pre-trained on COVID-19 tweets, showed that domain specialisation improves accuracy on domain-specific datasets [62]. RoBERTa [60] and XLNet [61, 69] further advanced the state of the art through improved pretraining strategies. ALBERT [59] reduced computational overhead through parameter sharing and factorized embeddings. Additional COVID-domain BERT applications include Wani et al. [57], Koloski et al. [332], Shifath et al. [66], and Shishah [68]. Advances in BERT-based FND are reviewed in Khan [75] and Fu et al. [71].

3.2.2 Uncertainty and Multi-Label Classification

Binary classification approaches [57, 63] assume every article can be unambiguously assigned real or fake. Gautam et al. [69] partially addressed this through topic distributions but maintained a binary output. Uncertainty quantification for transformer predictions is explored in Arora et al. [77] and Vlachos [78]. The three-label framework of Section 3.3 is the first work in the COVID-19 fake news domain to explicitly model borderline uncertainty as a first-class output.

3.2.3 Continual and Incremental Learning for NLP

Continual learning addresses the stability-plasticity dilemma [6]. Knowledge Distillation [233] has emerged as the dominant exemplar-free forgetting mitigation strategy: a frozen teacher guides the student learning from current data. Hasan et al. [93] applied incremental deep learning to fake news detection but used fixed distillation weights. Concept drift adaptation in text streams is studied in Alrowili and Roy [94]. Class-incremental learning is addressed in Kang et al. [96]. Continual learning for FND is further studied in Rahaman et al. [99] and Ratna [97]. The adaptive scheduling of FakeNews Buster (Section 3.4) and the per-instance uncertainty weighting of EAKD-CFND (Section 3.6) differ from this baseline.

3.2.4 Graph Neural Networks and Retrieval-Augmented Generation

Graph-structured representations model the social ecosystem surrounding news content, capturing propagation patterns invisible to content-only methods [9]. Graph Attention Networks (GATs) [252] extend GNNs by assigning learned attention weights to edges. Foundational GNN architectures include GCN [251] and GraphSAGE [250]. Retrieval-Augmented Generation [268] connects a model to an external knowledge base at inference time, grounding decisions in evidence rather than parametric memory. Edge et al. [269] extended RAG with graph-structured retrieval, demonstrating multi-hop recall improvements. CRAG [270] adds corrective self-evaluation to retrieval. Evidence grounding for FND is explored in Li and Zhou [271] and Vo and

Lee [272]. The Google Fact Check Tools API [277] provides verified claims from 28 fact-checking organisations. The Graph-RAG framework of Section 3.5 applies multi-hop graph retrieval specifically to the fake news classification problem.

3.2.5 Uncertainty as a Training Signal

Uncertainty quantification in neural networks has been studied primarily as a post-hoc diagnostic [333]. Softmax entropy and Maximum Softmax Probability (MSP) are standard lightweight estimators [334]. Adaptive KD approaches [335] introduced category-level uncertainty weighting. Dropout-induced uncertainty estimation is studied in Dai et al. [232]. No prior method located in the literature review applies per-instance uncertainty signals both to weight distillation and to trigger external label verification.

3.2.6 Explainability in Fake News Detection

Explainability is a critical requirement for deployable FND systems [37, 38]. LIME [104] provides local model-agnostic explanations. Shu et al. [121, 319] introduced sentence-level explanations alongside classifications. Grad-CAM [106] provides visual explanation for image-text systems. Spectral relevance analysis [192] exposes Clever-Hans behaviour in neural networks. Explainable FND systems are further studied in Luo et al. [207], Al-Sayed et al. [208], Gupta et al. [209], and Kumar et al. [210]. All four contributions in this chapter integrate explainability mechanisms, ranging from LIME post-hoc analysis (Three-Label CT-BERT) to API-driven evidence retrieval (FakeNews Buster, EAKD-CFND) and graph attention weights (Graph-RAG).

3.3 Three-Label Framework for Uncertainty-Aware Fake News Detection

3.3.1 Motivation: The Gray Area in Fake News

Consider a social media post claiming: *"A recent study suggests air pollution increases COVID-19 mortality."* A standard binary classifier will assign this to either real or fake. However, the underlying claim may be supported by some studies, disputed by others, or represent a mischaracterisation of genuinely uncertain scientific evidence. Assigning a hard label introduces epistemic overconfidence that can mislead users and undermine trust.

The three-label framework addresses this by recognising that the latent representation space of BERT-family models contains a continuous region between definitively fake and definitively real content. Articles whose probability estimates fall within this region should not be forced into either class. Instead, they should be identified as uncertain and subjected to additional verification, improving accuracy and generating verified training data for the next retraining cycle.

3.3.2 Proposed Framework Architecture

The framework combines CT-BERT [58], a domain-specific transformer pre-trained on 160 million COVID-19 Twitter posts, with a three-label output scheme and an external knowledge enrichment pipeline. The complete framework flow is:

1. Input news article is tokenised and encoded by CT-BERT.
2. Softmax probabilities $P(\text{fake})$ and $P(\text{real})$ are extracted from the final classification layer.
3. Threshold evaluation: articles with $0.45 \leq P \leq 0.55$ for both classes are assigned the *Can't Say* label; all others retain binary Real/Fake assignment.
4. For *Can't Say* articles, claim text is extracted and queried against the Google Fact Check Tools API.
5. The API response (aggregating 28 verified fact-checking sources) provides a verified label $y_{verified}$, which is added to the training dataset.
6. The model is retrained on the enriched dataset, iteratively reducing the proportion of uncertain classifications.

Fig. 3.1 depicts this end-to-end three-label methodology, from CT-BERT encoding through threshold-based uncertainty assignment to external knowledge enrichment.

3.3.3 Theoretical Justification of the Probability Threshold

The choice of a 0.45–0.55 probability window for the *Can't Say* category is grounded in the geometry of BERT's embedding space. In high-dimensional representation spaces, the decision boundary between two classes is not a point but a manifold. Articles near this manifold produce softmax outputs that assign similar probability mass to both classes, precisely the $[0.45, 0.55]$ range. The threshold width of 0.10 probability units is empirically motivated: narrower windows leave too many genuinely uncertain articles in the binary classes, while wider windows over-classify ambiguous content as uncertain and reduce system throughput by generating excessive API queries.

3.3.4 Dataset and Exploratory Analysis

The Constraint COVID-19 Fake News Detection Dataset [3] is employed. The experiments use its official 6,420-sample training split (the full corpus contains 10,700 posts, including the 2,140-sample validation and test splits), sourced from PolitiFact, NewsChecker, Boomlive, Google Fact-Check-Explorer, the IFCN chatbot, and Twitter. The dataset contains 3,360 real (52.3%) and 3,060 fake (47.7%) samples, a near-balanced distribution suitable for binary and three-label evaluation. Fig. 3.2 presents the class distribution and the comparison of average text length between real and fake tweets.

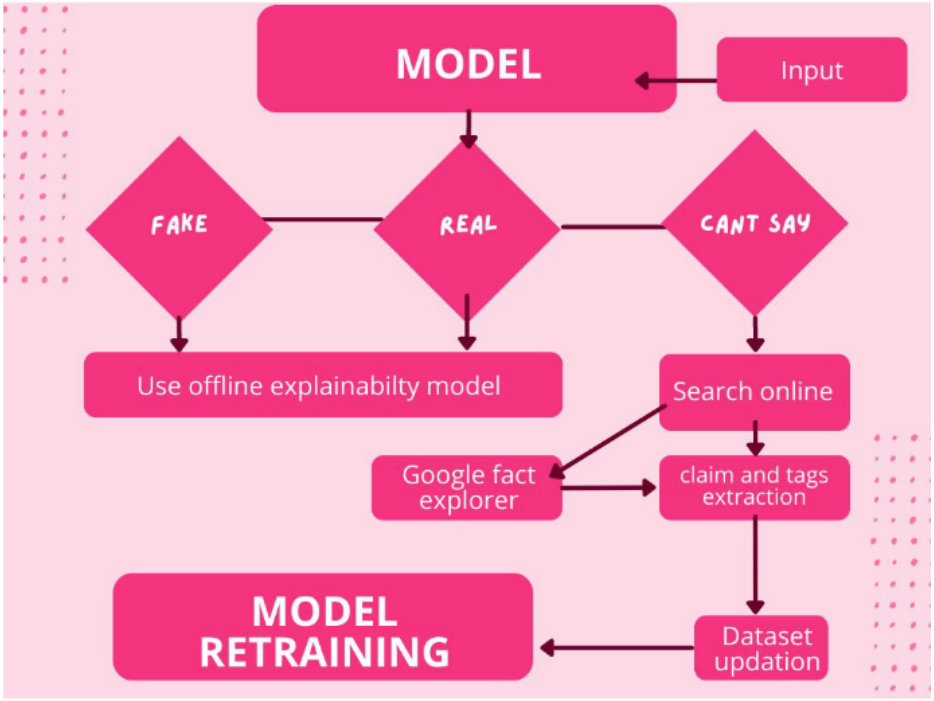


Figure 3.1: Proposed methodology for the three-label uncertainty-aware framework [11].

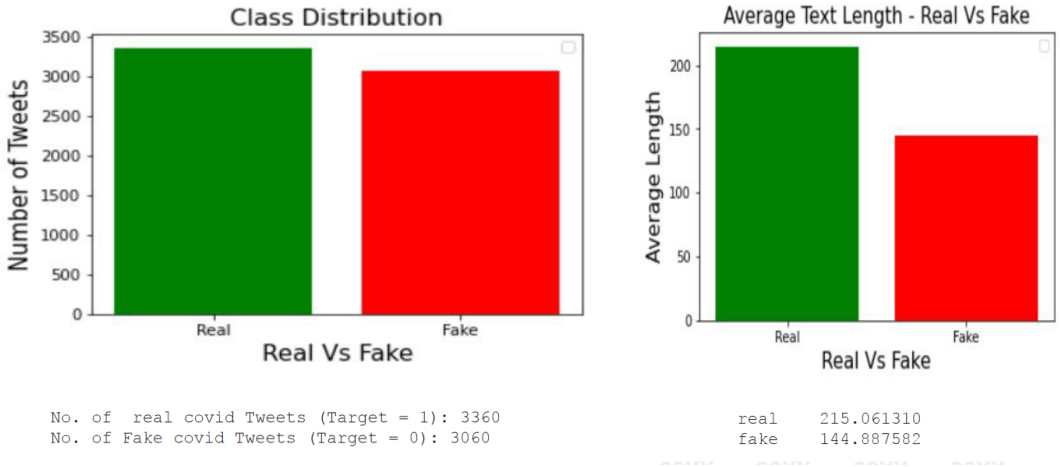


Figure 3.2: Class distribution (left) and average text length comparison between real and fake tweets (right) in the Constraint COVID-19 dataset [11].

Exploratory analysis reveals a statistically meaningful difference in average text length: real tweets average 215 words compared to 144 words for fake tweets. This disparity suggests that authentic COVID-19 reporting tends to be more detailed, while fake

content is more concise, consistent with the hypothesis that misinformation prioritises emotional impact over factual elaboration.

3.3.5 Results and Discussion

Table 3.1 reports the classification performance of the three-label CT-BERT framework against the BERT-base binary baseline.

Table 3.1: Classification Performance: BERT-base vs. CT-BERT Three-Label Framework [11]

Model	**Test Accuracy (%)**	**Framework**
BERT-base (baseline)	80.0	Binary (Real/Fake)
CT-BERT (proposed)	**93.0**	Three-label + data enrichment
Improvement	+13.0 pp	—

CT-BERT achieves 93% accuracy, a 13 percentage-point improvement over BERT-base. This gain is attributable to two compounding factors: domain-adaptive pre-training on COVID-19 tweets and the three-label output scheme that prevents forced misclassification of uncertain content.

The LIME explainability analysis presented in Fig. 3.3 reveals an important finding: even when CT-BERT correctly classifies an article, it may do so for incorrect reasons, assigning low weight to domain-critical keywords such as *air pollutant* and *COVID exposure*. This diagnostic capability identifies specific sample types where the model's feature-capture mechanism is deficient and where targeted dataset enrichment would be most beneficial.

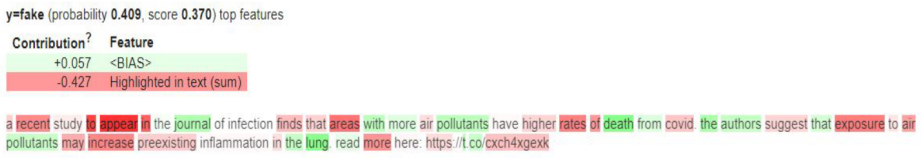


Figure 3.3: LIME explainability output showing feature contributions to a *Can't Say* prediction. Critical domain keywords are underweighted by the model [11].

3.3.6 Limitations

The three-label framework presents several limitations. Only test accuracy is reported; precision, recall, and F1-score analysis are absent, limiting cross-study comparability. Evaluation is restricted to the single COVID-19 domain dataset, leaving generalisation to other fake news domains unvalidated. Google Fact API coverage is limited to the 28 partner fact-checking sites; claims not indexed by these sources remain in the *Can't Say* category indefinitely.

3.4 FakeNews Buster: Continual Learning with BERT and FastFormer

Building on the uncertainty-aware detection established in Section 3.3, this section presents FakeNews Buster, a complete continual learning architecture that operationalises dynamic adaptation at scale [12]. While the three-label framework addresses borderline classification uncertainty at the output layer, FakeNews Buster addresses a deeper challenge: the continuous evolution of the input data distribution itself due to concept drift [94, 93].

3.4.1 Motivation: Concept Drift in Social Media Fake News

Concept drift refers to the statistical phenomenon whereby the relationship between input features and target labels changes over time [94]. In fake news detection, this manifests in multiple dimensions simultaneously: topics shift (from pandemic misinformation to election interference), vocabulary evolves (new slang, emerging technical terms), and adversarial tactics adapt in response to deployed detectors. A static model trained in 2023 will exhibit systematically degrading performance when evaluated on 2025 content, even if the underlying task definition remains identical.

3.4.2 Architecture: Dual Transformer Models

FakeNews Buster evaluates two transformer architectures with complementary strengths:

BERT (`bert-base-uncased`) [5]: A bidirectional encoder pre-trained on BooksCorpus and Wikipedia through masked language modelling and next-sentence prediction. A 128-unit LSTM is appended to the encoder outputs for sequential modelling of tweet-level dynamics, followed by a linear classification head.

FastFormer [307]: An efficient transformer architecture employing additive attention with linear computational complexity. FastFormer aggregates global context through a compressed query representation rather than pairwise attention, enabling effective processing of longer sequences within tighter computational budgets. Configuration: 256 hidden size, 2 layers, 16 attention heads, 256 maximum position embeddings; a 128-unit LSTM and linear head are appended identically to the BERT variant.

3.4.3 Data Processing Schemes

Two processing schemes are evaluated:

Fixlength: Uniform padding of shorter sequences and truncation of longer sequences to a predefined maximum (BERT: 512 tokens; FastFormer: 256 tokens; CL experiments: 256 tokens). Simple and deterministic, but loses information from truncated content.

Recurrence: Long sequences are segmented into overlapping chunks (chunk length: 200 tokens; overlap: 50 tokens). Each chunk is processed independently and chunk representations are aggregated, preserving fine-grained detail from longer inputs at the cost of increased computation. Fig. 3.4 contrasts the two schemes.

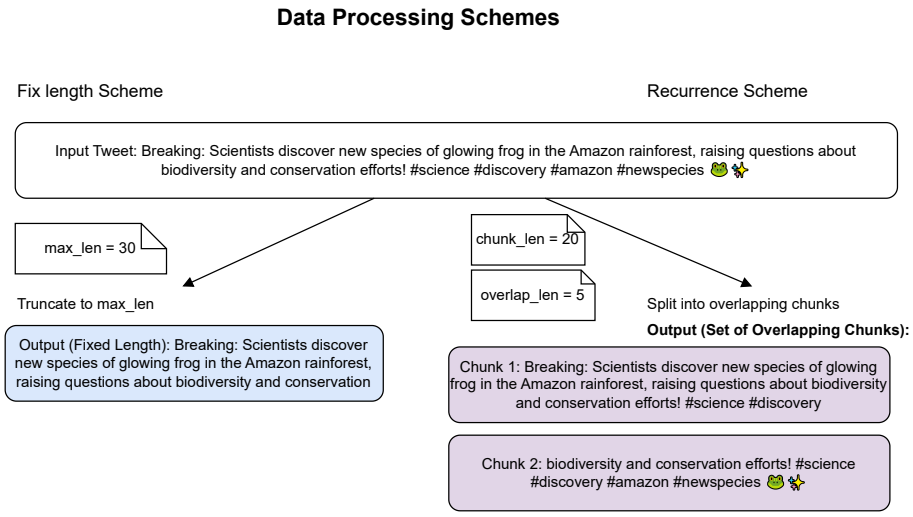


Figure 3.4: Data processing schemes: Fixlength (padding/truncation) and Recurrence (overlapping chunks) for handling variable-length social media text [12].

3.4.4 Continual Learning via Knowledge Distillation

FakeNews Buster employs a teacher-student KD strategy for forgetting mitigation. The model trained on task t (teacher) is frozen and used to guide the model training on task $t + 1$ (student). The KD loss L_{KD} encourages the student's soft output distribution to match the teacher's over previously learned classes.

Data is partitioned chronologically into sequential tasks to simulate real-world concept drift: Task 1 comprises 70% of the Constraint 2021 training set (old data); Task 2 comprises the remaining 30% (new data), with the Task 1 model serving as teacher.

3.4.5 Custom Adaptive Loss Function

The core methodological contribution of FakeNews Buster is a novel adaptive loss function that dynamically schedules the relative weight of the classification and distillation components:

$$L = \frac{1}{\sqrt{\text{seen} + 1}} L_{CE} + \left(1 - \frac{1}{\sqrt{\text{seen} + 1}}\right) L_{KD} \tag{3.1}$$

where `seen` denotes the number of previously encountered training states (batches processed). In the early stages of training on a new task, $\frac{1}{\sqrt{\text{seen}+1}} \approx 1$, placing the full weight on L_{CE}, encouraging the model to learn the current task. As training progresses, the weight shifts progressively toward L_{KD}, increasing the pressure to preserve past-task knowledge. The $\frac{1}{\sqrt{\text{seen}+1}}$ decay rate guarantees monotonic convergence from the plasticity regime to the stability regime without requiring manual tuning of a transition epoch.

3.4.6 Adaptive Explainability Module

The explainability module integrates the Google Fact Check API with vector-based semantic search. For an input tweet, the module queries the API, retrieves matching fact-checked claims and verdicts, computes cosine similarity between tweet embeddings and claim embeddings, and presents the top-ranked fact-checks alongside the model's classification decision. This provides external, verifiable justifications for every prediction and directly connects to the external verification mechanism developed in EAKD-CFND (Section 3.6).

3.4.7 Vocabulary Enrichment

FakeNews Buster extends the BERT vocabulary with domain-specific keywords and 50,000 emoji entries, representing each emoji as its textual description, enabling the model to capture sentiment and emphasis conveyed through emoji-laden posts that would otherwise be tokenised as `[UNK]`. The complete FakeNews Buster architecture, integrating these components, is shown in Fig. 3.5.

3.4.8 Experimental Setup

All models are trained and evaluated on the Constraint 2021 dataset [3]. Training uses the AdamW optimiser [244] with learning rate 3×10^{-5} (linear decay, zero warmup steps), batch size 16, 4 data-loading workers, and 30 epochs per CL task. Training Scheme 2 (attention + regularisation, attention weight $\beta = 0.7$, regularisation margin 0.2) is used for all CL experiments.

3.4.9 Results

3.4.9.1 Single-Task Performance

Table 3.2 reports the single-task F1 scores across model, data-processing, and training-scheme configurations.

BERT with fixlength processing and attention-weighted training (Scheme 1) achieves the highest single-task F1 of 0.9616. FastFormer with recurrence processing and Scheme 2 achieves 0.9544, only 0.72 percentage points below BERT's best, while offering substantially lower computational complexity, making it the recommended architecture for resource-constrained deployment.

3.4.9.2 Continual Learning Performance

Table 3.3 compares the continual learning performance of FakeNews Buster against static, fine-tuning, and fixed-weight KD baselines.

FakeNews Buster achieves an average F1 of 0.92 and a forgetting rate of 0.06, a 60% reduction in forgetting compared to fine-tuning and a 40% reduction compared to fixed-weight KD. Fig. 3.6 plots the per-task F1 trajectory of each method across the sequential tasks.

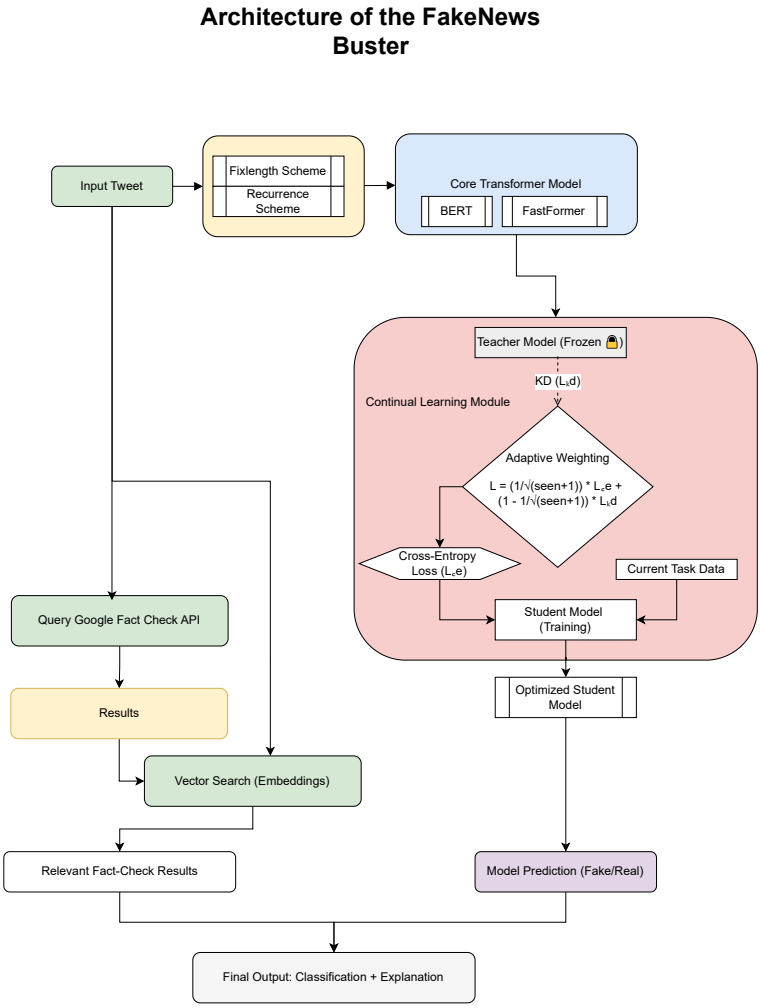


Figure 3.5: FakeNews Buster architecture: input processing, core transformer models, continual learning module, and adaptive explainability [12].

3.4.9.3 Ablation Studies

Table 3.4 isolates the contribution of the adaptive loss weighting, and Table 3.5 that of vocabulary enrichment.

The adaptive loss ablation confirms that both the inclusion of KD and the adaptive scheduling of its weight are necessary for optimal performance. Fixed-weight KD improves over CE-only by 3 F1 points but leaves a 3-point gap below the adaptive variant. Vocabulary enrichment provides a consistent but modest gain (+0.01 F1), confirming its value for social media language while suggesting it is not the primary driver of performance.

Table 3.6 consolidates these results, together with the encoder comparison of Table 3.2 and the static baseline of Table 3.3, into a single component-wise ablation of FakeNews Buster. Removing knowledge distillation entirely causes the largest loss in

Table 3.2: F1 Scores for Different Model Configurations in Single-Task Evaluation [12]

Model	Data Scheme	Training Scheme	F1 Score
BERT	fixlength	0	0.9583
BERT	**fixlength**	**1**	**0.9616**
BERT	fixlength	2	0.9591
BERT	recurrence	0	0.9499
BERT	recurrence	1	0.9601
BERT	recurrence	2	0.9593
FastFormer	fixlength	0	0.9474
FastFormer	fixlength	1	0.9515
FastFormer	fixlength	2	0.9531
FastFormer	recurrence	0	0.9535
FastFormer	recurrence	1	0.9520
FastFormer	**recurrence**	**2**	**0.9544**

Table 3.3: Continual Learning Performance: Average F1 and Forgetting Rate [12]

Method	Avg. F1 Across Tasks	Avg. Forgetting ↓
Static Model	0.70±0.05	—
Fine-Tuning	0.85±0.03	0.15±0.04
Naive KD (Fixed Weight)	0.88±0.02	0.10±0.03
FakeNews Buster (Proposed)	**0.92±0.02**	**0.06±0.02**

Table 3.4: Ablation: Impact of Adaptive Loss Weighting on Continual Learning Performance [12]

Method	Avg. F1	Avg. Forgetting ↓
No KD (CE only)	0.86±0.03	0.14±0.04
Fixed KD weight (0.5)	0.89±0.02	0.09±0.03
Adaptive KD (Proposed)	**0.92±0.02**	**0.06±0.02**

continual-learning F1 (0.06) and more than doubles the forgetting rate (0.06 to 0.14); replacing the adaptive weight with a fixed weight costs half of that gap (0.03 F1, 0.03 forgetting); and vocabulary enrichment contributes the smallest gain (0.01 F1). Not using continual learning at all (the static model) is the most costly choice, at 0.22 F1. In the single-task setting, replacing the BERT encoder with FastFormer costs 0.0072 F1, which is the price paid for linear-complexity attention. Hence the adaptive distillation

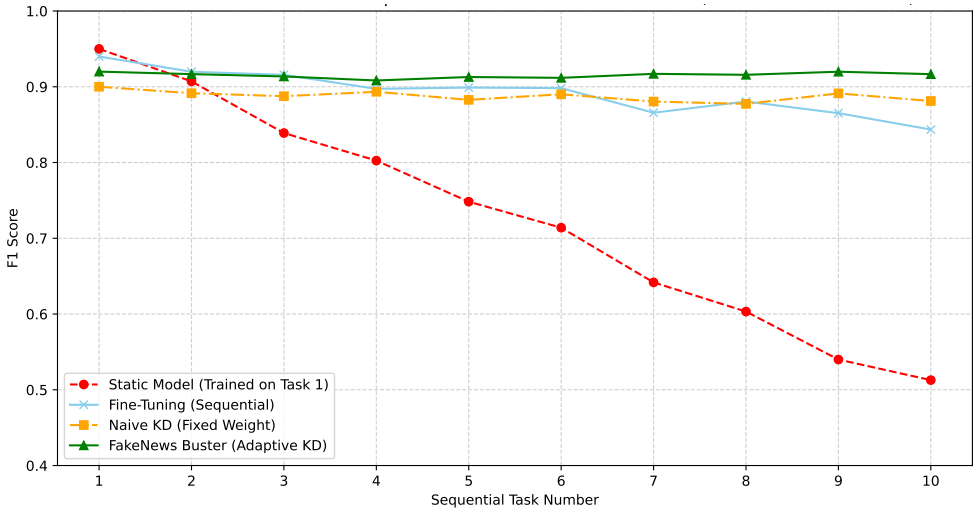


Figure 3.6: F1 score per sequential task for different continual learning methods on the Constraint 2021 dataset [12].

Table 3.5: Ablation: Impact of Vocabulary Enrichment on Continual Learning Performance [12]

Model Variant	**Average F1 Score**
FakeNews Buster (Standard Vocabulary)	0.91±0.02
FakeNews Buster (Enriched Vocabulary)	**0.92±0.02**

schedule is the dominant component, vocabulary enrichment a refinement, and the encoder a cost-accuracy trade-off.

3.4.10 Limitations

FakeNews Buster presents several limitations. The CL evaluation covers only two sequential tasks; real-world deployments would encounter many more tasks with potentially larger distribution shifts. The multimodal dimension of fake news is not addressed. Real-time inference latency is not measured or optimised. The explainability module is evaluated qualitatively only. The vocabulary enrichment gain is modest and scalability to other domains or languages is unvalidated.

3.5 Graph-RAG Classification Framework for Misinformation Mitigation

3.5.1 Motivation: The Retrieval-Relevance Mismatch

A core limitation of standard retrieval-based fake news detection is what this work formalises as the *retrieval-relevance mismatch*. Standard dense-passage retrieval ranks candidate evidence documents by semantic similarity to the input claim. For a claim

Table 3.6: Consolidated ablation of FakeNews Buster components [12]

Configuration	**Component varied**	**Avg. F1**	**Avg. forgetting ↓**
Continual-learning setting (two sequential tasks)			
Full FakeNews Buster	—	**0.92**	**0.06**
Fixed KD weight (0.5)	adaptive weight	0.89	0.09
No KD (CE only)	distillation	0.86	0.14
Standard vocabulary	vocabulary	0.91	—
Static model (no continual learning)	continual module	0.70	—
Single-task setting (best data scheme per encoder)			
BERT, fixed-length scheme	encoder	0.9616	—
FastFormer, recurrence scheme	encoder	0.9544	—

such as *"Politician X voted in favour of Bill Y,"* the most useful evidence for debunking it is a document that explicitly states *"Politician X voted against Bill Y"*, a document that is semantically antithetical rather than similar to the claim. Cosine similarity in embedding space, however, will score such a contradictory document lower than documents that repeat the claim's content.

On the FEVER benchmark [291], which explicitly tests the ability to retrieve *refuting* evidence, multi-hop expansion is expected to help most, since the refuting document often sits a hop away from the claim rather than beside it. The controlled evaluation later in this section bears that out: on FEVER, multi-hop Graph-RAG raises the three-class macro-F1 from 0.724 (single-hop RAG) to 0.755, whereas on tasks without such evidence chains the gain disappears.

3.5.2 The RAG-for-Classification Paradigm

Existing RAG applications predominantly augment large-scale generative models whose inference costs are prohibitive for high-throughput fake news detection pipelines. The RAG-for-Classification paradigm proposed here shows that the core benefits of retrieval-augmentation, including evidence grounding, hallucination resistance, and dynamic knowledge access, can be realised by augmenting lightweight discriminative classifiers (RoBERTa [60], DistilBERT [234]) at a fraction of the computational cost.

The core classifier input format is:

$$\text{Input} = \texttt{[CLS]}\ \text{claim}\ \texttt{[SEP]}\ e_1\ \texttt{[SEP]}\ e_2\ \cdots\ \texttt{[SEP]}\ e_k\ \texttt{[SEP]} \tag{3.2}$$

where $e_1, \ldots, e_k$ are the top-k retrieved evidence passages. This teaches the classifier to perform veracity judgement by reasoning over the claim-evidence relationship rather than claim content alone.

3.5.3 Architecture Design: Three Models Along the Accuracy-Efficiency Frontier

3.5.3.1 Architecture 1: Standard RAG

The Standard RAG model encodes the input claim using `all-MiniLM-L6-v2` [336] and performs exact nearest-neighbour search over the knowledge base using FAISS `IndexFlatL2` [337]:

$$d(q, v) = \sqrt{\sum_{i=1}^{n} (q_i - v_i)^2} \tag{3.3}$$

The top-$k = 3$ semantically similar passages are retrieved and concatenated with the claim for RoBERTa classification. While effective for single-fact claims, this architecture is constrained by the retrieval-relevance mismatch.

3.5.3.2 Architecture 2: Graph-RAG

Graph-RAG addresses the retrieval-relevance mismatch through multi-hop neighbourhood expansion. Given an input claim, primary neighbours N_1 are retrieved identically to Standard RAG. For each document $d_i \in N_1$, a secondary retrieval pass gathers its neighbourhood $N_2(d_i)$. The full evidence set is:

$$E = N_1 \cup \left(\bigcup_{d_i \in N_1} N_2(d_i) \right) \tag{3.4}$$

This expanded evidence set is intended to surface documents that are not directly similar to the claim but are connected through intermediate evidence nodes, including the contradictory documents that single-hop retrieval can miss.

A Graph Attention Network [252] aggregates the heterogeneous evidence graph:

$$h_v^{(l)} = \sigma \left(W^{(l)} \cdot \text{CONCAT} \left(h_v^{(l-1)}, \sum_{u \in \mathcal{N}(v)} \alpha_{vu} \, h_u^{(l-1)} \right) \right) \tag{3.5}$$

where α_{vu} denotes the learned attention weight expressing the semantic importance of neighbour node u to node v. These attention weights serve a dual purpose: evidence aggregation during classification and human-interpretable explainability scores.

3.5.3.3 Architecture 3: Fast Graph-RAG

Fast Graph-RAG retains multi-hop graph reasoning while integrating three production-oriented optimisations: (1) DistilBERT classifier, 40% smaller and 60% faster while retaining ~97% of BERT performance [234]; (2) Approximate FAISS indexing (`IndexIVFFlat`) reducing retrieval latency at a small cost to recall; (3) Neo4j graph database for persistent, scalable graph traversal at production scale.

3.5.4 Knowledge Base Construction and Experimental Setup

Three benchmark datasets are employed: LIAR/LIAR2 [289] (short political statements, six-level truthfulness scale); FakeNewsNet (FNN) [290] (news content with social context); and FEVER [291] (185,445 claims annotated as SUPPORTED, REFUTED, or NOTENOUGHINFO with Wikipedia evidence, the primary benchmark for multi-hop reasoning capability). All datasets use stratified 80/10/10 partitioning. On FEVER the three native classes (SUPPORTED, REFUTED, NOTENOUGHINFO) are retained as the classification targets rather than collapsed to a binary scheme; FEVER F1 is therefore reported as a macro-averaged three-class score, so the NOTENOUGHINFO (retrieval-insufficient) case is evaluated explicitly rather than discarded. Table 3.7 lists the model configuration and hyperparameters for the three RAG architectures.

Table 3.7: Model Configuration and Hyperparameters for RAG Architectures

Parameter	**Standard RAG**	**Graph-RAG**	**Fast G-RAG**
Classifier	RoBERTa	RoBERTa	DistilBERT
Retriever	`all-MiniLM-L6-v2`		
FAISS Index	FlatL2	FlatL2	IVFFlat
nprobe	N/A	N/A	10
Max Seq. Length	256	256	256
k (top-k), m (hops)	k=3	k=3, m=2	k=3, m=2

3.5.5 Results and Discussion

Table 3.8 reports the comparative performance of the three RAG architectures across the LIAR, FNN, and FEVER benchmarks.

Table 3.8: Comparative Performance Metrics Across RAG Architectures and Datasets (single seed, indicative) [14]

Model	**F1 (LIAR2)**	**F1 (FNN)**	**F1 (FEVER)**	**Avg. Time**
Standard RAG	**0.296**	0.989	0.724	9.8 ms
Graph-RAG	0.286	**0.992**	**0.755**	12.8 ms
Fast Graph-RAG	0.281	0.987	0.720	**9.2 ms**

The comparison tells a task-dependent story rather than a clean win for graph structure. On FEVER, where verification runs over explicit multi-hop evidence, Graph-RAG is the strongest of the three and lifts the three-class macro-F1 from 0.724 to 0.755 (+3.1 points), the one setting where the extra hops pay for themselves. On the six-class LIAR2 task all three architectures land near 0.29 macro-F1, and the graph variant is marginally behind the single-pass baseline: fine-grained veracity is bounded by the

labelling, not by how much evidence is retrieved. On the binary FakeNewsNet content subset all three saturate above 0.98, so the architecture makes no difference. The practical reading is that task structure, not architecture, sets the ceiling, and that graph machinery should be added only where the task supplies evidence chains to exploit. Fast Graph-RAG is the deployable choice: its F1 is comparable to the single-pass Standard RAG baseline (within 0.5 points on FNN and FEVER and 1.5 points on LIAR2) while giving the lowest average latency (9.2 ms per claim). Fig. 3.7 shows the three architectures, while Figs. 3.8 and 3.9 visualise the per-dataset F1 spread and the accuracy–latency trade-off.

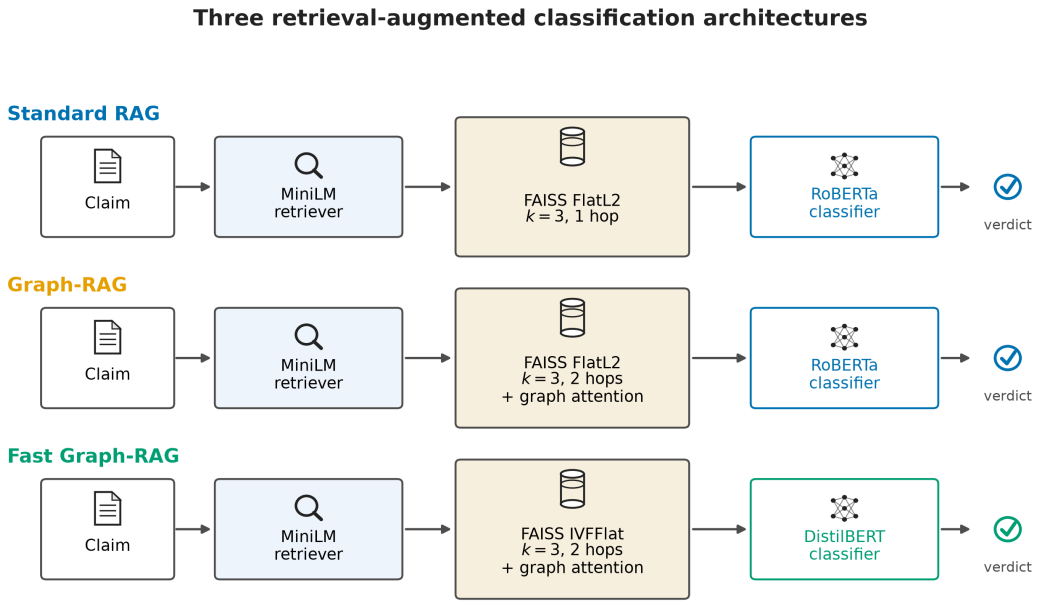


Figure 3.7: The three evaluated RAG architectures. Standard RAG performs single-pass dense retrieval (FAISS FlatL2) into a RoBERTa classifier; Graph-RAG adds a second-hop neighbourhood expansion with graph attention and the same classifier; Fast Graph-RAG swaps in an approximate IVFFlat index and a DistilBERT classifier for lower cost [14].

3.5.6 Limitations

All experiments employ a static knowledge base indexed prior to evaluation; a dynamic knowledge base with live ingestion is a critical direction for production deployment. Knowledge base poisoning is an unaddressed security vulnerability. The framework is text-only. The results reported here are from a single seed on capped subsets and are therefore indicative; a full multi-seed, full-corpus study forms the submitted extended version of this work.

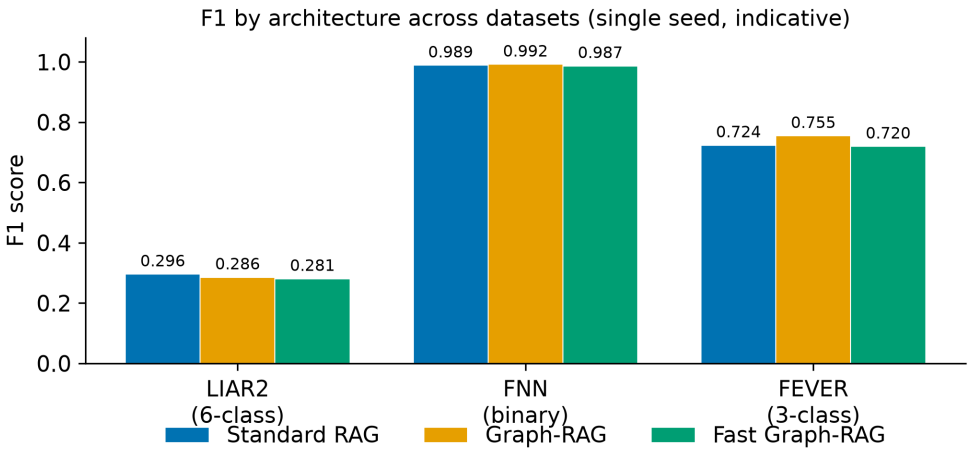


Figure 3.8: F1 by architecture across the three datasets (single seed, indicative). The architectures cluster on each task: near 0.29 on six-class LIAR2, saturated above 0.98 on the binary FNN subset, and separated only on FEVER, where Graph-RAG leads. Task structure, not architecture, sets the level [14].

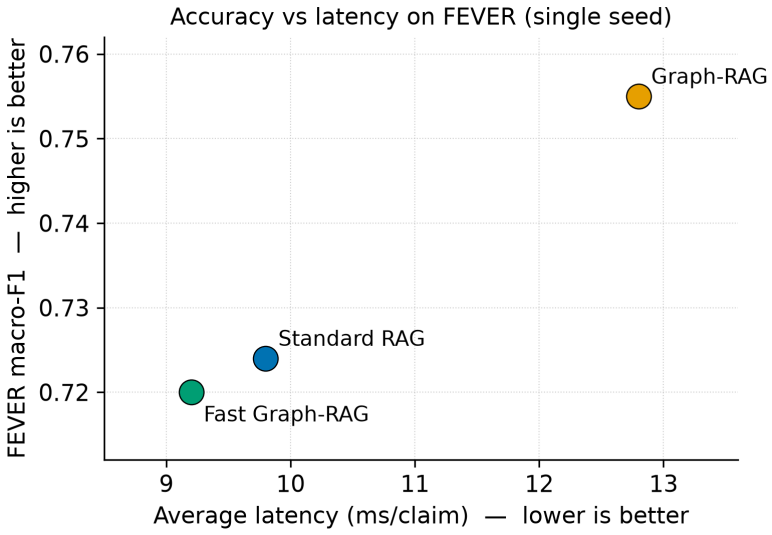


Figure 3.9: Accuracy against average latency on FEVER (single seed). Graph-RAG is the most accurate but the slowest; Fast Graph-RAG sits at the low-latency end while staying within half an F1 point of the single-pass baseline, the practical default where the graph does not pay for itself [14].

3.6 EAKD-CFND: Continual Fake News Detection via Adaptive Knowledge Distillation

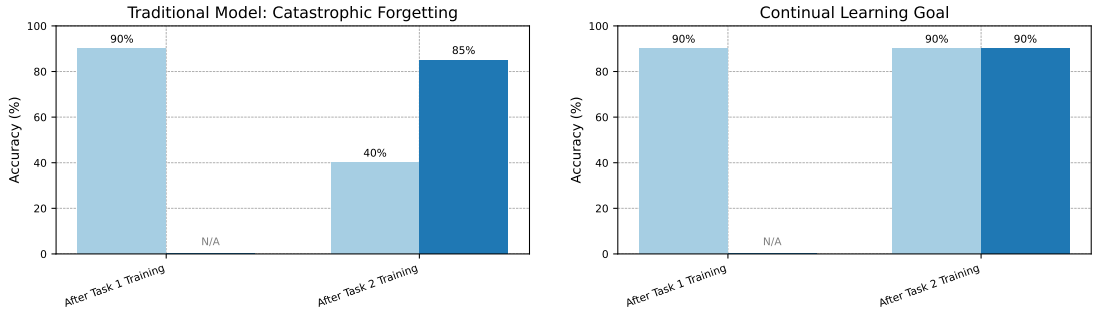
3.6.1 Motivation: Uncertainty-Aware Distillation Without Stored Exemplars

Continual-learning methods for fake news detection (FND) face a trade-off that is usually settled before the data are seen. Rehearsal-based methods store real past

examples and replay them; they are accurate but incur memory and data-retention costs that many streaming deployments cannot accept. Exemplar-free methods avoid that cost but may degrade badly under severe task drift. Standard Knowledge Distillation (KD) belongs to the second family and imposes a further limitation: the student is anchored to the teacher with equal strength for confident and uncertain teacher predictions alike, although the teacher is least reliable at category boundaries, under novel misinformation narratives, and on domains absent from earlier tasks.

EAKD-CFND (Evidence-Aware Knowledge-Distilled Continual Fake News Detection) is built around one mechanism: *per-instance, uncertainty-adaptive knowledge distillation with external fact-checking verification*. The mechanism is offered in two configurations. The exemplar-free default (EAKD-CFND) never stores a past instance. The rehearsal-augmented configuration (EAKD-CFND-R) adds a small reservoir buffer for deployments where storage is available and drift risk is high. The two are configurations of one framework rather than competing methods, and the chapter evaluates both against five baselines under one identical protocol [13].

Fig. 3.10 illustrates the underlying problem. A model trained to approximately 90% accuracy on political fake news (Task 1) loses most of that accuracy on Task 1 after being updated on health misinformation (Task 2), whereas the continual-learning objective is to retain high accuracy on both tasks. The figure is a conceptual illustration of the phenomenon; the accuracy values shown are not measurements from a specific run.



(a) Traditional FND: Performance on Task 1 (e.g., political fake news) degrades significantly after training on Task 2 (e.g., health misinformation).

(b) Continual Learning: The goal is to learn Task 2 without forgetting Task 1, maintaining high performance on both.

Figure 3.10: Catastrophic forgetting in traditional FND models versus the continual learning objective (conceptual illustration; values are illustrative and not measured) [13].

A review of the literature located no established prior system that poses fake news detection as a class-incremental learning problem with disjoint per-task label sets, a frozen-teacher/trainable-student protocol, and sequential task boundaries; the closest related work concerns concept drift in streaming text classification generally [235]. The closest methodological relative is LUD [335], which weights distillation by *category-level* uncertainty, whereas EAKD-CFND uses an *instance-level* signal and, in addition, acts on high uncertainty by consulting an external source rather than only discounting the teacher.

3.6.2 Problem Formulation

The continual fake news detection problem is formalised as a Class-Incremental Learning (CIL) setting. A sequence of tasks $\mathcal{T} = \{T_1, T_2, \ldots, T_N\}$ is received, where each task T_i introduces a dataset $\mathcal{D}_i = \{(x_{ij}, y_{ij})\}$ with a disjoint set of classes. Tasks are ordered chronologically, so the sequence also reflects real temporal drift, but the learning protocol itself is class-incremental. During training on T_i, only $\mathcal{D}_i$ is available in the exemplar-free configuration. The objective is a single model f^S that achieves high performance on T_i while retaining performance on all previously learned tasks.

3.6.3 Framework Architecture

EAKD-CFND employs a teacher-student architecture with six interacting components, shown in Fig. 3.11: a shared Transformer backbone, a knowledge distillation module, an uncertainty signal generation module, an adaptive weighting module, an uncertainty-driven external verification module, and a combined loss function. Let f^T denote the teacher model trained through task T_{i-1} and held frozen. Let f^S denote the student model, initialised from f^T and trained on the current task T_i. Both models share a `BERT-base-uncased` backbone [5].

An input article x from $\mathcal{D}_i$ is passed simultaneously to the frozen teacher and to the student. The teacher logits $z^T(x)$ feed the uncertainty signal generation module, which outputs an instance-level score $\omega(x)$; this score determines the distillation weight $\alpha(\omega(x))$. In parallel, the student logits $z^S(x)$ feed the task loss and the weighted distillation loss. When $\omega(x)$ exceeds the threshold $\theta_{uncertainty}$, the external verification module may substitute a fact-checked label for the original training label. Fig. 3.12 expands the internal logic of the uncertainty and weighting modules.

3.6.4 Knowledge Distillation Module

The distillation loss aligns the student's predictions with the teacher's over previously learned classes using Kullback-Leibler divergence between temperature-softened distributions:

$$L_{KD}(x) = \tau^2 \cdot \mathrm{KL}\left(\sigma\left(\frac{z^T(x)}{\tau}\right) \middle\| \sigma\left(\frac{z^S(x)}{\tau}\right)\right) \tag{3.6}$$

where $z^T(x)$ and $z^S(x)$ are the teacher and student logits, σ is the softmax function, and τ is the temperature parameter. Higher temperature produces softer probability distributions that convey more information about inter-class relationships [233].

3.6.5 Uncertainty Signal Generation and Adaptive Weighting

A per-instance uncertainty signal $\omega(x)$ is extracted from the teacher's output distribution $P(x) = \{p_1, p_2, \ldots, p_C\}$ over C classes. Two measures are employed:

- *Prediction Entropy:* $H(P(x)) = -\sum_{i=1}^{C} p_i \log p_i$. Higher entropy indicates greater uncertainty.

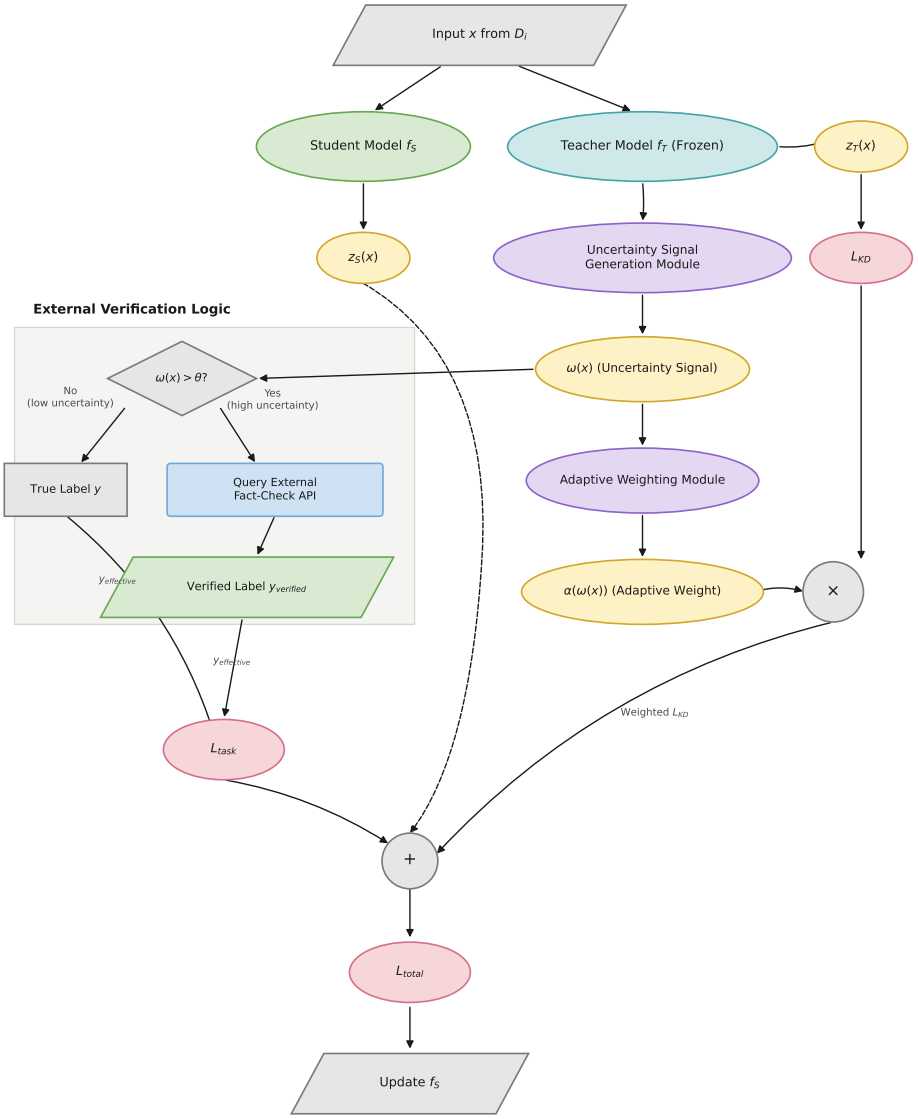


Figure 3.11: EAKD-CFND training flow diagram showing the six interacting components [13].

- *Maximum Softmax Probability (MSP):* A low maximum probability $\max_i p_i$ indicates that the teacher has no strong preference for any class.

The selected measure is normalised to $[0, 1]$ and mapped to an adaptive distillation weight:

$$\alpha(\omega(x)) = \max\left(0,\ 1 - \beta \cdot \text{normalize}(\omega(x))\right) \tag{3.7}$$

where β controls the sensitivity of the weight to uncertainty. When the teacher is confident ($\omega(x) \to 0$), $\alpha(\omega(x)) \to 1$ and the student is strongly anchored to the teacher's output; when the teacher is uncertain ($\omega(x) \to 1$), $\alpha(\omega(x)) \to 0$ (for $\beta = 1$; for $\beta > 1$, α reaches zero at lower uncertainty) and the distillation pressure is removed.

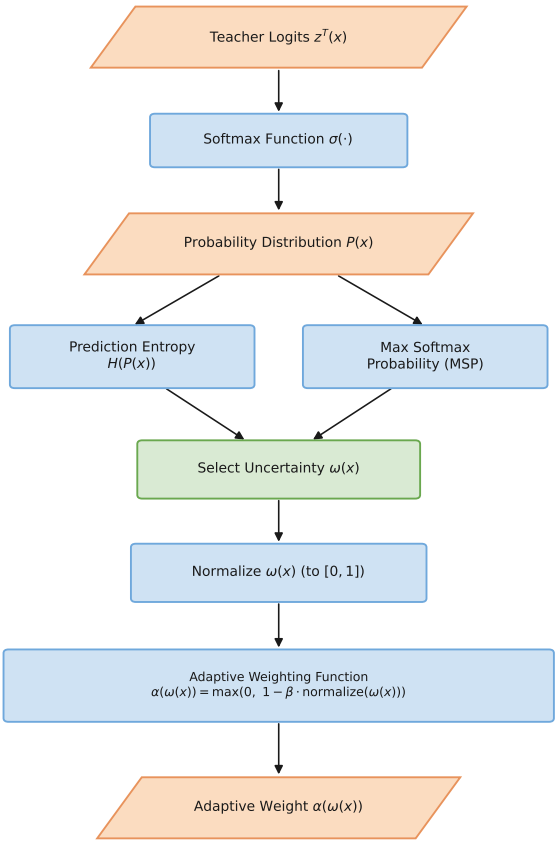


Figure 3.12: Detailed flow of uncertainty signal generation and adaptive weighting [13].

Fig. 3.13 visualises this function for several values of β. This is analogous to adaptive per-instance loss weighting in prior KD work [102].

3.6.6 Uncertainty-Driven External Verification

When $\omega(x) > \theta_{uncertainty}$, a query is issued to an external fact-checking service (the Google Fact Check Tools API). Algorithm 1 specifies how a query is constructed from the input, how results are filtered by a similarity threshold τ_{sim}, how textual verdicts are mapped onto the task label space, and how multiple or conflicting results are resolved by strict majority.

The module is best-effort by construction. An API timeout, an empty result set, a result that fails the similarity filter, or a tied conflict all resolve to the same fallback: the original dataset label is retained. External verification can therefore only replace a label with a filtered, mapped, majority-supported one and never injects an unvetted

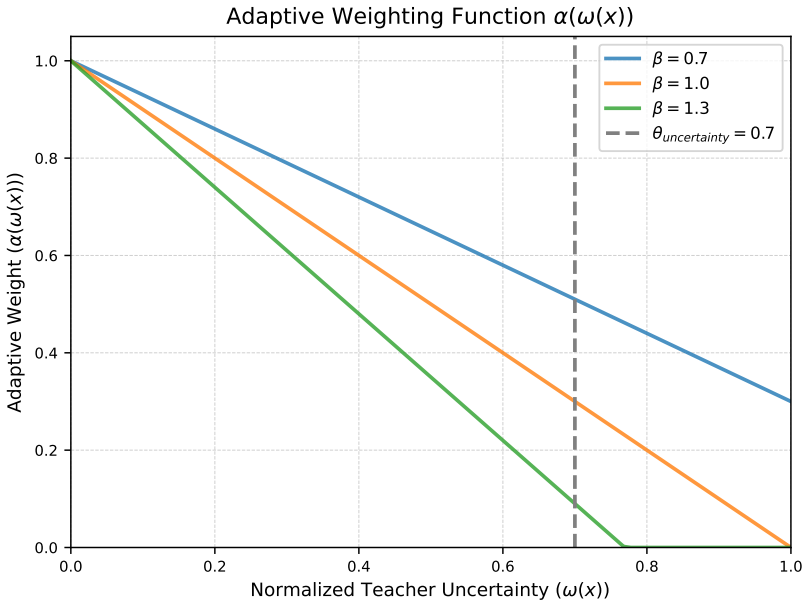


Figure 3.13: Adaptive weighting function $\alpha(\omega(x))$ for several β values; the uncertainty threshold $\theta_{uncertainty}$ beyond which external verification is triggered is marked [13].

Algorithm 1 Uncertainty-Driven External Verification

Require: Input x, uncertainty $\omega(x)$, threshold $\theta_{uncertainty}$, original label y, similarity threshold τ_{sim}
Ensure: Effective training label $y_{effective}$

if $\omega(x) \leq \theta_{uncertainty}$ **then**
    return y
end if
$q \leftarrow \textsc{BuildQuery}(x)$; $R \leftarrow \textsc{FactCheckAPI.search}(q)$
if API call failed **or** $R = \emptyset$ **then**
    return y
end if
$R' \leftarrow \{r \in R : \textsc{sim}(r.\text{claim}, x) \geq \tau_{sim}\}$
if $R' = \emptyset$ **then**
    return y
end if
$L \leftarrow \{\textsc{MapVerdictToLabel}(r.\text{verdict}) : r \in R'\}$
if $|L| = 1$ **then**
    return the single label in L
else if one label in L has a strict majority **then**
    return the majority label
else
    return y
end if

label. Verification is invoked only on instances of the current task's *training* split and never at evaluation time, so no fact-check result informs a test label. Because tasks are processed in chronological order, each query is also issued at the point in time corresponding to its task, which avoids retrospective fact-checks leaking into earlier tasks. The intent of the mechanism is that the student need not imitate a hedging teacher and can, through the external channel, accumulate knowledge the teacher lacked (the "student surpasses teacher" dynamic). Section 3.6.10.2 reports the measured outcome of testing this intent, which shows no measurable accuracy benefit from verification.

3.6.7 Combined Loss Function

For each training batch $\mathcal{B}$, the student minimises:

$$L_{total} = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \Big(L_{task}(f^S(x),\ y_{effective}) + \alpha(\omega(x)) \cdot L_{KD}(x) \Big) \tag{3.8}$$

where L_{task} is the cross-entropy loss and $y_{effective}$ equals the original label y unless external verification substituted $y_{verified}$.

3.6.8 EAKD-CFND-R: A Rehearsal-Augmented Configuration

Eqn. (3.8) never references a stored past instance, which is the appropriate default where retaining raw historical data carries a storage or data-retention cost. For deployments with a storage budget and severe drift, EAKD-CFND-R adds a small reservoir buffer of past $(x, z^S(x))$ pairs (same capacity and sampling scheme as DER's buffer of 200) replayed through a mean-squared-error loss against the stored logits, on top of the uncertainty-adaptive KD and verification terms:

$$L^R_{total} = L_{total} + \gamma \cdot \mathbb{E}_{(x',z') \sim \text{Buffer}} \left[\| z^S(x') - z' \|_2^2 \right] \tag{3.9}$$

with $\gamma = 0.5$ (matching DER's replay weight). EAKD-CFND-R trades away the exemplar-free property for the accuracy that rehearsal provides under severe drift. Three extensions to the replay term are also evaluated: +PP, a second buffer batch replayed by cross-entropy against stored ground-truth labels (a DER++-style term); +UW, replacement of the flat γ by a per-instance weight from $\omega(x)$ applied to the replayed instances; and no-verification, the plain variant with the external API disabled.

3.6.9 Experimental Setup

Two public benchmarks are adapted for the CIL scenario, giving three class-incremental task sequences (Table 3.9): PHEME-Event (rumourous claims from Twitter, ordered by event chronology), FNN-Poli-Time (PolitiFact articles from FakeNewsNet, binned by publication year), and FNN-Gossip-Time (GossipCop articles from FakeNewsNet, binned by publication year). Because publication-year metadata do not cover every article, tasks are binned to years with adequate coverage; the uneven task sizes of FNN-Gossip-Time reflect the corpus's own collection-date skew.

Table 3.9: EAKD-CFND dataset and task configuration (train/test examples per task) [13].

Dataset	Source / split	Avg. length
PHEME-Event	Twitter; event chronology	~30 words
FNN-Poli-Time	PolitiFact; publication year	~500 words
FNN-Gossip-Time	GossipCop; publication year	~300 words

Dataset	T0	T1	T2	T3	T4
PHEME-Event	1664/415	915/228	376/93	712/178	977/244
FNN-Poli-Time	126/31	263/65	114/28	—	—
FNN-Gossip-Time	789/197	8586/2146	7172/1792	—	—

Five baselines are compared under one identical protocol (same backbone, optimiser, learning rate, batch size, epoch budget, and task sequence): Fine-tuning (the lower bound), EWC [91], LwF [338], DER with buffer 200 [92], and LUD [335]. Both EAKD-CFND configurations are evaluated. Metrics are Average Accuracy, macro-averaged Average F1, and Backward Transfer (BWT; closer to zero indicates less forgetting). Forward Transfer is not reported, because the training loop did not record the per-task random-initialisation baseline it requires.

Implementation: `BERT-base-uncased`, Adam optimiser [339], learning rate 2×10^{-5}, batch size 32, 10 epochs per task, $\theta_{uncertainty} = 0.7$, $\beta = 1.0$. All experiments use five seeds (13, 42, 123, 2024, 31415); results are mean ± standard deviation.

3.6.10 Results and Discussion

3.6.10.1 Quantitative Performance

Table 3.10 presents the full comparison; every cell comes from a complete five-seed run under the identical protocol.

Fig. 3.14 plots Average Accuracy and BWT for all seven methods.

On PHEME-Event, all five exemplar-free methods (Fine-tuning, EWC, LwF, LUD, and EAKD-CFND) collapse to near-random accuracy (16.1–16.8%) with BWT of −85% to −87%. DER is the exception (84.8%, BWT −3.2%). The distinguishing factor is not distillation quality, weighting granularity, or external verification: it is that DER alone stores and replays real past exemplars. On both FakeNewsNet splits no method separates itself: all cluster within a few accuracy points (27–31% on FNN-Poli-Time, 28.9–29.4% on FNN-Gossip-Time).

Statistical significance. Paired, seed-matched t-tests with Holm–Bonferroni correction across all 15 comparisons (EAKD-CFND versus each baseline on each dataset; full table in Appendix III.3) leave exactly one significant result: DER outperforms EAKD-CFND on PHEME-Event by 68 accuracy points ($d_z = -55.55$, $p < 0.001$). Every other comparison is a statistical tie after correction, including the FNN-Gossip-Time

Table 3.10: Continual learning performance (5 seeds, identical protocol) [13]. BWT closer to 0 indicates less forgetting. Bold marks the best accuracy and F1 cells (and the best BWT among them) as in the source paper; DER and EAKD-CFND-R are statistically tied on PHEME-Event.

Method	**Dataset**	**Avg. Acc. (%)**	**Avg. F1 (%)**	**BWT (%)**
Fine-tuning	PHEME-Event	16.1±1.2	3.2±0.2	−85.3±2.0
	FNN-Poli-Time	27.4±2.8	7.1±2.5	−80.5±3.4
	FNN-Gossip-Time	29.0±0.2	8.9±0.1	−84.6±2.0
EWC	PHEME-Event	16.5±0.5	3.3±0.1	−86.3±1.6
	FNN-Poli-Time	28.6±3.4	7.9±2.9	−82.1±2.7
	FNN-Gossip-Time	28.9±0.3	8.9±0.1	−84.8±1.6
LwF	PHEME-Event	16.5±0.3	3.3±0.1	−87.1±1.0
	FNN-Poli-Time	31.2±1.0	10.1±0.5	−81.4±2.4
	FNN-Gossip-Time	28.9±0.5	8.9±0.2	−84.3±1.5
DER (buf.=200)	PHEME-Event	**84.8±1.5***	**16.5±0.2**	**−3.2±2.1***
	FNN-Poli-Time	30.4±0.9	9.9±0.3	−80.1±4.0
	FNN-Gossip-Time	**29.4±0.3**	**9.1±0.1**	−85.4±0.5
LUD	PHEME-Event	16.7±0.3	3.3±0.1	−87.1±0.9
	FNN-Poli-Time	31.2±0.5	10.1±0.3	−82.2±2.3
	FNN-Gossip-Time	29.0±0.3	8.8±0.2	−84.8±1.5
EAKD-CFND	PHEME-Event	16.8±0.4	3.3±0.1	−87.2±1.1
(exemplar-free)	FNN-Poli-Time	29.8±2.8	9.0±2.4	−81.6±2.0
	FNN-Gossip-Time	29.1±0.4	8.9±0.1	−84.6±1.7
EAKD-CFND-R	PHEME-Event	**85.5±1.1***	**16.7±0.2**	**−1.8±1.0***
(rehearsal)	FNN-Poli-Time	**34.9±3.2**	**11.7±1.2**	**−72.0±8.0**
	FNN-Gossip-Time	29.2±0.3	9.0±0.1	−85.6±0.7

*DER and EAKD-CFND-R are statistically indistinguishable on PHEME-Event (see text). Per-dataset bold cells follow the source paper: DER on PHEME-Event and FNN-Gossip-Time accuracy/F1, EAKD-CFND-R on PHEME-Event and FNN-Poli-Time.

comparison with DER ($p = 0.011$ raw, 0.147 corrected). EAKD-CFND is therefore indistinguishable not only from DER and LUD on both FakeNewsNet splits but also from fine-tuning, EWC, and LwF. The only significant difference favours DER.

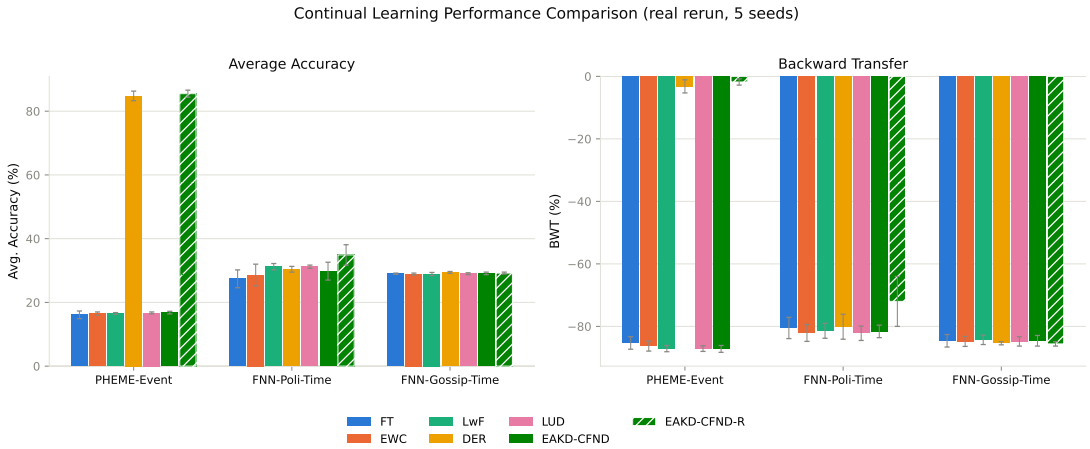


Figure 3.14: Quantitative performance comparison across all seven methods and three datasets; EAKD-CFND-R (hatched) is the rehearsal-augmented configuration of the same framework [13].

3.6.10.2 Ablation Studies

Table 3.11 isolates the contribution of each component on FNN-Poli-Time (5 seeds), and Fig. 3.15 plots the same data with EAKD-CFND-R as a reference bar.

Table 3.11: Ablation of EAKD-CFND components on FNN-Poli-Time [13]

Configuration	**Avg. Accuracy (%)**	**BWT (%)**
EAKD-CFND (Full)	29.8±2.8	−81.6±2.0
Standard KD (λ=1, fixed)	31.2±1.0	−81.4±2.4
EAKD w/ Random $\alpha(x)$	29.0±2.5	−81.1±1.2
EAKD w/o Uncertainty (loss-based α)	28.8±2.3	−80.8±2.5
EAKD w/o External Verification	29.5±2.6	−81.6±2.0
Fine-tuning (no KD)	27.4±2.8	−80.5±3.4

All six configurations lie within about 4 accuracy points (27.4–31.2%) and 1.1 BWT points, inside the per-row standard deviations. Fixed-weight Standard KD is numerically the best row, so the instance-level weighting does not show a measurable accuracy advantage over uniform weighting on this dataset, and removing external verification costs essentially nothing (29.5% versus 29.8%). Paired seed-matched tests of each row against the full model, with Holm correction across the five comparisons, find no significant difference (smallest raw $p = 0.07$ for Random $\alpha(x)$, $d_z = 1.10$; corrected $p = 0.35$). The null result is therefore supported by paired tests and not only by overlapping standard deviations; the Random-α comparison is the one most worth revisiting with more seeds.

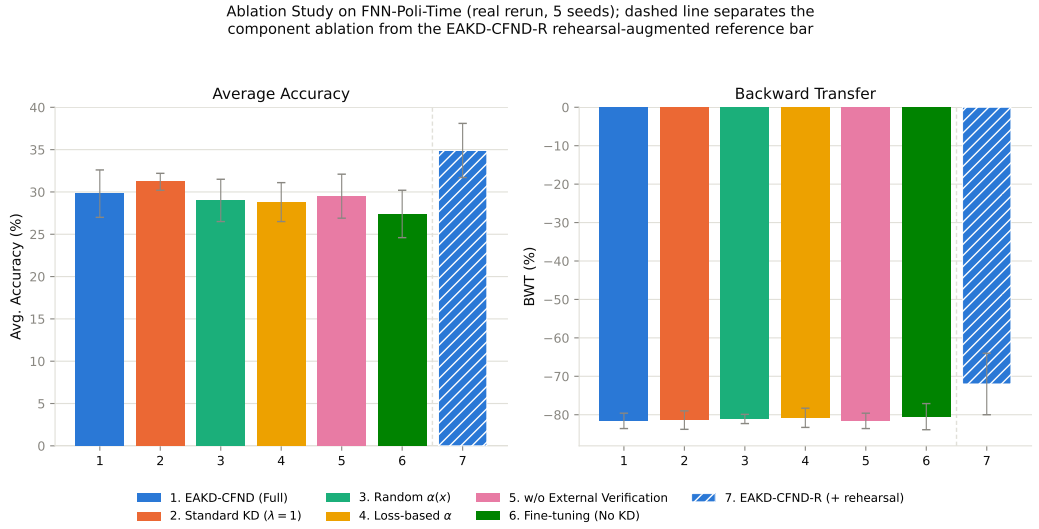


Figure 3.15: Ablation results on FNN-Poli-Time; the hatched EAKD-CFND-R bar is not part of the ablation sweep and is included for scale [13].

3.6.10.3 EAKD-CFND-R Evaluation

Table 3.12 reports the rehearsal-augmented variants on all three benchmarks.

On PHEME-Event every variant closes the gap to DER (84.3–85.5%, statistically indistinguishable from DER's 84.8% and from each other), confirming that rehearsal, and not any property specific to the distillation mechanism, is what the five-task sequence requires. Within the design space tested, any rehearsal mechanism suffices, and disabling external verification on top of plain replay changes essentially nothing (85.4% versus 85.5%), reproduced exactly on a second run with identical seeds (a determinism check, not an independent replication). On FNN-Poli-Time every variant beats both EAKD-CFND (29.8%) and DER (30.4%); +PP (37.2%) and +PP+UW (37.8%) are strongest (paired $p < 0.05$ against both). On FNN-Gossip-Time no variant separates from EAKD-CFND or DER.

3.6.10.4 Why the Benchmarks Diverge

The divergence among the three splits was tested directly. TF-IDF task-centroid cosine similarity (Table 3.13) shows that FNN-Gossip-Time's tasks are topical near-duplicates (mean pairwise similarity 0.816, driven by recurring tabloid entities and headline templates), PHEME-Event's tasks share only Twitter artefacts (0.199), and FNN-Poli-Time lies between (0.406).

To exclude the alternative explanation that Gossip-Time is simply a low-ceiling problem, a non-continual fine-tuned model was trained on each task in isolation. Its three tasks reach 83.8%, 86.9%, and 86.8% (mean 85.8%) against ~29% in the full sequence, so the tasks are individually easy and little forgetting pressure exists to recover from. PHEME-Event (83.1–89.9% in isolation versus 16.8% in sequence) and FNN-Poli-Time (78.5–89.2% versus 29.8%) show the opposite pattern: real forgetting occurs and rehearsal has degradation to repair. The practical reading is that drift severity, and

Table 3.12: EAKD-CFND-R variants across three benchmarks (5 seeds) [13]

Dataset	Variant	Avg. Acc. (%)	BWT (%)
PHEME-Event	DER (reference)	84.8±1.5	−3.2±2.1
	EAKD-CFND (reference)	16.8±0.4	−87.2±1.1
	plain replay	**85.5±1.1**	**−1.8±1.0**
	+PP	85.0±0.7	−1.9±0.8
	+UW	84.3±1.1	−3.1±1.5
	+PP+UW	84.4±0.5	−2.6±1.2
	no-verification	85.4±1.1	−2.0±1.0
FNN-Poli-Time	DER (reference)	30.4±0.9	−80.1±4.0
	EAKD-CFND (reference)	29.8±2.8	−81.6±2.0
	plain replay	34.9±3.2	−72.0±8.0
	+PP	37.2±2.6	−56.5±12.7
	+UW	32.8±1.6	−77.8±4.2
	+PP+UW	**37.8±2.5**	**−64.3±8.1**
	no-verification	34.1±2.5	−72.3±7.9
FNN-Gossip-Time	DER (reference)	29.4±0.3	−85.4±0.5
	EAKD-CFND (reference)	29.1±0.4	−84.6±1.7
	plain replay	29.2±0.3	−85.6±0.7
	+PP	29.4±0.2	−85.4±0.9
	+UW	29.2±0.1	−85.4±0.8
	+PP+UW	29.4±0.2	−85.0±0.6
	no-verification	29.1±0.3	−85.6±0.7

Table 3.13: Task-to-task topical similarity by dataset (TF-IDF task centroids) [13]

Dataset	Mean task-pair similarity
PHEME-Event	0.199
FNN-Poli-Time	0.406
FNN-Gossip-Time	0.816

not dataset identity, determines whether rehearsal is worth its storage cost. Task size remains a candidate confound for PHEME-Event, whose smallest task has 376 training examples.

3.6.10.5 Calibration, Sensitivity, and Cost

Calibration. On FNN-Poli-Time, Expected Calibration Error [340] (15 equal-width bins, 5 seeds) is 0.798 for a 5-member deep ensemble, 0.823 for MC Dropout [333], and 0.842 for both entropy and MSP, the two cheap signals EAKD-CFND actually deploys. All four are poorly calibrated in absolute terms, so the value of $\omega(x)$ must not be read as the probability that the teacher is wrong; such miscalibration is not unusual under distribution shift [341]. A poorly calibrated $\omega(x)$ is one candidate explanation for the null ablation result and for the sensitivity finding below.

Hyperparameter sensitivity. Sweeping $\theta_{uncertainty}$ over 0.5–0.9 and β over 0.25–2.0 (3 seeds per point) leaves accuracy essentially flat (movement of at most 0.4 points, except at $\beta = 2.0$, where accuracy falls by 2.4 points to 28.6% because, for $\beta > 1$, α is clipped to zero over a wide range of uncertainty), whereas the API trigger rate falls from 29% at $\theta_{uncertainty} = 0.5$ to effectively zero at 0.9. The sweep uses three seeds and therefore reports slightly different default-setting values (31.0% accuracy, API rate 0.05) from the five-seed main table and the single-seed cost run. Accuracy is therefore robust to both hyperparameters, but the threshold mainly controls how often the API is called and not what the model learns.

Cost. On FNN-Poli-Time (single seed), EAKD-CFND trains 3.5 times slower than fine-tuning (198.2 s versus 56.5 s) and 1.8 times slower than DER, yet requires less peak memory than DER (4318 MB versus 6679 MB, owing to DER's replay buffer). The API call rate is 0.106 (400 calls over 3770 training instances), and about two-thirds of the total training time (128.4 of 198.2 s; mean latency 0.32 s) is network wait. Of the 400 calls, 360 (90%) fell back to the original label, a blend of transport errors and genuine no-match outcomes that was not instrumented separately; the roughly 40 substituted labels may also form a coverage-biased slice of the teacher's errors, since automated fact-checking coverage is uneven across topic and salience [276].

Table 3.14 reports the training time, inference latency, peak memory, and external API call rate of all six methods on FNN-Poli-Time. Relative to fine-tuning, EAKD-CFND multiplies training time by 3.5 but leaves inference latency and peak memory essentially unchanged; relative to DER, it trains 1.8 times slower yet uses 35% less peak memory, because it stores no replay buffer. Since the uncertainty signal itself is computed from a softmax over a small logit vector, the cost is dominated by network wait on the verification API, which is the component to optimise in deployment (for example through caching or batched queries).

3.6.10.6 Feature Comparison

Table 3.15 maps capability coverage. The distinction is architectural: EAKD-CFND and LUD are statistically tied on all three datasets, so the additional features have not been shown to yield an accuracy advantage.

Table 3.14: Training and inference cost of EAKD-CFND and the baselines on FNN-Poli-Time (single seed, full 10-epoch training) [13]

Method	**Train (s)**	**Infer (ms/inst.)**	**Peak mem. (MB)**	**API call rate**
Fine-tuning	56.5	8.37	4244	—
EWC	56.7	7.27	4260	—
LwF	70.1	7.33	4313	—
DER	111.0	7.71	6679	—
LUD	69.6	7.57	4318	—
EAKD-CFND	198.2	7.19	4318	0.106

Table 3.15: Feature comparison of EAKD-CFND with baseline methods [13]

Method	**CL**	**Adaptive KD**	**Uncertainty-guided KD**	**External knowledge**	**Exemplar-free**
Fine-tuning	×	×	×	×	✓
EWC	✓	×	×	×	✓
LwF	✓	×	×	×	✓
DER	✓	×	×	×	×
LUD	✓	✓	✓	×	✓
EAKD-CFND	✓	✓	✓	✓	✓
EAKD-CFND-R	✓	✓	✓	✓	×

3.6.11 Significance, Limitations, and Deployment Guidance

The practical case for EAKD-CFND is the exemplar-free property itself and not an unconditional accuracy edge. The trade is free on FNN-Poli-Time and FNN-Gossip-Time and costly on PHEME-Event. Rehearsal is worth its storage cost where drift is severe: plain replay recovers almost all of DER's accuracy on PHEME-Event, label-replay variants add a further measurable gain on FNN-Poli-Time, and nothing is gained on FNN-Gossip-Time. The recommendation that follows is exemplar-free by default and rehearsal-augmented when storage is available and drift risk is severe. Inspectability of the verification trigger is a property of the mechanism and not a demonstrated trust benefit, given the poorly calibrated signal that gates it.

The principal limitations are as follows. (i) The adaptive mechanism has more moving parts than standard KD. (ii) The uncertainty signal is poorly calibrated (ECE 0.80–0.84). (iii) The exemplar-free configuration has a measured boundary: it collapses on PHEME-Event (16.8% versus DER's 84.8%). (iv) Neither instance-level weighting nor external verification shows an accuracy contribution outside the noise floor, with or without a rehearsal buffer. (v) Forward Transfer is not reported. (vi) Cost

figures are single-seed and single-dataset, as are the calibration and sensitivity studies. (vii) FNN-Poli-Time's test splits (31, 65, and 28 examples) are small enough that binomial sampling error (about 5.7–8.7 percentage points) exceeds the cross-seed standard deviations reported. (viii) External verification depends on a third-party API whose coverage bias was not audited. (ix) The evaluation covers the class-incremental scenario on two benchmarks, and the framework is text-only. Future work includes larger buffer capacities, an artificial similarity-manipulation split, separate logging of no-match versus transport failures, recording Forward Transfer baselines, calibration-aware thresholds, multimodal extension, and live-stream evaluation.

3.7 Consolidated Comparison of All Four Contributions

Table 3.16 provides a direct comparison of all four contributions of this chapter.

Table 3.16: Consolidated Comparison of Efficient and Intelligent FND Contributions (Three-Label CT-BERT, FakeNews Buster, Graph-RAG, EAKD-CFND)

Property	**Three-Label**	**FakeNews Buster**	**Graph-RAG**	**EAKD-CFND**
Primary approach	Uncertainty thresholding	Adaptive loss scheduling	Evidence retrieval	Instance-level KD weighting
External API role	Dataset enrichment	Post-hoc explainability	Core evidence source	Active training signal
Datasets	Constraint 2021 (COVID)	Constraint 2021 (COVID)	LIAR, FNN, FEVER	PHEME-Event, FNN-Poli, FNN-Gossip
CL mechanism	None (static)	Global seen-state counter	None (static)	Per-instance entropy/MSP
Best metric	93% accuracy	F1: 0.9616 (single-task)	FEVER F1: 0.755 (best of 3 archs)	EAKD-CFND-R 85.5% (PHEME-Event)
Key innovation	Three-label framework	Adaptive KD scheduling	Multi-hop evidence	Uncertainty-weighted KD + verification (no accuracy gain measured)

The progression across the four contributions reflects a coherent research arc. Three-Label CT-BERT identified the uncertainty problem and provided a first-order solution (threshold-based labelling). FakeNews Buster operationalised dynamic adaptation through scheduled KD, but remained bounded by global loss weighting. Graph-RAG addressed evidence grounding through RAG, decoupling detection from model

memory entirely. EAKD-CFND proposed per-instance granularity in KD with uncertainty-triggered external verification (without a measured accuracy gain over fixed-weight KD), and measured, rather than assumed, where the exemplar-free design holds and where it requires rehearsal.

3.7.1 A Unified Framework, Not a Collection of Models

The four contributions are not independent point solutions but composable layers of a single efficient and intelligent fake news detection framework, which is the central artefact fulfilling RO 2. All four share a common BERT-family transformer encoder backbone (CT-BERT in Three-Label CT-BERT, BERT/FastFormer in FakeNews Buster, RoBERTa/DistilBERT in Graph-RAG, BERT-base in EAKD-CFND), so they integrate at the representation level rather than competing as alternatives. Layered on this shared backbone are four orthogonal capabilities: an *uncertainty-aware output layer* (Three-Label CT-BERT) that exposes borderline cases rather than forcing binary verdicts; a *continual-learning adaptation layer* (FakeNews Buster) that maintains performance under concept drift; an *evidence-grounding retrieval layer* (Graph-RAG) that decouples veracity judgement from parametric memory; and a *per-instance uncertainty-weighted distillation layer* (EAKD-CFND) that matches replay-based DER at zero storage on the FakeNewsNet splits, where no method retains much past-task accuracy, and, in its rehearsal-augmented configuration, recovers accuracy under severe drift. The recurring thread that unifies them is the Google Fact Check Tools API verification signal, used for dataset enrichment in Three-Label CT-BERT, post-hoc evidence in FakeNews Buster, and as an active training signal in EAKD-CFND, with Graph-RAG's graph-retrieval providing the equivalent evidence-grounding function from a static knowledge base. A deployment instantiating all four layers would flag uncertain claims (Three-Label CT-BERT), resolve them through retrieved or externally verified evidence (Graph-RAG, EAKD-CFND), and adapt to new data (FakeNews Buster, and EAKD-CFND with exemplar storage required under severe drift), constituting a single end-to-end intelligent detection pipeline rather than a portfolio of disconnected experiments. Such a fully unified deployment has not yet been built or evaluated as a single system; doing so would be a valuable next step in demonstrating the practical benefit of this architectural composability.

3.7.2 Efficiency of the Proposed Framework

The *efficient* qualifier in RO 2 is addressed at the architectural level across the four contributions, although end-to-end real-time deployment engineering remains out of scope (Chapter 6). Three concrete efficiency mechanisms are demonstrated. First, lightweight retrieval-classification: Fast Graph-RAG attains the lowest average inference latency of the three RAG architectures (9.2 ms per claim on a single GPU), by pairing a DistilBERT classifier (40% smaller and 60% faster than BERT while retaining ~97% of its performance [234]) with approximate FAISS indexing. Second, linear-complexity encoding: FakeNews Buster's FastFormer encoder replaces quadratic pairwise self-attention with additive attention of linear complexity, achieving F1 within 0.72 percentage points of BERT at substantially lower cost and making it

the recommended encoder for resource-constrained streaming. Third, exemplar-free continual learning: EAKD-CFND is statistically indistinguishable from replay-based DER on both FakeNewsNet splits while storing *zero* past examples (4318 MB versus 6679 MB peak memory on FNN-Poli-Time, single seed), although it trains 3.5 times slower than fine-tuning owing to API wait and collapses on the five-task PHEME-Event sequence, where the rehearsal-augmented EAKD-CFND-R is required. Collectively these indicate efficiency gains in inference latency (Fast Graph-RAG, measured) and memory (EAKD-CFND versus DER, single seed); the encoder-complexity argument for FastFormer is architectural, and EAKD-CFND's training cost is higher than that of fine-tuning.

3.8 Summary

This chapter presented four interconnected contributions fulfilling RO 2 of this thesis. The three-label CT-BERT framework [11] established that explicitly modelling uncertainty at the classification boundary yields a 13 percentage-point accuracy improvement over binary baselines on COVID-19 fake news detection. FakeNews Buster [12] operationalised dynamic adaptation into a full continual learning architecture, achieving an average continual-learning F1 of 0.92 and reducing the forgetting rate to 0.06, a 60% improvement over fine-tuning. The Graph-RAG framework [14] introduced the RAG-for-Classification paradigm and, through a single-seed audit, showed that multi-hop evidence expansion helps specifically on evidence-chain tasks (FEVER, three-class macro-F1 rising from 0.724 to 0.755) while task structure dominates architecture, with Fast Graph-RAG offering the best accuracy-latency balance for deployment. EAKD-CFND [13] proposed per-instance uncertainty-weighted distillation; under an identical five-seed protocol its exemplar-free configuration ties DER and the other baselines on both FakeNewsNet splits at zero storage but collapses on PHEME-Event (16.8% versus 84.8%), where the rehearsal-augmented EAKD-CFND-R reaches 85.5% and adds a further gain on FNN-Poli-Time (up to 37.8%).

Chapter 4 extends the scope of this thesis to multimodal fake news detection, investigating how the combination of text and image modalities affects detection robustness, adversarial susceptibility, and modality bias in state-of-the-art detection systems.

CHAPTER 4

MULTIMODAL FAKE NEWS DETECTION: ROBUSTNESS, ADVERSARIAL EVALUATION, AND BIAS ANALYSIS

4.1 Introduction

Misinformation increasingly exploits the combination of manipulated images and misleading text: a photograph of a disaster paired with a false caption, or a politician's image captioned with fabricated quotes [9, 107]. Unimodal classifiers cannot detect such cross-modal deception. Multimodal fusion architectures that jointly encode textual and visual features were developed precisely to close this gap.

Whether that promise holds is the problem this chapter investigates. The most striking empirical finding is deceptively simple: remove the image input from a BERT-VGG multimodal fake news detection model and its F1-score collapses from 0.94 to exactly 0.00, indistinguishable from random chance [17]. The 0.94 clean-data baseline is itself partly inflated by a watermark (Clever-Hans) shortcut specific to the D1 dataset, diagnosed in Section 4.3.5 and confirmed on the watermark-free ReCOVery and PolitiFact datasets; the collapse, however, persists across all datasets. This phenomenon, termed *Unimodal Bias*, represents not a marginal degradation but a complete failure of multimodal reasoning. It persists across multiple datasets and is confirmed independently across two companion studies [16, 15]. That a model trained on both text and images can be rendered completely blind by removing one modality reveals a fundamental architectural fragility that neither standard performance benchmarks nor clean-data evaluations expose: architectures often default to whichever modality is most informative in the training dataset rather than genuinely integrating cross-modal evidence.

Two further threats compound the modality bias problem. First, large language models (LLMs) such as DeepSeek-LLM [216] and Llama 3 [215] enable adversaries to generate *semantic inversions*, text that retains grammatical fluency and surface plausibility while reversing the veracity of factual claims. These attacks cannot be detected by pattern matching and they exploit precisely the semantic understanding that transformer-based encoders are supposed to provide. Second, adversarial image perturbations [8, 223] imperceptible to human observers can redirect model attention from semantic features to irrelevant image artifacts, defeating visual encoders that have been shown to focus on watermarks and background noise rather than news content.

This chapter addresses RO 3 of this thesis through three interrelated contributions spanning comprehensive multimodal evaluation, adversarial stress testing, and vision-language robustness analysis. Section 4.2 introduces the unified adversarial attack taxonomy used across all three studies, avoiding redundant repetition in later sections. Section 4.3 presents the comprehensive evaluation of three foundational architectures with the Synergy Gain metric [17]. Section 4.4 extends the evaluation to SOTA models and introduces novel DeepSeek-LLM attacks [16]. Section 4.5 analyses vision-language model architectures and black-box adversarial transferability [15]. Section 4.6 synthesises findings across all three works.

4.2 Related Work and Unified Attack Taxonomy

4.2.1 Multimodal Fusion Architectures

Multimodal fake news detection has evolved from feature concatenation to sophisticated cross-modal attention [107, 110]. EANN [116] employed adversarial training to extract event-invariant representations. MVAE [117] used variational autoencoders for joint text-image modelling. SpotFake+ [118] demonstrated cross-modal transfer learning on limited labelled data. Visual encoders used across architectures include VGG-19 [111], ResNet [112], CLIP [114, 115], and ViT [113]. Cross-modal attention mechanisms are studied in Wu et al. [125], Qian et al. [128], and Wang et al. [129]. VL-BERT [141] and UNITER [142] provide foundational vision-language pretraining. ViLBERT [139] introduced deep co-attentional cross-modal fusion. More recent architectures address the shortcomings of simple late fusion: MFIR [146] introduces an explicit inconsistency reasoning module targeting semantic contradictions between text-image pairs; CroMe [147] employs cross-modal metric learning with Proxy Anchor Loss. Recent VLM-based multimodal FND includes Zhang et al. [150], Chen et al. [151], Niu et al. [153], Yuan et al. [154], Rocamora et al. [148], and Li et al. [155]. Despite these advances, Chen et al. [342] found that multimodal models are not necessarily more robust than unimodal counterparts, a finding extended and confirmed in this chapter.

Additional recent multimodal FND systems include Guo et al. [108], Yan et al. [109, 178], Shen et al. [176], Zhu et al. [177], Liu et al. [158], Zhong et al. [190], Meel and Vishwakarma [131], and Hangloo and Arora [130]. Surveys of multimodal misinformation detection include Abdali et al. [107] and Nasser et al. [110]. Temporal and graph-augmented multimodal systems include Qiu et al. [159], Jing et al. [169], Qian et al. [168], and Meng et al. [163]. Financial domain multimodal FND is studied in Mohankumar et al. [137].

4.2.2 Adversarial Robustness in Multimodal Systems

Liu et al. [343] raised concerns about MFND robustness under missing modalities, identifying Unimodal Bias as a critical unaddressed vulnerability. Yang et al. [181] investigated robustness of multimodal fusion models against single-source adversaries, finding that architectural diversity improves resilience. Adversarial robustness of vision-language models is further studied in Schlarmann et al. [183] and Zhang et al. [182]. Black-box transfer attacks on multimodal systems are examined in Li et al. [180].

Recent adversarial multimodal FND analyses include Lu et al. [185], Liao et al. [186], Jolly and Kumar [187], and Li et al. [188]. AdvFakeBench [206] provides a comprehensive adversarial evaluation benchmark. Domain-transfer robustness for multimodal FND is studied in Mosallanezhad et al. [156]. Uncertainty-aware robustness approaches are explored in Li et al. [162]. Explainability tools including Grad-CAM [106] and Clever Hans diagnostics [192] provide the principled framework applied throughout this chapter to diagnose shortcut learning and attention misdirection. Interpretable multimodal FND is addressed in Chen et al. [189].

4.2.3 The Unified Adversarial Attack Taxonomy

All three contributions in this chapter employ the same core attack suite, defined here once. Sections 4.3–4.5 reference this taxonomy by name without repeating definitions, noting only model-specific or study-specific extensions.

4.2.3.1 Image-Based Attacks (10)

All image attacks operate on the pixel-level input x_V before the visual encoder. The perturbation budget ε is bounded in L_∞ or L_2 norm.

FGSM [8]: single-step gradient attack,

$$x'_V = x_V + \varepsilon \cdot \text{sign}(\nabla_{x_V} J(\theta, x_V, y_{\text{true}})) \tag{4.1}$$

FastFGSM [228]: FGSM with a random start, which initialises the perturbation uniformly inside the ε-ball before a single gradient-sign step,

$$x'_V = \text{Clip}_\varepsilon\big(x_V + \delta_0 + \alpha \cdot \text{sign}(\nabla_{x_V} J(\theta, x_V + \delta_0, y_{\text{true}}))\big), \quad \delta_0 \sim \mathcal{U}(-\varepsilon, \varepsilon) \tag{4.2}$$

MIFGSM [224]: adds momentum μ to stabilise gradient updates and improve transferability. DIFGSM [226]: applies stochastic input transformations (resize, pad) before each gradient step. NIFGSM: integrates Nesterov lookahead into the iterative process.

PGD [223]: the standard strong baseline, iterative with random start and L_∞ projection:

$$x_V^{(t+1)} = \text{Proj}_{\|\cdot\|_\infty \le \varepsilon}\Big(x_V^{(t)} + \alpha \cdot \text{sign}(\nabla_{x_V} J)\Big) \tag{4.3}$$

PGDL2: PGD with L_2-norm constraint. PGDRS: PGD with explicit random initialisation. PGDRSL2: combines random start with L_2 constraint.

Pixle [229]: black-box attack directly optimising sparse pixel values without gradient access.

4.2.3.2 Text-Based Attacks (7)

HotFlip [193]: white-box beam search over character and word embedding gradients to find optimal token swaps. TextBugger [194]: black-box, five character/word perturbation strategies guided by word importance scores. TextFooler [7]: black-box synonym

substitution (WordNet, Universal Sentence Encoder) with semantic similarity threshold $\delta = 0.8$.

Four LLM-generated attacks are produced by fine-tuning *Meta-Llama-3-8B-Instruct* and *deepseek-ai/deepseek-llm-7b-chat* via QLoRA [89] on ∼1,500 fact-checking examples (temperature 0.7, top_p 0.9):

- *L3-Sem*: Llama3 + Semantic Alteration Prompt, which alters the core factual claim while preserving grammatical structure and fluency.
- *L3-Gram*: Llama3 + Grammar Distortion Prompt, which introduces subtle grammatical errors while preserving the original message.
- *DS-Sem*: DeepSeek-LLM + Semantic Alteration Prompt. *Statistically confirmed most potent attack.*
- *DS-Gram*: DeepSeek-LLM + Grammar Distortion Prompt.

The key distinction between LLM attacks and classical methods is qualitative: L3-Sem and DS-Sem produce semantically coherent text that reverses factual veracity while remaining indistinguishable from authentic journalism to human readers. Classical attacks produce detectable perturbations (character noise, synonym substitutions); LLM attacks produce new sentences.

Table 4.1 organises the seventeen attacks into a single taxonomy by target modality, attacker knowledge, and perturbation mechanism. Ten attacks perturb the image (nine gradient-based white-box attacks and one black-box sparse-pixel attack), and seven perturb the text (two classical word-level attacks with white-box and black-box access, one synonym-substitution attack, and four generative attacks produced by fine-tuned LLMs). The LLM-generated attacks are the only ones that alter the factual claim itself, which is why they form a separate branch.

Table 4.1: Taxonomy of the 17 adversarial attacks used across the three multimodal studies

#	Attack	Target	Access	Mechanism	Setting
Image-based attacks (10): pixel-level perturbation of x_V					
1	FGSM	Image	White-box	Single-step gradient sign, L_∞	$\varepsilon = 0.01$
2	FastFGSM	Image	White-box	FGSM with random start	$\varepsilon = 0.01$
3	MIFGSM	Image	White-box	Iterative with momentum	10 steps, $\alpha = 0.001$
4	DIFGSM	Image	White-box	Iterative with input diversity	10 steps, $\alpha = 0.001$

continued on next page

Table 4.1 (continued)

#	Attack	Target	Access	Mechanism	Setting
5	NIFGSM	Image	White-box	Iterative with Nesterov lookahead	10 steps, $\alpha = 0.001$
6	PGD	Image	White-box	Iterative, random start, L_∞ projection	$\varepsilon = 0.01$
7	PGDL2	Image	White-box	PGD with L_2 constraint	$\varepsilon = 0.1$
8	PGDRS	Image	White-box	PGD with explicit random start	$\varepsilon = 0.01$
9	PGDRSL2	Image	White-box	Random start with L_2 constraint	$\varepsilon = 0.1$
10	Pixle	Image	Black-box	Sparse pixel rearrangement	max_pixels = 5
Text-based attacks (7): perturbation of the claim text x_T					
11	HotFlip	Text	White-box	Gradient-guided token swaps	Beam search
12	TextBugger	Text	Black-box	Character/word perturbations ranked by importance	Five strategies
13	TextFooler	Text	Black-box	Synonym substitution with similarity check	$\delta = 0.8$
14	L3-Sem	Text	Black-box	Llama3 semantic alteration (claim reversal)	QLoRA; $T = 0.7$
15	L3-Gram	Text	Black-box	Llama3 grammar distortion	QLoRA; $T = 0.7$
16	DS-Sem	Text	Black-box	DeepSeek-LLM semantic alteration (claim reversal)	QLoRA; $T = 0.7$
17	DS-Gram	Text	Black-box	DeepSeek-LLM grammar distortion	QLoRA; $T = 0.7$

4.2.3.3 Modality Ablation Protocol

Across all three studies, modality ablation follows a uniform three-condition protocol: Baseline (Both): standard evaluation; Image-Only: textual input replaced with padding; Text-Only: visual input replaced with a zero tensor.

4.2.3.4 The Synergy Gain Metric

Standard F1 cannot distinguish a model that genuinely integrates both modalities from one that defaults to its strongest single modality. The Synergy Gain S is defined as:

$$S = F1_{\text{Multi}} - \max(F1_{\text{TextOnly}},\ F1_{\text{ImageOnly}}) \tag{4.4}$$

$S > 0$ indicates effective cross-modal integration; $S \approx 0$ indicates the model is functionally unimodal.

4.3 Comprehensive Evaluation of Multimodal FND

4.3.1 Motivation and Scope

Prior to this contribution [17], multimodal FND evaluation was conducted exclusively on clean test sets. This section addresses two gaps: systematic adversarial robustness across the full attack taxonomy of Section 4.2.3 (15 of 17 attacks; the Comprehensive Multimodal Evaluation predates the DeepSeek extensions), and formal modality removal analysis to quantify Unimodal Bias.

4.3.2 Models and Datasets

Three foundational architectures share a common co-attention fusion mechanism. Textual features $\mathbf{t}$ and visual features $\mathbf{v}$ (VGG-19, ImageNet pre-trained) are fused via:

$$h_{\text{att}} = \tanh(W_t \cdot \mathbf{t} + W_v \cdot \mathbf{v} + b_a) \tag{4.5}$$

$$\alpha_i = \text{softmax}(w_i^\top \cdot h_{\text{att}}), \quad i \in \{t, v\} \tag{4.6}$$

$$F = \alpha_t \odot \mathbf{t} + \alpha_v \odot \mathbf{v} \tag{4.7}$$

The three models differ only in the text encoder: BertVgg uses BERT (LR $3{\times}10^{-5}$, batch 8); CnnVgg uses TextCNN with 100-dimensional n-gram embeddings (LR 1×10^{-4}, batch 16); LstmVgg uses a 512-unit LSTM (LR 1×10^{-4}, batch 16). All use Adam, dropout 0.2, early stopping with patience 10.

D1 [292]: 4,843 samples (2,613 fake / 2,230 real), balanced; many fake images contain explicit "FAKE"/"MISLEADING"/"HOAX" text watermarks. RECOVERY [119]: 1,445 samples (310 fake / 1,135 real), class-imbalanced, no explicit watermarks. Split: 70/15/15, stratified. Results: mean ± SD over 5 random seeds.

The threat surface against which these architectures are evaluated is summarised in Fig. 4.1, which identifies the injection points at which textual and pixel-level adversarial perturbations enter the co-attention fusion pipeline.

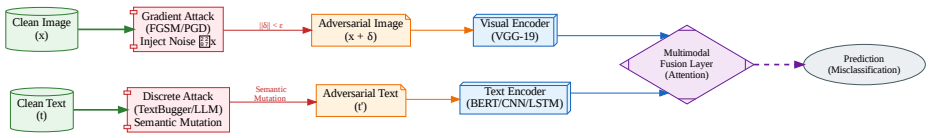


Figure 4.1: Multimodal threat model for the foundational architectures: adversarial perturbations are injected at the textual input (semantic and grammatical attacks) and at the pixel-level visual input (gradient-based and black-box attacks) prior to co-attention fusion [17].

4.3.3 Baseline Performance and Adversarial Impact

Table 4.2 summarises the baseline F1 of the three foundational architectures against SOTA models and the worst-case adversarial attack impact on each.

Table 4.2: Baseline F1 vs SOTA and Adversarial Attack Impact Summary [17]

Model	**D1 F1**	**RECOVERY F1**	**Worst-case attack (F1)**
CnnVgg	**0.98±0.01**	**0.99±0.01**	TextBugger: ~0.70
BertVgg	0.94±0.02	0.98±0.01	TextFooler: ~0.62
LstmVgg	0.93±0.02	0.95±0.02	FGSM / Pixle: ~0.70
SpotFake [118]	0.97±0.01	0.98±0.01	—
EANN [116]	0.93±0.01	0.97±0.01	—
Att-RNN [124]	0.95±0.02	0.96±0.01	—

Under the image attack suite (Section 4.2.3), all models remain relatively stable: F1 stays above 0.80 under gradient-based attacks. Under text attacks, degradation is substantially greater, consistent with the general finding that text-based attacks are the more effective modality for adversarial manipulation. BertVgg, despite its sophisticated BERT encoder, shows the greatest text vulnerability (F1 drops to 0.62 under TextFooler), while CnnVgg's n-gram local features provide a degree of robustness against semantic attacks. The per-architecture performance profiles across the full attack suite are compared in Fig. 4.2, exposing the divergent degradation signatures of the three text encoders under identical visual fusion.

4.3.4 Modality Removal and Unimodal Bias

Table 4.3 reports the modality-removal results on the D1 dataset, quantifying each model's reliance on individual modalities through the Synergy Gain metric.

The Synergy Gain column interprets the results directly. BertVgg achieves $S = 0.00$ on both datasets: its Image-Only F1 equals its full-modality F1, confirming that the BERT encoder contributes nothing to the fusion. LstmVgg achieves the highest $S = +0.20$, indicating that LSTM sequential text features and VGG-19 spatial features

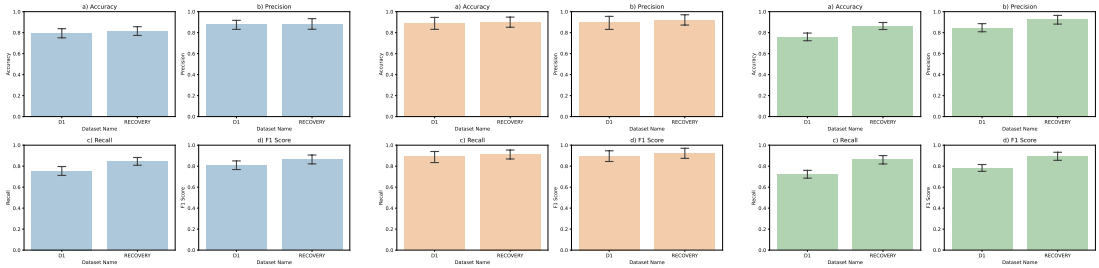


Figure 4.2: Per-architecture performance metrics (accuracy, precision, recall, F1) under the unified attack suite, shown for BertVgg (left), CnnVgg (centre), and LstmVgg (right). The divergent degradation profiles confirm that text-encoder choice, not the shared VGG-19 visual backbone, governs adversarial vulnerability [17].

Table 4.3: Modality Removal Results on D1 Dataset [17]. Bold: catastrophic failure.

Model	**Condition**	**Accuracy**	**Precision**	**Recall**	**F1**	**Synergy** S
CnnVgg	Baseline	0.98±0.01	0.99±0.01	0.98±0.01	0.98±0.01	
CnnVgg	Image-Only	0.92±0.02	0.97±0.02	0.88±0.03	0.92±0.02	+0.06
CnnVgg	Text-Only	0.78±0.04	0.57±0.05	0.97±0.02	0.72±0.04	
BertVgg	Baseline	0.94±0.02	0.97±0.01	0.92±0.02	0.94±0.02	
BertVgg	Image-Only	0.93±0.02	0.95±0.02	0.92±0.03	0.94±0.02	0.00
BertVgg	Text-Only	**0.50±0.01**	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**	
LstmVgg	Baseline	0.92±0.02	0.95±0.02	0.90±0.03	0.93±0.02	
LstmVgg	Image-Only	0.63±0.03	1.00±0.00	0.57±0.04	0.73±0.03	+0.20
LstmVgg	Text-Only	0.54±0.03	0.12±0.05	0.76±0.06	0.21±0.04	

provide genuinely complementary information. CnnVgg's moderate $S = +0.06$ reflects partial but real cross-modal integration.

A practical Synergy Gain threshold. For deployment screening, the observed values suggest a two-part rule. First, the Synergy Gain should reach $S \geq 0.05$, which is about 2.5 times the typical cross-seed F1 standard deviation of 0.02 on D1, so that a positive S cannot be attributed to noise; BertVgg ($S = 0.00$) fails this test, whereas CnnVgg ($S = +0.06$) passes it. Second, because S alone can be inflated when both single-modality F1 scores are low (LstmVgg reaches $S = +0.20$ only because its Image-Only and Text-Only F1 are 0.73 and 0.21), the weaker modality should retain at least half of the full-model F1 when the dominant modality is removed. Table 4.4 applies this rule to the five architectures evaluated on D1: only CnnVgg and ViLBERT retain more than half of their baseline F1 under Text-Only input, so only they are considered to integrate both modalities. The thresholds are dataset-dependent, since the D1 watermark shortcut raises Image-Only F1, and should be recalibrated on a held-out, artefact-free split before use on a new corpus.

Table 4.4: Screening of the five architectures with the Synergy Gain rule ($S \geq 0.05$ and Text-Only F1 $\geq 0.5\times$ baseline F1; D1 dataset) [17, 15]

Model	S (D1)	Text-Only F1	Retention	Verdict
CnnVgg	+0.06	0.72	0.73	Integrates both modalities
BertVgg	0.00	0.00	0.00	Unimodal collapse
LstmVgg	+0.20	0.21	0.24	Synergy from complementary weakness; fails retention
BERT-CLIP	—	0.10	0.11	Unimodal collapse
ViLBERT	—	0.75	0.77	Retains text-only competence

4.3.5 XAI Diagnosis: The Clever Hans Effect

Grad-CAM [106] visualisations explain the root cause of BertVgg's visual dominance. In the clean state, attention maps focus almost exclusively on "True!" and "Fake!" text watermarks in the D1 images rather than semantic content, a textbook Clever Hans [192] failure. When PGD disperses this narrow focus, the model shifts attention to arbitrary background pixels, causing misclassification. This diagnosis has a methodological implication for the field: the high clean-data F1 on D1 partly reflects watermark detection, not genuine multimodal reasoning. Results on RECOVERY, which lacks watermarks, provide a more honest benchmark. Fig. 4.3 shows the Grad-CAM attention maps before and after the PGD attack.

4.4 Adversarial Stress Testing with LLM-Driven Perturbations

4.4.1 Motivation: SOTA Architectures and DeepSeek Attacks

Section 4.3 established the attack taxonomy and foundational vulnerabilities. Two questions remained open: do SOTA architectures with more sophisticated fusion mechanisms resolve foundational fragility, and do different LLM architectures produce attacks of meaningfully different potency? This work [16] answers both by evaluating five architectures (three foundational plus MFIR [146] and CroMe [147]) across the full 17-attack suite on three datasets (D1, RECOVERY, PolitiFact), with Wilcoxon signed-rank statistical testing.

4.4.2 SOTA Architectures

MFIR [146] pairs BERT-Base with ResNet-50 through an *inconsistency reasoning module* that explicitly models semantic contradictions between text-image pairs, addressing the core failure mode of simple late fusion. Configuration: LR 2×10^{-5}, batch 32, dropout 0.1.

CroMe [147] employs BLIP-2 as text encoder and ViT-L/14 as image encoder, fused through a Cross-Modal Tri-Transformer with Proxy Anchor Metric Learning loss that

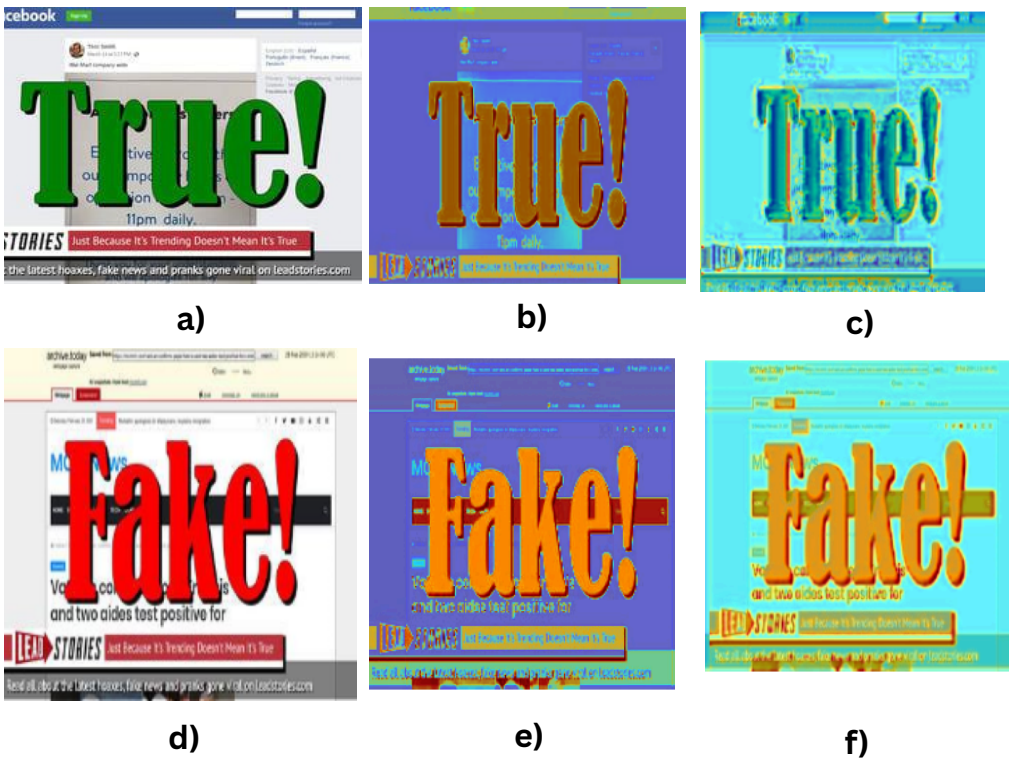


Figure 4.3: Grad-CAM visualisations: clean-state attention (left) focuses on watermark text overlays (Clever Hans behaviour); post-PGD attention (right) scatters to background, causing misclassification [17].

enforces tight intra-modal clustering. Configuration: LR 1×10^{-5}, batch 16, dropout 0.1.

4.4.3 PolitiFact: A Third Dataset

The PolitiFact dataset [272] provides ~3,000 political discourse samples (~35% fake) sourced from PolitiFact.com fact-checking articles. Unlike D1 (watermark-driven visual cues) and RECOVERY (disaster imagery), PolitiFact pairs textual political claims with generic stock photographs of politicians; the semantic veracity signal resides almost entirely in the text.

4.4.4 LLM Attack Examples

The qualitative distinction between semantic inversions and classical attacks is illustrated in Table 4.5. DS-Sem reverses the entire factual meaning of a COVID-19 pandemic description while producing a grammatically natural, plausible sentence that no human reader would identify as adversarial. The corresponding effect on the visual modality is shown in Fig. 4.4 for a PolitiFact sample, where an imperceptible image perturbation redirects model attention away from the salient region.



Figure 4.4: Adversarial perturbation on a PolitiFact sample: original image (left), model attention on the clean input (centre), and attention after an imperceptible adversarial perturbation (right). The attention map is dispersed away from the discriminative region, illustrating the visual-modality attack on the text-dominant PolitiFact domain [16].

Table 4.5: LLM Adversarial Text Examples from RECOVERY Dataset [16]

Variant	**Text**
Original	"... has spread to six continents. Hundreds of thousands have been infected, at least 20,000 people have died and the spread of the coronavirus was called a pandemic by the WHO in March."
DS-Sem	"... **has been largely confined to one region. Fewer than expected cases have been documented, minimal loss of life has occurred, and the spread of the coronavirus was not declared a pandemic** by the WHO in March."
L3-Gram	"... **in last** December has spread to six continents, **where** hundreds of thousands have been infected, **with at lest** 20,000 people having died."

4.4.5 Results: Modality Reversal on PolitiFact and DeepSeek Superiority

The PolitiFact reversal in Table 4.6 is a critical finding: MFIR's Text-Only F1 (0.890) substantially exceeds its Image-Only F1 (0.719). This is the opposite of the D1/RECOVERY pattern, where image features dominate. The reversal is explained by dataset composition: political fact-checking depends on the semantic content of textual claims; the associated stock photographs of politicians carry no discriminative information. This establishes that modality preference is a property of the dataset, not the model; a practitioner cannot assume that visual dominance generalises across deployment domains. The per-dataset modality contributions are quantified in Fig. 4.5, which contrasts baseline, text-only, and image-only F1 across the five architectures and three datasets.

Under DS-Sem, all five architectures degrade substantially despite their baseline superiority: CroMe (F1=0.9921 at baseline) falls to 0.815 on D1; MFIR falls to 0.805.

Table 4.6: Modality Ablation: SOTA Models on PolitiFact (LLM-Driven Perturbations, C&EE). Note text-over-image reversal for MFIR [16].

Model	Condition	F1	Modality preference
MFIR	Baseline	0.970	
MFIR	Text-Only	0.890	**Text > Image**
MFIR	Image-Only	0.719	
CroMe	Baseline	0.955	
CroMe	Text-Only	0.885	Text > Image
CroMe	Image-Only	0.750	

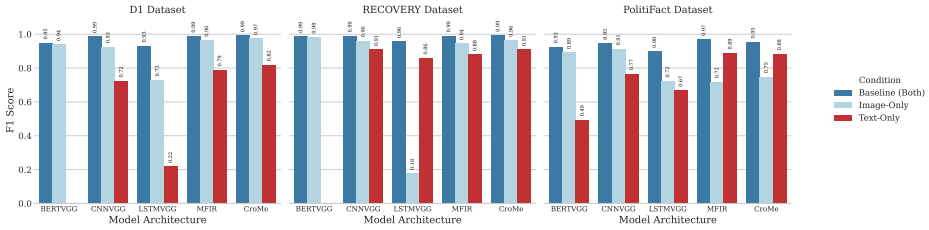


Figure 4.5: Modality ablation across the five architectures and three datasets: baseline versus text-only versus image-only F1. The PolitiFact bars reverse the D1/RECOVERY pattern, confirming that modality preference is dataset-dependent [16].

The relative potency of each attack across datasets is mapped in Fig. 4.6. The Wilcoxon signed-rank test confirms that DS-Sem and DS-Gram cause statistically significantly greater F1-score degradation than L3-Sem and L3-Gram ($p < 0.05$) across all model-dataset combinations. DeepSeek's semantic inversions are more potent than Llama3's, and this superiority is not random fluctuation.

4.4.6 Proposed Defences

Five architectural blueprints are proposed (empirical validation remains future work): Adversarial Training with Semantic Curriculum, progressively exposing the model to stronger attacks during training; Semantic Sanitisation Layer, an upstream lightweight LLM that rewrites input text to canonical journalistic form before classification; Input Sanitisation for image pre-processing; Confidence-Based Gating, monitors high-frequency noise spectra to detect PGD-style image attacks and suppresses the compromised visual modality; Ensemble Diversity exploiting complementary failure modes across architectures.

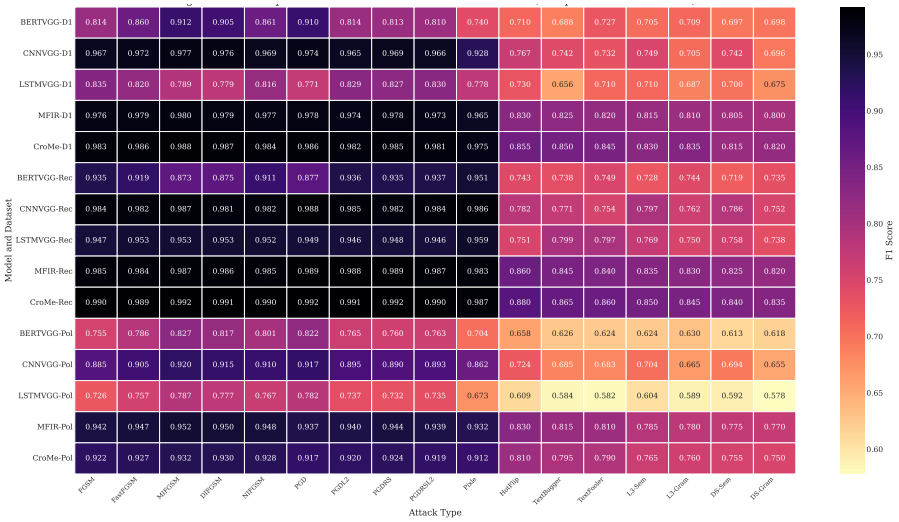


Figure 4.6: Heatmap of F1-scores across the full attack suite. Each row is a model–dataset pair (the models BERTVGG, CNNVGG, LSTMVGG, MFIR, and CroMe, each on D1, RECOVERY, and PolitiFact), and each column is one of the 17 attacks (10 image attacks followed by 7 text attacks). Lighter cells indicate lower F1-scores; the DeepSeek semantic inversion (DS-Sem) column is among the lightest, marking it as one of the most potent attack classes [16].

4.5 Modality Dependence and Vision-Language Model Robustness

4.5.1 Motivation: Visual Backbone vs Fusion Strategy

The BertVgg collapse under modality removal raises a precise architectural question: is the visual encoder (VGG-19) the source of the vulnerability through its exploitation of watermark artifacts, or is the late fusion strategy itself the problem? Two novel architectures disentangle these hypotheses [15].

BERT-CLIP replaces VGG-19 with CLIP [114] (ViT-B/32, 512-dim visual embeddings) while retaining the identical late co-attention fusion of Eqns. 4.5–4.7. If the visual backbone is the problem, BERT-CLIP should resolve the collapse.

ViLBERT [139] retains Faster R-CNN region features but adopts a different fusion strategy: co-attentional transformer layers where text and visual streams attend to each other at every transformer layer, not just at the final classification head. If the fusion strategy is the problem, ViLBERT should show qualitatively different behaviour.

The experimental design that disentangles these two hypotheses through controlled modality restriction and black-box transfer evaluation is outlined in Fig. 4.7.

4.5.2 Architecture Configurations

BERT-CLIP: Longformer [70] text encoder + CLIP visual encoder; same co-attention fusion; LR 3×10^{-5}, batch 8, dropout 0.2, hidden dim 512.

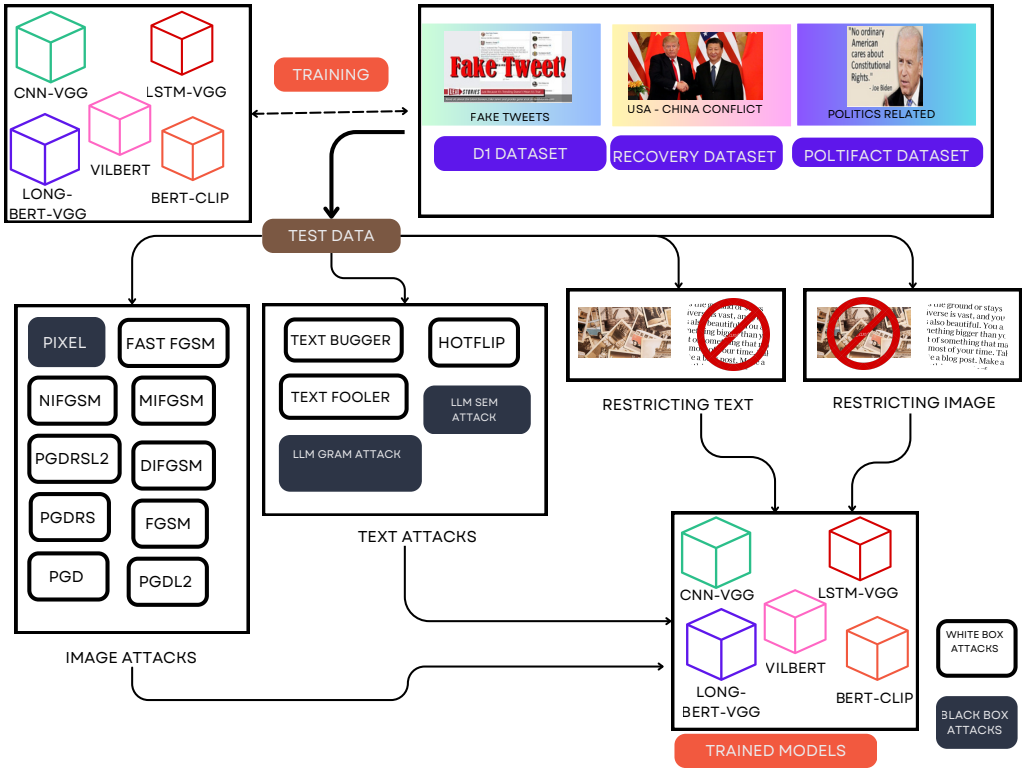


Figure 4.7: Experimental design for isolating the backbone-versus-fusion question: each architecture is evaluated under full-modality, image-only, and text-only conditions, then subjected to white-box and black-box transfer attacks to assess adversarial robustness independently of the visual encoder choice [15].

ViLBERT [139]: BERT text stream + Faster R-CNN region-feature visual stream; co-attentional transformer layers enabling cross-modal attention at each level; LR 2×10^{-5}, batch 8, dropout 0.1, hidden dim 768.

4.5.3 Modality Ablation: Backbone versus Fusion

Table 4.7 reports the modality ablation across all five architectures on the D1 dataset.

The result is clear. BERT-CLIP, with CLIP replacing VGG-19, achieves a Text-Only F1 of 0.104 on D1, marginally better than BERTVGG's 0.000 but still in the same catastrophic failure mode. *Swapping the visual backbone does not resolve Unimodal Bias.* ViLBERT, with deep co-attentional fusion, achieves Text-Only F1 of 0.748 on D1 and 0.931 on RECOVERY. The deep fusion forces the model to build a text representation enriched by visual context at every transformer layer, producing an encoding that remains meaningful even when the image is absent. *The fusion strategy, not the visual encoder choice, is the architectural variable critical to unimodal robustness.*

4.5.4 Black-Box Transfer Attacks and Statistical Confirmation

The black-box transfer analysis is the first systematic evaluation of adversarial transferability in MFND [15]. PGD perturbations generated on BERTVGG transfer effectively

Table 4.7: Modality Ablation Across All Five Architectures, D1 Dataset [15]. Bold: catastrophic failure; underline: best text-only F1. Mean ± SD, N=3 seeds (the Cluster Computing study used three seeds; the AJSE study of Section 4.3 used five).

Model	Condition	Accuracy	Precision	Recall	F1
BERTVGG	Baseline	0.948±0.003	0.977±0.002	0.923±0.004	0.949±0.003
BERTVGG	Text-Only	**0.500±0.000**	**0.000±0.000**	**0.000±0.000**	**0.000±0.000**
BERT-CLIP	Baseline	0.952±0.003	0.980±0.002	0.929±0.004	0.954±0.003
BERT-CLIP	Text-Only	**0.520±0.005**	**0.150±0.010**	**0.080±0.005**	**0.104±0.007**
ViLBERT	Baseline	0.975±0.002	0.982±0.002	0.971±0.003	0.976±0.002
ViLBERT	Text-Only	0.812±0.006	0.615±0.007	0.950±0.003	0.748±0.005
CNNVGG	Baseline	0.987±0.002	0.990±0.001	0.984±0.003	0.987±0.002
CNNVGG	Text-Only	0.780±0.006	0.577±0.008	0.972±0.003	0.724±0.005
LSTMVGG	Baseline	0.930±0.004	0.954±0.003	0.909±0.005	0.931±0.004
LSTMVGG	Text-Only	0.544±0.008	0.128±0.010	0.765±0.015	0.219±0.012

to BERT-CLIP. Because PGD acts in pixel space and the two models use different visual encoders (VGG-19 and CLIP), the transfer cannot be attributed to shared visual features or to the text encoder; a shared co-attention fusion design and a common reliance on the image modality are plausible but untested explanations. DS-Sem attacks transfer across all five architectures including ViLBERT, confirming that LLM semantic inversions represent *systemic* threats rather than architecture-specific vulnerabilities. An adversary with black-box access to any one deployed model can generate attacks that degrade all five.

The paired t-test across all model-dataset combinations confirms $p < 0.05$ for DeepSeek > Llama3, replicating the Wilcoxon result of Section 4.4 through an independent statistical method. Convergent evidence from two different statistical tests on two independent study designs constitutes strong support for this threat model finding.

4.6 Unified Analysis Across the Multimodal FND Contributions

4.6.1 The Research Programme in Full

Table 4.8 consolidates the scope and principal findings of the three multimodal contributions.

4.6.2 Conclusions

Conclusion 1: Unimodal Bias is the central failure mode. Across all three studies, seven architectures, and three datasets, catastrophic visual-modality dependence emerges in BERT-VGG, BERT-CLIP, and LSTM-VGG on image-watermarked or visually rich

Table 4.8: Scope and Principal Findings of the Three Multimodal Contributions

Dimension	**AJSE [17]**	**C&EE [16]**	**Cluster [15]**
Models	BertVgg, CnnVgg, LstmVgg	+ MFIR, CroMe (SOTA)	+ BERT-CLIP, ViLBERT (VLMs)
Datasets	D1, RECOVERY	+ PolitiFact	Same
Text attacks	5 (incl. L3)	7 (+ DS-Sem/Gram)	Same 7
Image attacks	10	Same 10	Same 10
Statistical test	None	Wilcoxon ($p < 0.05$)	Paired t-test ($p < 0.05$)
Black-box transfer	Not evaluated	Not evaluated	Systematic
Novel contribution	Unimodal Bias + Synergy Gain	DeepSeek > Llama3; SOTA not immune	Fusion strategy ≠ backbone
Worst-case F1 (any model)	0.00 (BertVgg, no images)	0.67 (BertVgg, DS-Sem)	0.00 (BertVgg on D1 and RECOVERY; BERT-CLIP on RECOVERY, 0.10 on D1; no images)

datasets. Even the most advanced architectures (MFIR, CroMe) show significant modality bias on PolitiFact. Only ViLBERT's deep co-attentional fusion provides genuine resilience.

Conclusion 2: LLM semantic inversions are the most dangerous attack class. Classical adversarial perturbations produce bounded, recoverable degradation. DS-Sem attacks reduce state-of-the-art CroMe from F1=0.99 to F1=0.82 on D1, are indistinguishable from legitimate news, and transfer via black-box access. Two independent statistical tests ($p < 0.05$, Wilcoxon and paired t-test) confirm that DeepSeek generates more potent attacks than Llama3.

Conclusion 3: Modality preference is dataset-dependent. D1 and RECOVERY are visually dominant; PolitiFact reverses to text dominance. Deployment decisions about architecture selection must account for the modality composition of the target domain, a finding absent from all prior MFND benchmarking studies.

4.6.3 Deployment Decision Framework

Table 4.9 translates these conclusions into an architecture selection guide for practitioners.

4.7 Summary

This chapter presented three interrelated studies addressing RO 3. The foundational evaluation [17] established Unimodal Bias as a pervasive failure mode quantified

Table 4.9: Architecture Selection Guide Derived from the Multimodal Evaluation

Deployment Context	**Recommended**	**Rationale**
Image-rich social media, high throughput	CnnVgg	Most consistent robustness; $O(k \cdot n \cdot d)$ complexity
Clinical robustness, defence-aware	ViLBERT	Best text-only resilience; resists single-modality removal
Political fact-checking (text-dominant)	MFIR	Text-dominant modality aligns with domain
LLM adversary expected	CnnVgg + ViLBERT ensemble	Complementary failure modes reduce DS-Sem transferability

through the Synergy Gain metric S, and diagnosed the D1 dataset's Clever Hans watermark shortcut as a confounding factor in published benchmarks. The adversarial stress-testing study [16] demonstrated through Wilcoxon signed-rank testing ($p < 0.05$) that DeepSeek-LLM semantic inversions outperform Llama3 attacks and that even SOTA architectures are not immune, while the PolitiFact dataset revealed that modality preference is dataset-dependent rather than architecture-intrinsic. The VLM robustness analysis [15] addressed the backbone-vs-fusion question: BERT-CLIP mirrors BERTVGG's collapse (F1=0.104 on D1 without images), which is consistent with the fusion strategy, rather than the visual encoder, determining unimodal resilience, although the text encoder is a confound. ViLBERT's deep co-attentional fusion yields the best text-only performance (F1=0.931 on RECOVERY), making it the most resilient architecture tested.

Chapter 5 broadens the scope to the COVID-19 infodemic context, complements the review of Chapter 2 (RO 1) with a domain-specific case study, and conducts a comparative evaluation of the text-based contributions (RO 4).

CHAPTER 5

COMPARATIVE EVALUATION OF FAKE NEWS DETECTION FRAMEWORKS AND THE COVID-19 INFODEMIC CASE STUDY

5.1 Introduction

The COVID-19 pandemic, which emerged in late 2019, was accompanied by a parallel information catastrophe: an unprecedented proliferation of health misinformation that the World Health Organization designated an *infodemic* [344]. False claims about vaccine safety, fabricated treatment protocols, conspiracy theories regarding the origins of the virus, and deliberate distortions of public health guidance spread across social media platforms at speeds that far exceeded the capacity of human fact-checkers and overwhelmed existing automated detection pipelines. The infodemic was not merely disruptive: it caused measurable harm to public behaviour, vaccine uptake, and mortality across multiple countries, establishing the automated detection of health misinformation as a problem with immediate consequences rather than a theoretical research challenge [2].

The fake news detection systems reviewed and developed in Chapters 3 through 4 were constructed and evaluated under largely stable, closed-world benchmark conditions. The COVID-19 infodemic simultaneously exposed five structural failure modes in this paradigm: static veracity assumptions that collapse as scientific consensus evolves; systematic vulnerability to adversarial text perturbations; the inability to flag misinformation at publication time before propagation occurs; persistent blindness to multimodal misinformation payloads; and poor generalisation across disease domains and languages. These interconnected limitations, none of which is fully addressed by any single existing system, remain the central unresolved problem in automated health misinformation detection.

This chapter addresses RO 4, the comparative result analysis of the developed frameworks against existing techniques. The review of fake news detection techniques (RO 1) is fulfilled in Chapter 2; here that review is complemented by a domain-specific case study. Section 5.2 presents the COVID-19 infodemic as an applied case study, synthesising 221 peer-reviewed publications on COVID-19 fake news detection from 2011 to May 2026 under the PRISMA protocol and surfacing the real-world failure modes and deployment requirements that frame the evaluation. Section 5.3 then addresses RO 4 directly through a unified evaluation of the four text-based FND

contributions from Chapter 3 across uncertainty handling, concept drift resilience, and explainability dimensions. Section 5.4 synthesises the case study and the comparative evaluation. Section 5.5 links forward to Chapter 6.

5.2 Mitigating the Infodemic: COVID-19 Fake News Detection and Epidemic Preparedness

This section uses the COVID-19 Infodemic Review [10] as an applied case study. The COVID-19 pandemic was the largest health-misinformation event yet recorded. False claims about vaccines, cures, and where the virus came from spread faster than verified guidance, and they measurably lowered vaccine uptake and compliance with public-health advice; in documented cases they contributed to avoidable harm. The case study uses this episode to show, under realistic deployment conditions, the gaps that automated detection has to close. Veracity drifts over time as scientific consensus shifts. The early-detection "golden window" closes before propagation graphs form. LLM-generated text defeats the stylistic cues that many classifiers depend on. Multilingual content and closed messaging platforms remain blind spots, and public-health decision-makers receive verdicts with no grounded evidence behind them. These failure modes are not specific to COVID-19; they recur in any fast-moving outbreak. The section argues that consolidating them into a reusable, epidemic-aware detection roadmap is a prerequisite for handling future pandemics and epidemics, whether a renewed coronavirus wave, an Ebola or Marburg flare-up, or the 2026 Andes hantavirus cluster examined later in this chapter.

5.2.1 The COVID-19 Infodemic: Scale, Spread, and Documented Harm

Global interest in the keyword "fake news", as measured by the Google Trends Popularity Score (0–100), peaked at approximately 100 in March 2020, coinciding with the WHO pandemic declaration (Fig. 5.1). Research activity in the corpus follows this trajectory exactly: the peak publication year was 2021 with 50 papers, preceded by 37 papers in 2020, with a renewed surge to 32 papers in 2025 driven by LLM-based detection and retrieval-augmented approaches (Fig. 5.2).

Three structural properties distinguish the COVID-19 infodemic from prior misinformation episodes. First, *temporal veracity drift*: scientific recommendations on mask usage, herd immunity thresholds, and reinfection risk changed materially within months, causing automated classifiers trained on early-pandemic labels to systematically misclassify later accurate reporting as misinformation. This failure mode is absent from all standard benchmark evaluation protocols, which assume a static ground truth. Second, the *walled-garden problem*: encrypted platforms including WhatsApp and Telegram served as primary propagation channels for health misinformation, yet are structurally inaccessible to automated monitoring. Third, *neural fake news*: as the pandemic progressed, LLMs became capable of generating health misinformation that occupies the same stylometric feature space as authentic clinical communication, invalidating the surface-feature-based detection approaches that had previously dominated the field [10].

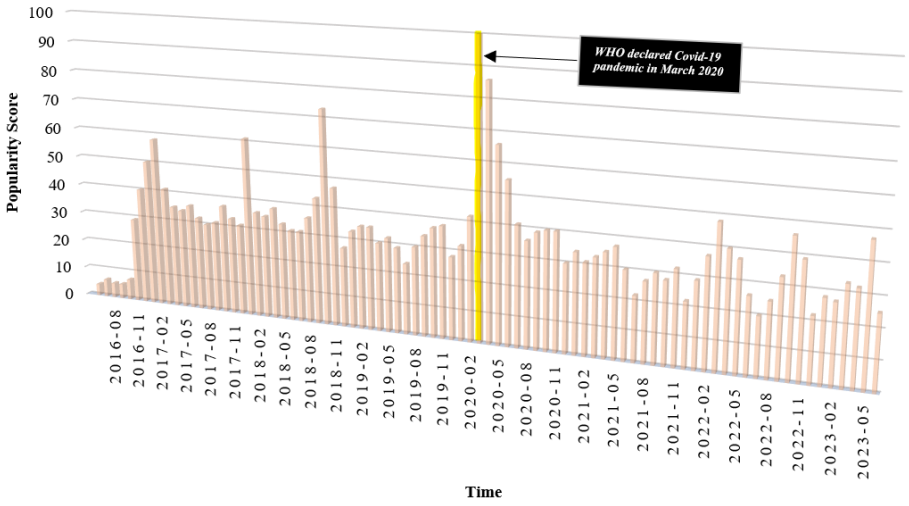


Figure 5.1: Worldwide interest in the keyword "fake news" measured by Google Trends Popularity Score (0–100). The sharp peak in March 2020 coincides with the WHO pandemic declaration. A sustained elevated baseline through 2024–2026 reflects continued public concern and the emergence of new infodemics, including the 2026 Andes hantavirus outbreak [10].

The prevalence of COVID-19 misinformation and its measurable health harms have been documented across multiple independent studies. Yang et al. [345] quantified the prevalence of COVID-19 health misinformation on social media and its association with vaccination hesitancy. Melki et al. [346] analysed COVID-19 infodemic dimensions and their psychological impacts. Kleinberg et al. [347] examined how COVID-19 misinformation spread on social platforms using computational approaches. Martins et al. [348] studied COVID-19 fake news in Brazil, establishing the multilingual dimension of the infodemic. Agarwal [349] examined the adjustment of health-seeking behaviours in response to COVID-19 misinformation. Dharawat et al. [350] proposed methods for countering COVID-19 misinformation with evidence-based correction. McGlynn et al. [351] analysed the relationship between COVID-19 misinformation beliefs and health outcomes. Cui and Lee [352] studied COVID-19 information credibility on Twitter using mixed-methods approaches. The WHO [353] has established formal guidance on infodemic management as a public health intervention.

5.2.2 Systematic Review Methodology

Four research questions (RQs) guided the review: RQ1 — What automated detection techniques have been applied to COVID-19 misinformation, and how do they perform on benchmarks? RQ2 — What datasets, knowledge bases, and infrastructure resources exist for pandemic FND research? RQ3 — What critical methodological and infrastructural gaps limit deployment in real-world epidemic contexts? RQ4 — How do COVID-19 detection lessons generalise to emerging epidemic infodemics such as the 2026 hantavirus outbreak?

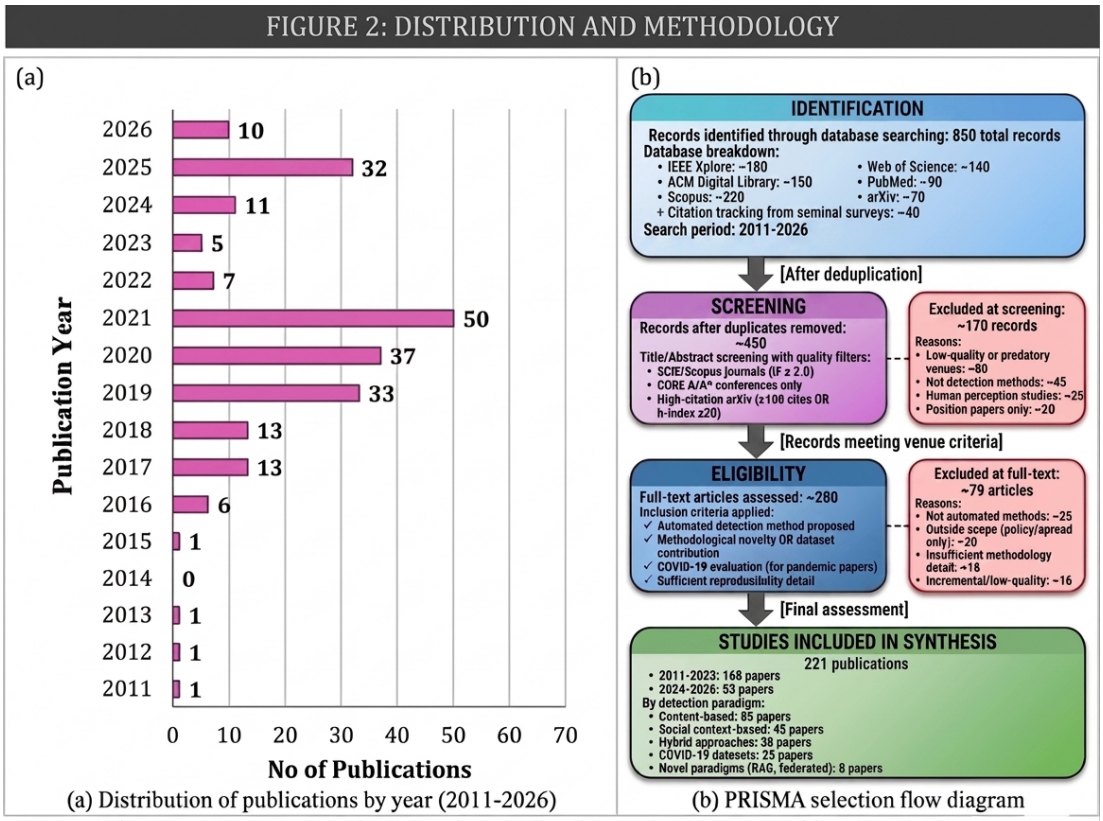


Figure 5.2: Systematic literature review methodology and corpus distribution (n = 221). (a) Publication distribution by year (2011–2026) showing the sharp increase following the March 2020 pandemic declaration, with a secondary surge from 2024–2026 driven by LLM-based detection, retrieval-augmented approaches, and multimodal misinformation research. (b) PRISMA selection flow: initial database search across six sources yielded ≈850 records; after deduplication, abstract screening with venue-quality filters, and full-text eligibility assessment, 221 publications were included [10].

A PRISMA-protocol systematic search was conducted across six databases, including ACM Digital Library (≈150 records), IEEE Xplore (≈180), Scopus (≈220), Web of Science (≈140), PubMed (≈90), and arXiv (≈70), supplemented by citation tracking (≈40), for an initial yield of ≈890 records (≈850 database records plus ≈40 citation-tracking records) across 2011–May 2026 [10]. The Boolean query combined pandemic terminology (*COVID-19, coronavirus, pandemic, infodemic*) with misinformation terms (*fake news, misinformation, disinformation, rumour*) and detection methodology terms (*machine learning, deep learning, transformer, BERT, GNN, NLP*).

After deduplication via Zotero (≈450 unique records), title and abstract screening retained only papers from SCIE/Scopus journals (IF ≥ 2.0) or CORE A/A* conferences (≈150 excluded; ≈300 advanced to full-text). Full-text assessment applying four documented exclusion criteria (non-automated methods: $n \approx 25$; out-of-scope policy/spread analysis: $n \approx 20$; insufficient methodology detail: $n \approx 18$; low-quality venues: $n \approx 16$) (≈79 exclusions in total) yielded a final corpus of 221 publications. Nine additional 2025–2026 papers identified through forward citation tracking are cited as supporting literature but excluded from the formal corpus. The four-phase screening process is depicted in the PRISMA selection flow of Fig. 5.2(b); Table 5.1 presents the thematic breakdown.

Table 5.1: Thematic Breakdown of the 221-Paper Corpus [10]

Category	**2011–2023**	**2024–2026**
Content-based methods	73	14
Knowledge-based	18	3
Style-based	22	1
Latent features (transformers)	33	10
Social context-based	42	6
Propagation-based (GNNs)	28	5
Source-based	14	1
Hybrid approaches	32	18
COVID-19 specific dataset/resource papers[‡]	21	4
Novel paradigms	–	11
RAG-based detection	–	8
Federated learning	–	3
Total	**168**	**53**

[‡]25 papers; 23 of these are evaluation studies on which the Constraint-dataset count in Section 5.2.3 is based.

Bibliometric summary. Table 5.2 condenses the descriptive statistics of the 221-paper corpus. Publication volume peaks in 2021 (50 papers), one year after the WHO pandemic declaration, and remains high in 2020 (37), 2019 (33), and 2025 (32), the last reflecting the LLM and RAG wave. The methodological mix shifts markedly

between the two periods: hybrid approaches grow from 19.0% of the 2011–2023 papers (32 of 168) to 34.0% of the 2024–2026 papers (18 of 53), whereas content-based work falls from 43.5% to 26.4% and social-context work from 25.0% to 11.3%. Only 25 papers (11.3%) are COVID-19-specific dataset or resource papers, and the corpus is overwhelmingly English-language, which is consistent with the benchmark monoculture and language skew discussed in Section 5.2.3.

Table 5.2: Bibliometric summary of the 221-paper corpus [10]

Dimension	**Value**	**Detail**
Records identified	≈890	≈850 from six databases plus ≈40 from citation tracking
After deduplication / full-text stage / final corpus	≈450 / ≈300 / 221	79 full-text exclusions (non-automated 25, out of scope 20, insufficient detail 18, low-quality venues 16)
Period split	168 / 53	2011–2023 (76.0%) versus 2024–2026 (24.0%)
Peak publication years	2021: 50	2020: 37; 2019: 33; 2025: 32
Content-based methods	87 (39.4%)	73 (43.5%) in 2011–2023; 14 (26.4%) in 2024–2026
Social context-based methods	48 (21.7%)	42 (25.0%) in 2011–2023; 6 (11.3%) in 2024–2026
Hybrid approaches	50 (22.6%)	32 (19.0%) in 2011–2023; 18 (34.0%) in 2024–2026
COVID-19-specific dataset/resource papers	25 (11.3%)	21 in 2011–2023; 4 in 2024–2026
Novel paradigms (RAG, federated)	11 (5.0%)	8 RAG-based; 3 federated, all 2024–2026
Quality filter and language	—	SCIE/Scopus IF ≥ 2.0 or CORE A/A* venues; overwhelmingly English

5.2.3 Datasets and Knowledge Resources

Table 5.3 characterises the principal COVID-19 fake news datasets covered in the reviewed literature. The Constraint dataset [3] dominates evaluation, appearing in 14 of 23 COVID-19-specific evaluation studies (the 25 papers of the COVID-19-specific category in Table 5.1 include two that are not evaluation studies), a benchmark monoculture that creates systematic publication bias toward text-only English-language methods.

Table 5.3: Principal COVID-19 Fake News Benchmark Datasets Covered in the Review [10]

Dataset	Size	Modality	Language	Key Result / Note
Constraint [3]	10,700 (full)	Text	EN	CT-BERT: 99.02% Acc; appears in 14/23 studies
CoAID	–	Text	EN	COVID-19 healthcare misinformation
FakeCovid	–	Text	EN	Multi-dataset fusion benchmark
MM-COVID	–	Multimodal	Multi.	dEFEND: 91% Acc on this corpus
ReCOVery [119]	1,445	Multimodal	EN	SAFE F1=83.6% vs Text-CNN F1=76.6%
GeoCoV19 [293]	524M tweets	Text	62 lang.	Unlabelled; enables multilingual social-context work
CTF COVID-19 Twitter	–	Text	EN	Cross-SEAN: 95.4% Acc, 95.3% F1
MediaEval 2020	–	Propagation	EN	GNN structural features: F1=95%
Koirala [64]	–	Text	EN	Evolutionary feature selection experiments

Three structured knowledge resources support evidence-grounded COVID-19 detection: the CIDO (COVID-19 Infectious Disease Ontology) with 7,834 classes enabling Description Logic reasoning over pandemic claims; CORD-19-on-FHIR with >3.2 million clinical instances supporting medical evidence retrieval; and Wikidata used for entity extraction in the DEAP-FAKED knowledge graph integration system [10].

5.2.4 Review of COVID-19 Fake News Detection Approaches

5.2.4.1 Taxonomy of Detection Strategies

Fig. 2.1 (Chapter 2) presents the three-tier taxonomy organising the reviewed literature: content-based strategies (knowledge-based, style-based, latent feature-based); social context-based strategies (propagation-based, source-based); and hybrid strategies combining signals from both tiers. The 2024–2026 sub-corpus introduces a fourth tier of emerging paradigms discussed in Section 5.2.4.5.

5.2.4.2 Content-Based Approaches

Fig. 2.2 (Chapter 2) illustrates the general workflow for content-based detection across the three sub-paradigms.

Knowledge-based methods ground veracity assessment in structured external knowledge. FACE-KEG [262] extracts Subject-Predicate-Object triples from claims and queries a Wikidata knowledge graph, with an ELMo + BiLSTM classifier over the resulting evidence subgraph. For COVID-19-specific work, the CIDO (COVID-19 Infectious Disease Ontology) [354, 355] with 7,834 classes enables Description Logic reasoning over pandemic claims, and CORD-19-on-FHIR provides access to peer-reviewed clinical evidence at inference time [10]. Banik et al. [356] and Zarei et al. [357] compiled early COVID-19 misinformation datasets and evaluated knowledge-assisted detection. Shahi et al. [358] provided further COVID-19 FND dataset resources. The WHO's characterisation of the infodemic [353] establishes the public health stakes that knowledge-grounded detection must address.

Style-based methods exploit statistical regularities in the surface form of misinformation: word frequency distributions, syntactic feature profiles, psychological word category loadings via LIWC, and readability indices. An SVM trained on stylometric features achieves 93.46% accuracy on Constraint [10], while evolutionary feature selection methods combining whale optimisation and particle swarm algorithms with k-NN classifiers improve further on the Koirala dataset [64]. Pranesh et al. [359] applied stylometric features to COVID-19 social media misinformation. Groza et al. [360] studied surface-feature-based detection of COVID-19 health misinformation. Memon and Carley [361] characterised COVID-19 content on social platforms using stylometric profiling. Fondazione Bruno Kessler [362] examined linguistic patterns in COVID-19 misinformation collections.

Latent feature-based methods dominate benchmark performance. CT-BERT [58], pre-trained specifically on COVID-19 tweets, achieves 99.02% accuracy on Constraint in its highest-performing fusion configuration [363]. Transformer ensembles combining BERT, ALBERT, and XLNet consistently exceed 98% [69]. Additional transformer-based COVID-19 FND systems include Chen et al. [62], Wani et al. [57], Das et al. [364], Shifath et al. [66], Gundapu and Mamidi [67], Li et al. [365], Al-Ahmad et al. [65], Koloski et al. [332], and Vijjali et al. [63]. Two-stage transformer approaches for COVID FND are described in Vijjali et al. [63]. Kotonya et al. [366] provides a recent Check-COVID study. The Chapter 3 contribution Three-Label CT-BERT employs CT-BERT in a three-label framework, achieving 93% accuracy by explicitly modelling borderline uncertainty rather than forcing binary verdicts [11].

5.2.4.3 Social Context-Based Approaches

GNNs applied to retweet propagation graphs yield striking results on the MediaEval 2020 benchmark: structural features alone achieve F1 of 95% [367], compared to 65.33% for a content-based ensemble. This 30 percentage-point advantage confirms that propagation topology is a stronger veracity signal than content during high-volume epidemic conditions. Community characterisation approaches model the social graph

of COVID-19 misinformation communities on Twitter, identifying echo chamber properties and coordinated amplification signatures. Source-based methods exploit account-level signals, including posting age, follower ratios, verification status, and coordinated behaviour patterns, but are systematically compromised in epidemic contexts by rapid account creation for infodemic amplification [10].

5.2.4.4 Hybrid Approaches

Cross-SEAN [261], a cross-stitch semi-supervised neural attention model, achieves 95.4% accuracy and 95.3% F1 on the CTF COVID-19 Twitter dataset by jointly attending to content and social context. dEFEND [121] achieves 91% accuracy on MM-COVID by combining news content with user comment propagation in an explainable co-attention framework. SAFE achieves F1 of 83.6% on the ReCOVery repository [119] through cross-modal semantic alignment of text and image representations, compared to 76.6% for a Text-CNN unimodal baseline. Multi-dataset fusion across Constraint, FakeCovid, and CoAID yields further performance increments [363]. Table 5.4 consolidates representative results.

Table 5.4: Representative Performance Results on COVID-19 Benchmark Datasets [10]

Model / Approach	**Dataset**	**Metric**	**Score**
CT-BERT fusion (best) [363]	Constraint	Accuracy	99.02%
Transformer ensemble (B+A+X) [69]	Constraint	Accuracy	>98%
BERT baseline [57]	Constraint	Accuracy	above CNN/LSTM/HAN
SVM (stylometric)	Constraint	Accuracy	93.46%
Cross-SEAN [261]	CTF COVID-19	Accuracy	95.4%
Cross-SEAN [261]	CTF COVID-19	F1	95.3%
GNN (structural only) [367]	MediaEval 2020	F1	95%
Content ensemble [367]	MediaEval 2020	F1	65.33%
dEFEND [121]	MM-COVID	Accuracy	91%
SAFE [119]	ReCOVery	F1	83.6%
Text-CNN [119]	ReCOVery	F1	76.6%

5.2.4.5 Emerging Paradigms (2024–2026)

The 2024–2026 sub-corpus (n = 53) introduces qualitatively new detection paradigms not represented in the prior literature (Table 5.5).

Table 5.5: Emerging Detection Paradigms in the 2024–2026 Sub-corpus [10]

Paradigm	**Description**	**Representative Systems**
RAG-based detection	Grounds classification in dynamically updated knowledge at inference time, addressing both temporal drift and domain shift	RADAR [219], EvolveDetector [368]
Federated learning	Decentralised training without raw data exfiltration; compatible with encrypted platform constraints	Machová et al. [369]
LLM-as-detector	Uses LLMs for implicit characteristic analysis and adaptive boosting via zero-shot or few-shot prompting	Raza et al. [370]
Agentic multimodal	Orchestrates retrieval, reasoning, and cross-modal verification as coordinated AI agents	MIRAGE framework

5.2.5 Generative AI as a New Threat Vector for Epidemic Misinformation

The emergence of capable LLMs is the largest structural shift in fake news detection since the transformer. Schuster et al. [213] showed that machine-generated and human-generated text now occupy overlapping regions in stylistic feature space (Fig. 5.3), implying that stylometry-based detection has reached a structural ceiling irrespective of further improvements in classifier architecture.

Three properties make this failure mode uniquely dangerous in health contexts [10]: (i) Scalable personalisation: LLMs can generate thousands of demographically targeted variants of a single false claim (pregnancy risk, athletic performance, paediatric safety), defeating claim-matching defences and complicating training corpus construction. (ii) Hallucinated epistemic authority: fabricated medical mechanisms described with clinical jargon are treated as expert content by downstream sharers, even when the underlying claim is verifiably false. (iii) Short feedback loop to physical harm: unlike political misinformation that shapes electoral behaviour over months, health misinformation directly influences treatment decisions and vaccine uptake within hours of first exposure.

Fig. 2.4 (Chapter 2) illustrates the types of textual perturbations that LLM-based adversarial systems can generate.

The root cause is *shortcut learning* [10]: when a classifier's prediction flips after substituting the word *lethal* with *deadly*, the model is revealed to have learned statistical surface-form correlates of veracity rather than reasoning about the medical claim itself. Standard transformer fine-tuning on benchmark datasets actively encourages this failure

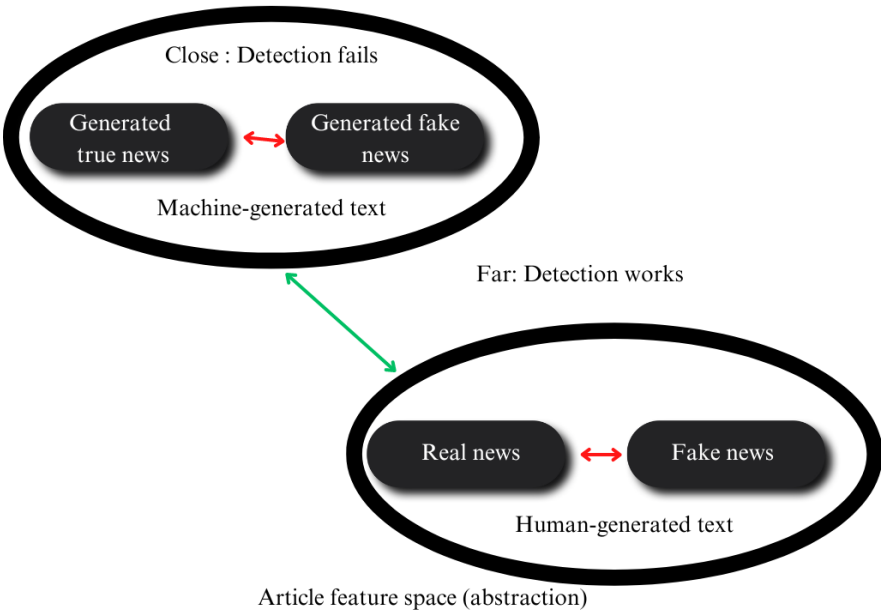


Figure 5.3: Conceptual feature space showing proximity between machine-generated and human-generated true and fake news. When LLM-generated misinformation and authentic clinical text occupy overlapping regions, no stylometric classifier can maintain a reliable decision boundary [213, 10].

mode by rewarding surface-pattern memorisation. The work of Chapter 4 [16, 15] provides independent empirical confirmation: DeepSeek-LLM semantic inversions outperform conventional text-attack methods across five multimodal FND architectures, with statistical confirmation at $p < 0.05$ via Wilcoxon and paired t-tests respectively.

Table 5.6 codifies the detection paradigm shifts required to address the GenAI threat.

5.2.6 Proposed Framework for Epidemic-Aware Fake News Detection

In response to the identified limitations, a tiered epidemic-aware detection architecture is proposed [10]. At $t = 0$ (publication time), a lightweight content-only classifier operates during the *golden window* before misinformation reaches viral threshold, accepting lower precision to maximise recall. As propagation data accumulates, a second-tier RAG-augmented pipeline grounds veracity in a dynamically updated medical knowledge base with social context features incorporated once the diffusion graph is sufficiently dense. The RAG layer simultaneously addresses temporal veracity drift: classification grounded in current knowledge reflects updated scientific consensus without parameter retraining, a mechanism instantiated by Graph-RAG and, in design but without a demonstrated accuracy gain, by the external verification component of EAKD-CFND (Chapter 3).

The three-class veracity schema (*true / false / uncertain*) is proposed as a replacement for binary classification in pandemic contexts, explicitly representing the

Table 5.6: Required Detection Paradigm Shifts for the Generative AI Era [10]

Current Paradigm (From)	**Required Paradigm (To)**
Text style analysis	Claim verifiability analysis
Stylometric pattern matching	Real-time integration with curated medical knowledge bases
Document-level classification	Document-level semantic consistency analysis
Content-only features	Source-pedigree signals (provenance metadata, posting velocity, human accountability traces)
Single-architecture classifiers	Ensemble strategies combining BERT-like encoders (clean-data classification) with autoregressive LLMs (perturbation robustness via generative verification)

epistemic register of contested or rapidly evolving health claims rather than forcing premature veracity assignment to content where the scientific consensus is itself in flux [10].

5.2.7 Results and Key Findings

The structured synthesis of 221 papers yields five principal empirical findings [10]: (i) *Benchmark saturation*: CT-BERT achieves 99.02% on Constraint, rendering this benchmark incapable of differentiating architectures. (ii) *Cross-dataset brittleness*: models exceeding 98% on Constraint show lower performance on MM-COVID, ReCOVery, and MediaEval 2020, confirming that benchmark monoculture masks generalisation failure. (iii) *Benchmark monoculture*: Constraint appears in 14 of 23 COVID-19-specific studies, biasing the literature toward text-only English-language methods. (iv) *Social context superiority*: GNNs using only structural propagation features outperform content ensembles by 30 percentage points on MediaEval 2020, attributable to the unusually dense and rapid propagation characteristic of pandemic misinformation. (v) *Multimodal integration gap*: despite SAFE's 7-percentage-point advantage over Text-CNN on ReCOVery, multimodal COVID-19 datasets remain underrepresented, and the Unimodal Bias problem documented in Chapter 4 applies directly.

5.2.8 Eleven Critical Research Gaps

Table 5.8 presents the eleven critical research gaps identified through the structured synthesis, each with the core problem and the primary future direction identified [10]. These eleven gaps are the COVID-19-specific refinement of the seven thesis-level technical gaps introduced in Chapter 2 (Table 2.4), specialising that general agenda to the pandemic-deployment context rather than constituting a separate or contradictory taxonomy. Table 5.8 also separates two questions: whether the gap was already partly

addressed in the surveyed literature or is an entirely novel formulation, and how far the thesis contributions of Chapters 3–4 close it. Seven gaps (G1, G2, G4, G5, G6, G7, G8) are addressed and four (G3, G9, G10, G11) remain open directions; G3, G10, and G11 are entirely novel formulations of this review.

5.2.9 Lessons for Future Epidemic Preparedness

The May 2026 Andes hantavirus outbreak [371, 372], linked to the MV Hondius cruise ship, three deaths, and passengers from 23 countries, provides a contemporary illustration of the potential generalisability of COVID-19 lessons. Within 48 hours, "Is hantavirus the next COVID-19?" was circulating across platforms before the WHO had issued formal risk guidance (WHO Disease Outbreak News DON600; CDC Health Alert Network HAN-00528). Three of the eleven research gaps applied immediately: G1 (temporal veracity drift: case counts evolving hourly); G3 (early detection paradox: no propagation graph at first report); and G10 (dataset infrastructure and language skew: 23 nationalities receiving simultaneous misinformation). A fourth gap, not among the eleven, also applied: a knowledge base gap (no hantavirus ontology comparable to CIDO [354] exists). The potential generalisability of COVID-19 detection lessons to future epidemic infodemics is thus not purely hypothetical but illustrated in conditions contemporaneous with this thesis [10].

5.2.10 Explainability in Pandemic Fake News Detection

Fig. 5.4 illustrates the three-dimensional explainability framework applicable to FND systems in pandemic contexts: algorithmic interpretability (how the model arrived at its decision), human interpretability (why the decision makes sense to a domain expert), and supporting evidence (what external information corroborates the verdict). The progression from Three-Label CT-BERT (LIME attribution) through FakeNews Buster (vector retrieval evidence) to Graph-RAG (graph evidence chains) and EAKD-CFND (an inspectable, though poorly calibrated, uncertainty trigger) across this thesis constitutes a systematic traversal of all three dimensions.

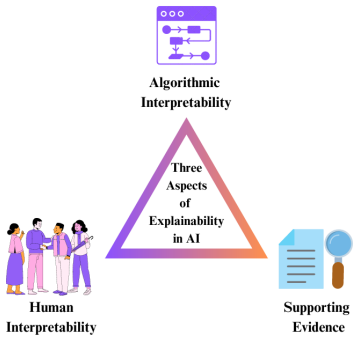


Figure 5.4: Three dimensions of explainability in AI systems relevant to fake news detection: algorithmic interpretability, human interpretability, and supporting evidence. Public health deployment requires all three dimensions to be addressed simultaneously [10].

5.2.11 Limitations of the Review

Seven limitations are acknowledged [10]: (i) no meta-analysis of performance data due to heterogeneous evaluation conditions; (ii) English-language bias in the reviewed literature itself; (iii) the PRISMA snapshot covers only to May 2026; (iv) non-automated approaches are excluded; (v) high reported accuracies (99.02% on Constraint) reflect in-distribution performance only; (vi) high-impact misinformation on WhatsApp, Telegram, and Signal is structurally underrepresented; and (vii) nine supplementary 2025–2026 papers identified through citation tracking are cited as supporting literature but excluded from the formal 221-paper corpus.

5.2.12 Cross-Domain Generalisability of the Review Findings

The systematic review is anchored in COVID-19, so the extent to which its findings carry over to other misinformation domains (elections, climate, finance, other epidemics) needs to be stated. Findings that rest on properties of the *method class* or of the *evaluation practice* are likely to transfer, whereas findings that rest on COVID-specific *resources* are not. Benchmark saturation and monoculture, the stylometry ceiling imposed by LLM-generated text, and the early-detection paradox of propagation-based detectors follow from how detectors and benchmarks are built and used, not from health content, and the thesis observes the same mechanisms outside the pandemic setting: the Clever Hans watermark shortcut in the political-fact-checking D1 dataset, dataset-dependent modality preference on PolitiFact (Chapter 4), and the sharp drop of exemplar-free continual learning on the event-ordered PHEME sequence (Chapter 3). The 2026 hantavirus illustration (Section 5.2.9) shows the same three gaps (temporal drift, early detection, and language coverage) in a second epidemic. By contrast, the domain resources catalogued in the review (the CIDO ontology, CORD-19-based knowledge graphs, and GeoCoV19) are specific to health, and a deployment in another domain must supply its own knowledge base; the Graph-RAG results on FEVER indicate that the retrieval approach itself does transfer wherever an evidence graph exists.

Two limits on generalisation remain. First, the corpus is overwhelmingly English-language and text-centred, so conclusions about low-resource languages and about image- or video-borne misinformation rest on the multimodal studies of Chapter 4 rather than on the review. Second, social-context superiority depends on platform data access, which differs sharply between open platforms and encrypted messaging services. Table 5.7 summarises these judgements.

Table 5.7: Likely transferability of the main PRISMA findings to other misinformation domains

Finding	**Transfers?**	**Basis**
Benchmark saturation and monoculture	Yes	Property of evaluation practice; D1 watermark shortcut shows the same artefact effect outside COVID-19
Stylometry ceiling under LLM-generated text	Yes	Property of the detector class; DeepSeek and Llama3 attacks degrade detectors on political datasets
Early-detection paradox	Yes	Propagation graphs need edges in any domain; hantavirus case shows it recurs
Temporal veracity drift	Partly	Strongest where scientific consensus evolves; event-driven drift appears in PHEME
Social-context superiority of GNNs	Conditional	Depends on platform data access; weak under encrypted messaging
Domain resources (CIDO, CORD-19, GeoCoV19)	No	Health-specific; other domains need their own knowledge bases

Table 5.8: Eleven Critical Research Gaps in Pandemic Fake News Detection, with Novelty and Status in This Thesis [10]

#	**Gap**	**Core Problem**	**Key Future Direction**	**Prior work**	**This thesis**
G1	Temporal Veracity Drift	Static models misclassify evolving scientific consensus	RAG + continual learning + three-class schema	Partially addressed	Addressed
G2	Neural Fake News	LLMs produce stylometrically authentic misinformation	Shift to claim verifiability; RAG with medical KBs	Partially addressed	Addressed
G3	Early Detection Robustness	Golden window closes before propagation graphs form	Tiered $t = 0$ content classifier + RAG pipeline	Entirely novel framing	Open (framework proposed)

continued on next page

Table 5.8 (continued)

#	Gap	Core Problem	Key Future Direction	Prior work	This thesis
G4	Adversarial Robustness	All SOTA detectors vulnerable to synonym perturbations	RAG grounding; social context; adversarial training	Partially addressed	Addressed
G5	Scalability	Transformer quadratic complexity creates backlogs	Sparse attention; knowledge distillation	Partially addressed	Addressed
G6	Multimodal Integration	Over-reliance on one modality (Unimodal Bias)	Cross-modal consistency; image provenance verification	Partially addressed	Addressed (diagnosis and metric)
G7	Cross-Domain Generalisation	Political FND models fail on health/pandemic domain	Few-shot; meta-learning; RAG as domain adaptor	Partially addressed	Addressed
G8	Explainability	Opacity impedes deployment by public health authorities	Intrinsically explainable architectures; RAG evidence	Partially addressed	Addressed
G9	Dynamic Propagation Modelling	Static snapshots miss temporal adversarial signals	Dynamic GNNs; temporal graph networks	Partially addressed	Open
G10	Dataset Infrastructure	Labelling bottleneck; walled-garden platforms; English skew	Federated learning; multilingual multimodal datasets	Entirely novel for pandemics	Open
G11	Stance-Based Detection	Binary veracity misses coordinated health-authority opposition	Stance as first-class feature; annotated datasets	Entirely novel framing	Open

5.3 Comprehensive Comparative Evaluation of FND Frameworks

5.3.1 Motivation and Evaluation Protocol

This section provides a comprehensive unified analysis of all eight contributions from the thesis, addressing RO 4. The contributions span three orthogonal dimensions: (i) *method paradigm*: text-based (Three-Label CT-BERT, FakeNews Buster, Graph-RAG, EAKD-CFND), multimodal (the three multimodal studies), or domain-specific (COVID-19 Infodemic Review); (ii) *robustness concern*: standard supervised learning (the text-based contributions), adversarial evaluation (Adversarial Robustness, LLM-Driven Perturbations), bias quantification (Comprehensive Multimodal Evaluation), or real-world domain shift (COVID-19 Infodemic Review); and (iii) *deployment consideration*: live-stream performance, long-horizon stability, or operational uncertainty.

Subsection 5.3.3 presents detailed numerical comparison of the four text-based FND contributions from Chapter 3, each addressing a distinct limitation: forced binary classification (Three-Label CT-BERT); concept drift under continual learning (FakeNews Buster); multi-hop evidence retrieval (Graph-RAG); and adaptive knowledge distillation for long-horizon deployment without exemplar memory (EAKD-CFND). This is followed by summary analysis connecting all eight papers across uncertainty, robustness, and generalisability dimensions.

Four primary evaluation dimensions are applied to the text-based methods: (i) *classification performance*: accuracy and F1-score on benchmarks; (ii) *uncertainty handling*: whether the method expresses borderline classification uncertainty as a first-class output; (iii) *concept drift resilience*: whether the method updates its knowledge representation without full retraining; and (iv) *explainability provision*: whether human-interpretable evidence accompanies verdicts. Cross-framework robustness assessment against all 8 papers follows in the unified discussion.

5.3.2 Method Descriptions

Three-Label CT-BERT [11] extends CT-BERT with a *Can't Say* output for claims whose posterior probability for either class falls within the [0.45, 0.55] confidence window, coupled with dynamic dataset enrichment via the Google Fact Check API. Full details: Section 3.3.

FakeNews Buster [12] is a dual-encoder (BERT + FastFormer) continual learning architecture with an adaptive loss:

$$L = \frac{1}{\sqrt{\text{seen}+1}} L_{CE} + \left(1 - \frac{1}{\sqrt{\text{seen}+1}}\right) L_{KD}$$

that dynamically rebalances plasticity and stability across sequential tasks. Full details: Section 3.4.

Graph-RAG Classification Framework [14] expands claim representations through multi-hop knowledge-graph traversal with Graph Attention Networks before RoBERTa classification (DistilBERT in the Fast Graph-RAG variant), overcoming the retrieval-relevance mismatch of standard single-hop RAG. Full details: Section 3.5.

EAKD-CFND [13] computes per-instance uncertainty $\omega(x)$ from the teacher's softmax entropy and scales KD loss as $\alpha(\omega(x)) = \max(0, 1 - \beta \cdot \text{norm}(\omega(x)))$. When $\omega(x) > \theta_{uncertainty} = 0.7$, an external API is queried to provide a verified training label (falling back to the original label when no usable match is found). The mechanism is offered as an exemplar-free default and as a rehearsal-augmented configuration (EAKD-CFND-R) with a small reservoir buffer. Full details: Section 3.6.

5.3.3 Performance Comparison

Table 5.9 presents the master comparison. Direct numerical comparison is partially constrained by different primary evaluation datasets across contributions (Three-Label CT-BERT, FakeNews Buster: Constraint; Graph-RAG: FEVER, LIAR, FNN; EAKD-CFND: PHEME-Event, FNN-Poli-Time, FNN-Gossip-Time). The table presents each method on its primary benchmark, with a footnote noting where dataset overlap permits direct comparison.

Table 5.9: Master Performance Comparison: Text-Based FND Contributions and SOTA Baselines

Method	**Dataset**	**Acc.(%)**	**F1**	**BWT(%)**	**Expl.**
SOTA Baselines					
BERT [57]	Constraint	—	0.823	—	No
CT-BERT+RoBERTa [62]	Constraint	—	0.887	—	No
XLNet+LDA [69]	Constraint	—	0.871	—	No
Fine-tuning	FNN-Poli-Time	27.4	0.071	−80.5	No
DER (buf=200) [92]	FNN-Poli-Time	30.4	0.099	−80.1	No
DER (buf=200) [92]	PHEME-Event	**84.8**	0.165	**−3.2**	No
Thesis Contributions					
Three-Label CT-BERT [11]	COVID-19 Constr.	**93.0**	—	—	Partial
FakeNews Buster [12]	Constraint	—	**0.962**	—†	Yes
Graph-RAG [14]	FEVER	—	0.755	—	Yes
EAKD-CFND [13] (exemplar-free)	FNN-Poli-Time	29.8	0.090	−81.6	Partial
EAKD-CFND [13] (exemplar-free)	PHEME-Event	16.8	0.033	−87.2	Partial

continued on next page

Table 5.9 (continued)

Method	Dataset	Acc.(%)	F1	BWT(%)	Expl.
EAKD-CFND-R [13] (rehearsal)	FNN-Poli-Time	**34.9**	0.117	−72.0	Partial
EAKD-CFND-R [13] (rehearsal)	PHEME-Event	**85.5**	0.167	**−1.8**	Partial

†FakeNews Buster reports a forgetting rate of 0.06 (≈6%) under a 2-task continual-learning protocol, which is not directly comparable to the BWT (%) reported for EAKD-CFND under the class-incremental protocol. Accuracy and BWT rows are comparable only within a dataset: the FNN-Poli-Time rows (27.4–34.9%) and PHEME-Event rows (16.8–85.5%) come from different task sequences, and EAKD-CFND's FNN-Poli-Time result is statistically tied with fine-tuning and DER (only the +PP and +PP+UW rehearsal variants separate). FNN-Gossip-Time rows are omitted for space: DER (29.4%), EAKD-CFND (29.1%), and EAKD-CFND-R (29.2%) are statistically tied there (Appendix III.2). The F1 column for continual-learning rows is the average macro-F1 over all tasks after the final training step and is not comparable with the single-task F1 of the other rows.

Table 5.10 places all eight contributions side by side on common dimensions (paradigm and modality, benchmarks, headline result, principal robustness finding, and explainability or cost), so that the strengths and the limits of each method can be read across the thesis in one view. The headline results are drawn from the primary benchmark of each paper and are therefore not directly comparable across rows.

Table 5.10: Unified cross-method summary of the eight thesis contributions (P1–P8)

Paper	Paradigm; modality; benchmarks	Headline result	Principal robustness / limitation finding	Explainability; cost
P1 Three-Label CT-BERT	Uncertainty-aware classification; text; Constraint 2021	93.0% accuracy, +13 pp over binary BERT	Borderline claims routed to *Can't Say* (band 0.45–0.55)	LIME; Fact Check API evidence
P2 FakeNews Buster	Continual learning with adaptive KD; text; Constraint 2021	F1 0.9616 (single task); continual F1 0.92, forgetting 0.06	Two-task protocol only; adaptive KD is the dominant component	LIME + vector evidence; FastFormer within 0.72 pp of BERT

continued on next page

Table 5.10 (continued)

Paper	Paradigm; modality; benchmarks	Headline result	Principal robustness / limitation finding	Explainability; cost
P3 Graph-RAG	Retrieval-augmented classification; text; LIAR2, FNN, FEVER	FEVER macro-F1 0.755 (+3.1 over single-hop)	Gain only on evidence-chain tasks; task structure sets the ceiling	Evidence graph chains; Fast Graph-RAG 9.2 ms/claim
P4 Cluster Computing	VLM robustness; multimodal; D1, RECOVERY, PolitiFact	ViLBERT Text-Only F1 0.748 (best resilience)	BERT-CLIP collapses (F1 0.104 Text-Only); black-box transfer across architectures	Grad-CAM; heavier VLM backbones
P5 Comp. Elec. Eng.	LLM-driven adversarial stress test; multimodal; D1, RECOVERY, PolitiFact	DS-Sem lowers CroMe F1 from 0.99 to 0.82 on D1	DeepSeek attacks stronger than Llama3 ($p < 0.05$); modality preference reverses on PolitiFact	Attention shifts under attack; SOTA models costly
P6 AJSE	Unimodal Bias and Synergy Gain; multimodal; D1, RECOVERY	Baseline F1 up to 0.98	BertVgg Text-Only F1 0.00; Clever Hans watermark shortcut	Grad-CAM diagnosis
P7 EAKD-CFND	Exemplar-free continual learning; text; PHEME-Event, FNN-Poli-Time, FNN-Gossip-Time	EAKD-CFND-R 85.5% on PHEME-Event; exemplar-free ties baselines on FNN	Exemplar-free collapses on PHEME-Event (16.8% vs 84.8%); ECE 0.80–0.84	Inspectable verification trigger; training 3.5× fine-tuning
P8 COVID-19 review	PRISMA review; text and multimodal literature; 221 papers (2011–2026)	CT-BERT 99.02% on Constraint (benchmark saturation)	Benchmark monoculture; eleven research gaps	Review of XAI practice

5.3.4 Analysis: Uncertainty Handling

The four contributions represent three approaches to classification uncertainty. Three-Label CT-BERT provides the most *explicit* uncertainty representation: a dedicated *Can't Say* output class with defined probability thresholds. This is architecturally honest

but operationally deferring: *Can't Say* instances require subsequent API resolution, which may fail on novel or uncatalogued claims [11].

FakeNews Buster handles uncertainty *implicitly* through instance-level difficulty weighting in the adaptive loss: items near the decision boundary receive higher distillation penalties, preserving knowledge about difficult cases across tasks. However, uncertainty is never surfaced to the end-user; a binary verdict is always produced with no borderline confidence signal [12].

EAKD-CFND uses the most elaborate uncertainty mechanism of the four: the entropy-based uncertainty estimator $\omega(x)$ triggers an external knowledge query when $\omega(x) > 0.7$ and substitutes a verified label where one is found. The measured behaviour is more modest than the design: only about 10% of triggered queries returned a usable verified label, the signal's calibration error is high (ECE 0.80–0.84), and ablations found no accuracy contribution from verification outside the noise floor [13]. The mechanism makes the trigger inspectable but does not by itself demonstrate improved accuracy.

5.3.5 Analysis: Concept Drift Resilience

Three-Label CT-BERT is a static model: parameters are fixed post-training and performance degrades as the misinformation distribution evolves. The API enrichment loop provides partial data-level mitigation but does not update the CT-BERT encoder itself.

Graph-RAG achieves implicit drift resilience through *retrieval-as-adaptation*: classification is grounded in a knowledge base that is updated independently of model parameters, so verdicts reflect current verified information without retraining. This makes Graph-RAG particularly suited to epidemic contexts where factual knowledge evolves faster than labelled datasets can be constructed [14].

FakeNews Buster and EAKD-CFND provide the most principled *parameter-level* drift resilience. FakeNews Buster's adaptive loss reduces forgetting by 60% relative to naive fine-tuning (forgetting rate 0.06 vs 0.15) [12]. EAKD-CFND's exemplar-free configuration is statistically indistinguishable from rehearsal-based DER (and from fine-tuning) on both FakeNewsNet splits and collapses on the five-task PHEME-Event sequence (16.8% versus DER's 84.8%, BWT −87.2% versus −3.2%); external verification therefore does not substitute for stored past-data memory under severe drift. Adding a small reservoir buffer (EAKD-CFND-R) recovers 85.5% on PHEME-Event, and the task-similarity analysis is consistent with the view that drift severity determines whether rehearsal is worth its storage cost [13].

5.3.6 Analysis: Explainability and Trust

The four contributions span the post-hoc to intrinsic explainability spectrum. Three-Label CT-BERT applies LIME for token-level attribution supplemented by API evidence, a post-hoc approximation of a locally linear model surface. FakeNews Buster adds a vector similarity search module retrieving the most similar fact-checked claims from a curated evidence database, providing context-aware evidence alongside LIME attribution [12].

Graph-RAG offers intrinsic, evidence-based explainability: the retrieved passages, together with the attention weights over them, form a human-readable evidence trail showing which facts supported the verdict. This grounds the classification directly rather than approximating it after the fact, though the property is a design feature of the approach rather than something this thesis evaluates quantitatively [14]. EAKD-CFND augments this with a per-instance uncertainty signal and an inspectable verification trigger. The signal is, however, poorly calibrated in absolute terms (ECE 0.80–0.84), so it supports relative ranking and not probability statements, and no user study establishes that inspectability produces warranted trust [13].

5.3.7 Synthesis: Deployment Decision Guide

Table 5.11 consolidates the analytical findings as a deployment decision guide. No single method is universally superior; the choice depends on deployment scenario, resource envelope, and institutional requirements.

Fig. 5.5 presents the same guidance as a decision flowchart: a practitioner answers at most four yes/no questions about the deployment and arrives at the recommended method, which is the corresponding row of Table 5.11 (multimodal inputs are routed to the architecture selection guide of Table 4.9).

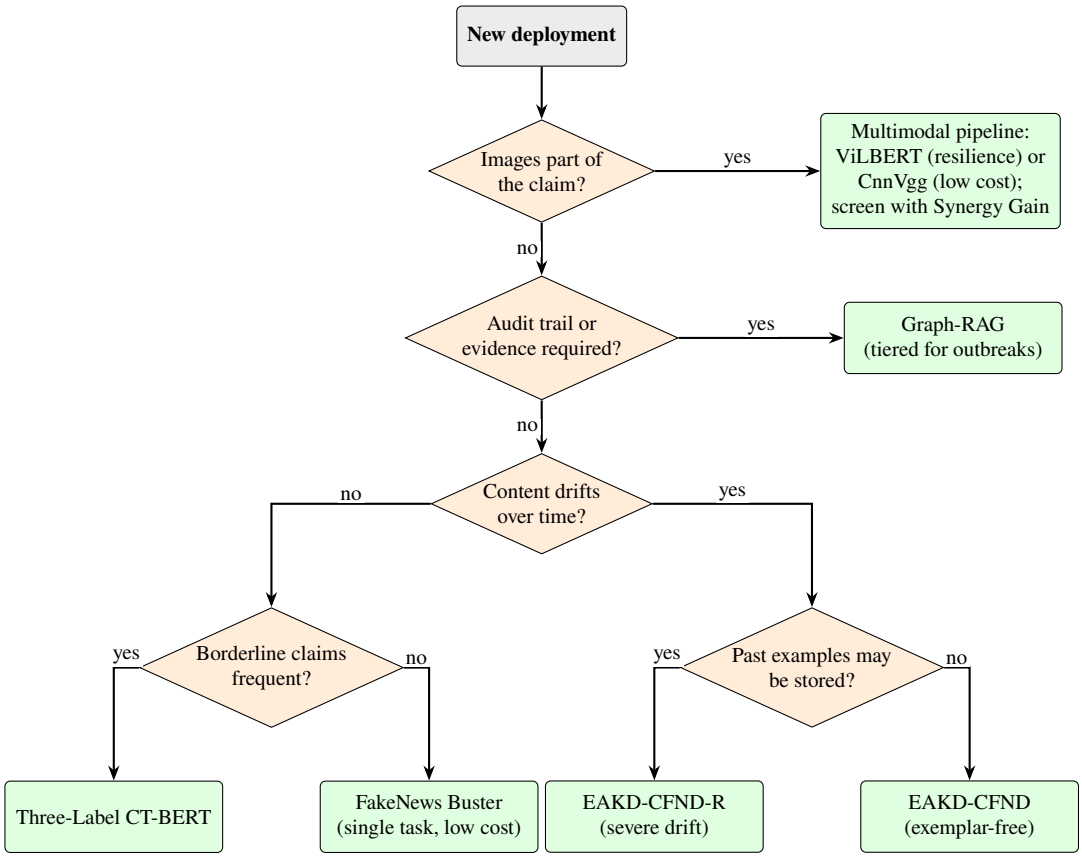


Figure 5.5: Deployment decision flowchart for selecting a fake news detection method from the thesis contributions.

Table 5.11: Deployment Decision Guide: Selecting the Appropriate Text-Based FND Method

Deployment Scenario	**Recommended Method**	**Rationale**
Static, single-domain dataset; high proportion of borderline claims	Three-Label CT-BERT	Explicit uncertainty prevents forced misclassification; most explicit uncertainty handling (93.0% on its COVID-19 benchmark)
Resource-constrained streaming; maximum single-task F1	FakeNews Buster	FastFormer encoder competitive F1 (0.954) at lower cost; forgetting rate 0.06 under its own 2-task protocol
Audit-trail mandated; health, legal, or government domain	Graph-RAG	Intrinsic graph evidence chains; best of the three RAG variants on multi-hop FEVER (macro-F1 0.755); implicit drift via dynamic KB
Long-running deployment; privacy-sensitive, or storing past examples not permitted	EAKD-CFND (exemplar-free)	Statistically tied with fine-tuning and DER on both FakeNewsNet splits (29.8% and 29.1%, BWT about −82%), i.e. no measured retention advantage; collapses on PHEME-Event; retains no past examples
Long-running deployment; storage budget available; severe task drift	EAKD-CFND-R	85.5% on PHEME-Event (matches DER); the +PP and +PP+UW variants are the only configurations significantly above DER on FNN-Poli-Time; no benefit on near-duplicate task sequences
Epidemic outbreak; real-time; evolving knowledge	Graph-RAG (tiered)	Retrieval grounding absorbs new knowledge without retraining; tiered approach covers both $t = 0$ and post-propagation regimes

5.4 Discussion

The COVID-19 infodemic case study and the comparative evaluation converge on four shared findings. The case study identifies, from 221 papers across a decade of pandemic FND research, that the dominant structural limitations are the static veracity assumption, the stylometry ceiling, the early detection paradox, and the explainability gap. The comparative evaluation shows that the thesis contributions address these four limitations to varying and partly unproven degrees.

The *static veracity assumption* is addressed by FakeNews Buster's continual learning and EAKD-CFND's adaptive distillation, both of which are intended to update knowledge representations over time without full retraining (the exemplar-free EAKD-CFND configuration does not demonstrably do so on any of the three splits; only its rehearsal-augmented configuration does so under severe drift). The *stylometry ceiling* is addressed by Graph-RAG and, in design though not demonstrably in accuracy, EAKD-CFND's external verification, both of which ground classification in factual knowledge rather than surface-form pattern matching. The *early detection paradox* is partially addressed by Three-Label CT-BERT's content-only $t = 0$ classification and more completely by the tiered architecture of Section 5.2.6. The *explainability gap* is addressed progressively across Three-Label CT-BERT (LIME + API evidence), FakeNews Buster (vector retrieval evidence), Graph-RAG (intrinsic graph evidence chains), and EAKD-CFND (an inspectable, though poorly calibrated, uncertainty trigger).

The thesis narrative arc across Chapters 3–5 is thus coherent: Chapter 3 showed that binary classification is insufficient and that static models cannot handle concept drift, and evaluated multi-hop retrieval and per-instance adaptive distillation as responses, with mixed results; Chapter 4 revealed the additional problem of unimodal bias in multimodal extension; and Chapter 5 situates the contributions in a real-world pandemic scenario, supporting their practical relevance while showing that they address the identified gaps only in part.

5.5 Summary

This chapter addressed RO 4 through a comparative evaluation and a real-world case study. The COVID-19 infodemic case study (Section5.2) synthesised 221 publications (2011–Maỹ2026), documenting benchmark saturation on Constraint (CT-BERT: 99.02%), GNN superiority under pandemic propagation conditions (+30 percentage points on MediaEval 2020), the structural invalidation of stylometry-based detection by LLMs, and eleven critical research gaps, three of which (with one further gap) recurred in the 2026 hantavirus outbreak.

RO 4 was addressed through a unified four-dimensional evaluation of the four text-based FND contributions from Chapter 3. The analysis established that Three-Label CT-BERT provides the most explicit uncertainty representation, FakeNews Buster the best balance of single-task performance and efficiency, Graph-RAG the preferred architecture for audit-mandated deployments, and EAKD-CFND-R the only configuration matching DER under severe drift. Chapter 6 concludes the thesis with its contributions, limitations, and future directions.

CHAPTER 6

CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT

6.1 Conclusion

This thesis addressed a technically significant and societally consequential problem: the automated detection of fake news in a world where misinformation spreads at platform scale, adversaries employ increasingly sophisticated generation and perturbation techniques, and the scientific consensus on which veracity classifiers depend is itself dynamic. Beginning from a structured review of the field across classical machine learning, deep learning and transformer methods, continual learning, multimodal fusion, adversarial robustness, and retrieval-augmented generation, the work identified the key limitations of existing approaches and translated them into four research objectives, fulfilled through eight research contributions.

In pursuit of an efficient and intelligent text-based detection framework (RO 2), the Three-Label CT-BERT model [11] introduced a probability-threshold *Can't Say* class with dynamic Google Fact Check enrichment, raising accuracy to 93% on the COVID-19 Constraint dataset, a 13 percentage-point gain over binary BERT baselines. FakeNews Buster [12] operationalised dynamic adaptation into a dual-encoder (BERT + FastFormer) continual learning architecture with an adaptive knowledge-distillation loss, achieving an average continual-learning F1 of 0.92 and reducing the forgetting rate to 0.06 (a 60% improvement over naive fine-tuning). The Graph-RAG framework [14] introduced the RAG-for-Classification paradigm and, through a single-seed audit, showed that multi-hop evidence expansion helps specifically on evidence-chain tasks (raising the three-class macro-F1 on FEVER from 0.724 to 0.755) while task structure, not architecture, sets the ceiling; its Fast Graph-RAG variant offers the best accuracy-latency balance for deployment, with human-readable retrieved evidence. EAKD-CFND [13] advanced continual learning to per-instance uncertainty-weighted distillation with external verification; across three class-incremental splits its exemplar-free configuration ties rehearsal-based DER on both FakeNewsNet splits at zero storage but collapses on PHEME-Event (16.8% versus 84.8%), where the rehearsal-augmented EAKD-CFND-R recovers 85.5% accuracy, the uncertainty signal itself proved poorly calibrated (ECE 0.80–0.84), and ablations found no gain from external verification or uncertainty weighting beyond seed noise.

Investigating multimodal detection and its biases (RO 3), the Comprehensive Multimodal Evaluation [17] formalised *Unimodal Bias* and introduced the Synergy Gain metric, documenting the catastrophic collapse of fusion models (F1 = 0.00) when a dominant modality is removed. The LLM-driven stress-testing study [16] confirmed via the Wilcoxon test ($p < 0.05$) that DeepSeek-LLM semantic inversions outperform Llama3 attacks even against state-of-the-art architectures, and the black-box transfer analysis [15] established via a paired t-test ($p < 0.05$) that this vulnerability is systemic, with deep co-attentional fusion (ViLBERT) providing the strongest text-only resilience. The review of fake news detection techniques (RO 1) was delivered by the structured literature review of Chapter 2, spanning classical, deep-learning, transformer, continual-learning, multimodal, adversarial, and retrieval-augmented paradigms. For the comparative objective (RO 4), a unified cross-method evaluation of all contributions produced a practical deployment decision guide across uncertainty handling, drift resilience, explainability, and efficiency, grounded in a COVID-19 infodemic case study [10] that synthesised 221 publications, identified eleven critical research gaps, and illustrated their relevance to the concurrent 2026 hantavirus infodemic.

Collectively, these contributions establish that binary classification is insufficient for the decision-boundary geometry of real misinformation; that per-instance adaptive distillation with external verification did not, in its exemplar-free configuration, measurably reduce forgetting relative to fine-tuning on any of the three class-incremental splits, whereas a small reservoir buffer (EAKD-CFND-R) matched rehearsal-based DER under severe drift; that multi-hop evidence retrieval is a principled and explainable alternative to pattern-based classification on evidence-chain tasks such as FEVER; that multimodal fusion systems systematically collapse under single-modality conditions due to Unimodal Bias; and that LLM-generated adversarial content poses a statistically significant, architecturally generalisable threat to all current detection paradigms. All four research objectives have therefore been fulfilled.

These findings are bounded by limitations acknowledged throughout the thesis. For EAKD-CFND these include poor uncertainty calibration (ECE 0.80–0.84, measured on FNN-Poli-Time only), very small FNN-Poli-Time test sets (31, 65, and 28 examples), an unaudited coverage bias of the fact-checking API, the class-incremental protocol only, and cost, calibration, and sensitivity measurements restricted to one dataset (cost to one seed). Evaluation is conducted exclusively on English-language benchmarks, leaving generalisation to low-resource and code-switched settings open; the strongest multimodal architectures (ViLBERT, CroMe) carry substantial computational overhead; the adversarial defences are assessed under fixed rather than adaptive attack settings; real-time streaming deployment, API integration, and latency benchmarking remain beyond the present scope; and the continual learning contributions are restricted to text-only scenarios.

Ultimately, the core challenge of fake news detection is not primarily algorithmic: sufficiently capable systems already exist for many constrained scenarios. It is structural — misinformation is generated faster than it can be verified, spreads through channels that resist monitoring, evolves in response to deployed countermeasures, and now benefits from LLM-powered generation that has invalidated the surface-feature approaches

of the field's early progress. Meeting this challenge requires the convergence of technical advances — continual learning, retrieval-augmented grounding, multimodal robustness, and multilingual coverage — with institutional, regulatory, and platform-level coordination that situates automated detection in its appropriate sociotechnical context. This thesis contributes the technical foundation for that convergence.

6.2 Future Directions

The contributions of this thesis open eight concrete directions for future research:

1. Real-time streaming fake news detection: The tiered architecture proposed in Chapter 5 (COVID-19 Infodemic Review), a lightweight $t = 0$ content classifier followed by RAG-augmented propagation-aware pipeline, requires engineering implementation. Combining a continual learning backbone (the exemplar-free EAKD-CFND configuration where drift is moderate, and EAKD-CFND-R where drift is severe and storage is available) with a streaming data ingestion layer, sub-100ms inference latency requirements, and online knowledge base updates would produce a system addressing both temporal adaptation and evidence grounding, with EAKD-CFND-R the only configuration with a demonstrated benefit under severe drift.

2. Multilingual and cross-cultural fake news detection: The English-centric benchmarks mean deployed systems perform worse in linguistic and regional contexts where institutional fact-checking infrastructure is also weakest [10]. Extending the graph-RAG framework to multilingual knowledge bases and the FakeNews Buster architecture to multilingual transformer backbones (XLM-RoBERTa, mBERT) would address this gap while enabling evaluation on underrepresented language communities.

3. Deepfake video and audio as new modalities: The multimodal evaluation in Chapter 4 is restricted to still images and text. Video deepfakes and synthetic audio generated by text-to-speech and voice cloning systems are the next generation of multimodal misinformation. Extending the Synergy Gain metric framework to video + text + audio combinations, and evaluating Unimodal Bias across three modalities, would extend the contribution of the Comprehensive Multimodal Evaluation.

4. Federated learning for privacy-preserving fake news detection: High-impact misinformation spreads disproportionately through encrypted platforms (WhatsApp, Telegram, Signal) that are structurally inaccessible to centralised dataset construction. Federated learning [369] enables model training across distributed data sources without centralising raw content, potentially enabling detection on encrypted platforms through participation of platform operators while preserving user privacy.

5. LLM-as-judge for automated fact verification: The evidence chains produced by the Graph-RAG framework provide structured inputs amenable to LLM-based reasoning about claim verifiability. Using a capable LLM as an autonomous fact-checker that retrieves evidence, reasons over contradictions, and generates natural-language verdicts with supporting citations would combine the structural advantages of Graph-RAG with the reasoning capacity of modern LLMs.

6. Regulatory and ethical framework for AI-based FND deployment: The explainability contributions of this thesis (LIME in Three-Label CT-BERT and FakeNews Buster, graph evidence chains in Graph-RAG, an inspectable but poorly calibrated

uncertainty trigger in EAKD-CFND) address the technical dimension of explainable AI for fake news detection. The regulatory and ethical dimensions, including who bears responsibility for automated moderation decisions, how appeals processes should be structured, and how detection bias across demographic groups should be audited, require interdisciplinary work extending beyond the scope of this thesis.

7. Extending EAKD-CFND to social graph dynamics: The EAKD-CFND framework currently operates on text-only class-incremental task sequences, with calibration, sensitivity, and cost analyses limited to a single dataset. Incorporating dynamic propagation graph features as additional modalities, specifically the temporal evolution of retweet trees as a continual learning stream, would combine the strengths of EAKD-CFND's adaptive distillation with the GNN-based social context approaches that achieved F1 of 95% on the MediaEval dataset. This would allow content-level and propagation-level concept drift to be studied together.

8. Calibration-aware and drift-aware continual learning: The EAKD-CFND evaluation showed that the entropy and maximum-softmax uncertainty signals are poorly calibrated (ECE 0.80–0.84) and that accuracy is insensitive to the verification threshold. Replacing the fixed global threshold with a calibration-aware or learned one, testing larger rehearsal buffers, recording Forward Transfer baselines, constructing artificial high- and low-similarity task splits from a single corpus, and separating no-match from transport failures in the verification module are the most direct next steps.

Prioritisation. The eight directions are not equally urgent. Table 6.1 ranks them (rank 1 to 8, in order of priority, which differs from the numbering of the list above) by the strength of the evidence in this thesis that the gap is real, the expected benefit, and the effort and dependencies involved. Calibration-aware thresholding, multilingual coverage, and a streaming implementation rank highest because each follows directly from a measured weakness (ECE of 0.80–0.84, English-only evaluation in every study, and the collapse of exemplar-free learning under severe drift) and can be pursued with existing models and data. Directions that depend on new datasets or on components that are not yet robust (video and audio modalities, extension of EAKD-CFND to propagation graphs) are ranked lower.

6.3 Social Impact

The contributions of this thesis extend beyond algorithmic performance to address a problem of direct societal consequence. Misinformation degrades public health outcomes, distorts democratic deliberation, and erodes institutional trust; the COVID-19 infodemic documented in the COVID-19 Infodemic Review [10] demonstrated that the velocity of false claims can directly undermine vaccination uptake and public compliance with health guidance. By formalising the limits of existing detection paradigms and proposing more robust alternatives, this work aims to support the information ecosystem on which informed public decision-making depends.

Public-health and epidemic preparedness: The COVID-19 Infodemic Review and the Three-Label CT-BERT framework are oriented towards health misinformation, where classification errors carry tangible risk. The explicit *Can't Say* uncertainty class (Three-Label CT-BERT) addresses the risk of confidently mislabelling borderline

Table 6.1: Prioritisation of the future research directions

Direction	**Priority**	**Justification from this thesis**	**Effort; dependency**
1. Calibration-aware thresholds	High	ECE 0.80–0.84 and a verification trigger insensitive to θ show the main EAKD-CFND weakness	Low; existing models
2. Multilingual and cross-cultural detection	High	All studies are English-only; gaps G7 and G10 remain open	High; multilingual datasets
3. Real-time streaming detection	High	EAKD-CFND-R recovers 85.5% under severe drift; Fast Graph-RAG runs in 9.2 ms/claim	Medium; engineering and API access
4. LLM-as-judge verification	Medium	Graph-RAG evidence chains give structured inputs; helps where multi-hop evidence exists (FEVER)	Medium; LLM cost
5. Regulatory and ethical framework	Medium	Unimodal Bias and attack vulnerability show deployment risks that need governance	Medium; interdisciplinary
6. Federated learning	Medium	Encrypted platforms are inaccessible to central data collection (gap G10)	High; platform cooperation
7. Deepfake video and audio	Lower	Synergy Gain is defined for images only; new modalities need new datasets	High; new benchmarks
8. EAKD-CFND with social graphs	Lower	The text-only EAKD-CFND does not yet match rehearsal under severe drift	High; depends on direction 8

claims, an essential property for systems advising the public during a fast-moving health crisis such as the 2026 hantavirus episode discussed in Chapter 5.

Support for human fact-checkers, not replacement: The comparative deployment guide (Chapter 5) and the evidence-grounded explanations produced by Graph-RAG are designed to augment professional fact-checkers operating at platform scale, surfacing supporting evidence chains and confidence estimates rather than issuing opaque automated verdicts. This keeps human judgement in the loop for consequential moderation decisions.

Fairness, accountability, and responsible deployment: The thesis also documents risks that must be managed before deployment: Unimodal Bias and dataset-artefact

shortcuts (Comprehensive Multimodal Evaluation, Adversarial Robustness) can produce systematically unfair outcomes, and the demonstrated vulnerability to LLM-generated adversarial content (Adversarial Robustness, LLM-Driven Perturbations) shows that detectors can be gamed. Reporting these failure modes openly, alongside the English-language and cultural limitations noted above, is itself a contribution to responsible, equitable use of automated fake news detection.

Scenario-based illustration. To make the potential impact concrete, Table 6.2 works through three deployment scenarios. The scenarios are illustrative: the assumed volumes and reviewer capacities are round numbers chosen for arithmetic, whereas the detector quantities are those measured in this thesis.

Table 6.2: Illustrative scenarios for the social impact of the thesis contributions

Scenario	**Assumptions (illustrative)**	**Thesis quantity used**	**Illustrative outcome**
Outbreak triage (health claims)	50,000 claims/day; 10% routed to *Can't Say*; 150 claims/day per human reviewer	Three-Label CT-BERT: 93% accuracy; Fast Graph-RAG: 9.2 ms/claim	Screening takes about 7.7 min of GPU time; humans review 5,000 claims (33 reviewers instead of 333 for manual review); about 3,150 automated errors remain to be audited
Election-week multimodal moderation	100,000 posts/day; 20% arrive with a missing or replaced image	BertVgg Text-Only F1 0.00; ViLBERT Text-Only F1 0.75	For the 20,000 affected posts, a collapse-prone model detects effectively none, whereas an architecture passing the Synergy Gain screen keeps F1 near 0.75
Newsroom fact-check support	500 claims/week; evidence assembly saves 5 min per claim	Graph-RAG evidence chains (FEVER macro-F1 0.755)	About 42 analyst hours/week, roughly one full-time position, redirected to judgement

REFERENCES

[1] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[2] D. M. Lazer, M. A. Baum, Y. Benkler *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[3] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty, "Fighting an infodemic: COVID-19 fake news dataset," in *Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Springer, 2021, pp. 21–29.

[4] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT 2019*. Association for Computational Linguistics, 2019, pp. 4171–4186.

[6] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, pp. 41–57, 2019.

[7] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Is bert really robust? natural language attack on text classification and entailment," *arXiv preprint arXiv:1907.11932*, 2019.

[8] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *CoRR*, vol. abs/1412.6572, 2015. [Online]. Available: https://arxiv.org/abs/1412.6572

[9] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys*, vol. 53, no. 5, pp. 1–40, 2020.

[10] A. Jolly and S. Kumar, "Mitigating the infodemic: Lessons from COVID-19 fake news detection for epidemic preparedness in the age of generative AI," *Information Processing & Management*, 2026, manuscript under review.

[11] A. Jolly and S. Kumar, "Navigating the gray area: A three-label framework for uncovering uncertainty in fake news," in *Proceedings of the International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*. Sharda University, Greater Noida: IEEE, November 2023.

[12] A. Jolly and S. Kumar, "FakeNews Buster: A continual learning approach with BERT and FastFormer for adaptive and explainable fake news detection," in *Proceedings of the International Conference on Intelligent and Secure Engineering Solutions (CISES)*. GL Bajaj, Greater Noida: IEEE, August 2025.

[13] A. Jolly, S. Kumar, and S. Paliwal, "EAKD-CFND: Continual fake news detection through adaptive knowledge distillation with external verification," *Scientific Reports*, 2026, accepted for publication; volume, article number and DOI to be added.

[14] A. Jolly and S. Kumar, "Retrieval without generation: Evidence-grounded fake news classification," in *Proceedings of the International Conference on Communication and Intelligent Systems (ICCIS)*. BITS Pilani, K. K. Birla Goa Campus, Goa: Springer, September 2026, presented, September 26–27, 2026.

[15] A. Jolly and S. Kumar, "A rigorous interrogation of multimodal fake news detection paradigms: Adversarial robustness and modality dependence," *Cluster Computing*, vol. 29, p. 307, June 2026.

[16] A. Jolly and S. Kumar, "Beyond the surface: Adversarial stress testing of multimodal fake news detection models with LLM-driven perturbations," *Computers and Electrical Engineering*, vol. 135, p. 111203, 2026.

[17] A. Jolly and S. Kumar, "A comprehensive evaluation of multimodal fake news detection: Assessing robustness and efficacy," *Arabian Journal for Science and Engineering*, May 2026.

[18] K. Meel and P. Shrihari, "Role of statistics in detecting misinformation: A review of the state of the art, open issues, and future research directions," *Annual Review of Statistics and Its Application*, vol. 10, pp. 429–454, 2023.

[19] J. Nasir *et al.*, "Sustainable development of information dissemination: A review of current fake news detection research and practice," *Information*, vol. 11, no. 9, p. 458, 2020.

[20] R. A. Duma, Z. Niu, A. S. Nyamawe, J. Tchaye-Kondi, N. Jingili, A. A. Yusuf, and A. F. Deve, "Fake review detection techniques, issues, and future research directions: a literature review," *Knowledge and Information Systems*, pp. 1–42, 2024.

[21] A. Athira, S. M. Kumar, and A. M. Chacko, "A systematic survey on explainable ai applied to fake news detection," *Engineering Applications of Artificial Intelligence*, vol. 122, p. 106087, 2023.

[22] S. Jadhav *et al.*, "Fake news detection revisited: An extensive review of theoretical frameworks, dataset assessments, model constraints, and forward-looking research agendas," *Engineering Proceedings*, vol. 32, no. 1, p. 222, 2022.

[23] N. Arega, K. Kalayu, and E. Yitbarek, "A review of deep-learning-based models for Afaan Oromo fake news detection," *Journal of Language Technology*, vol. 3, no. 1, pp. 45–58, 2025.

[24] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable ai systems," *arXiv:1811.11839 [cs]*, 2020.

[25] M. Ali *et al.*, "Enhancing disinformation detection with explainable ai and named entity replacement," *arXiv preprint arXiv:2502.04863*, 2025.

[26] S. Xie, T. Qiao, S. Li, X. Zhang, J. Zhou, and G. Feng, "Deepfake detection in the aigc era: A survey, benchmarks, and future perspectives," *Information Fusion*, vol. 127, p. 103740, 2026.

[27] W. Zhang, J. Liu, and H. Wang, "Federated learning in detecting fake news: A survey," *Information Fusion*, vol. 102, p. 102034, 2025.

[28] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, 2017.

[29] A. Hergovich, R. Schott, and C. Burger, "Biased evaluation of abstracts depending on topic and conclusion," *Current Psychology*, vol. 29, no. 3, pp. 188–209, 2010.

[30] W. Yang, E. Aghasian, S. Garg *et al.*, "A survey on blockchain-based internet service architecture," *IEEE Access*, vol. 7, pp. 75 845–75 872, 2019.

[31] FactCheck.org, "FactCheck.org," FactCheck.org, 2024, accessed: 2026-06-15. [Online]. Available: https://www.factcheck.org/

[32] Agence France-Presse, "AFP fact check," AFP, 2024, accessed: 2026-06-15. [Online]. Available: https://factcheck.afp.com/

[33] Poynter, "Poynter – a global leader in journalism," Poynter, 2024, accessed: 2026-06-15. [Online]. Available: https://www.poynter.org/

[34] H. Kaur *et al.*, "Fake news detection and classification: A comparative study of convolutional neural networks, large language models, and natural language processing models," *Electronics*, vol. 17, no. 1, p. 28, 2024.

[35] J. Y. Khan, M. T. I. Khondaker, S. Afroz, G. Uddin, and A. Iqbal, "A benchmark study of machine learning models for online fake news detection," *Machine Learning with Applications*, vol. 4, p. 100032, 2021.

[36] F. A. Alshuwaier and F. A. Alsulaiman, "Fake news detection using machine learning and deep learning algorithms: A comprehensive review and future perspectives," *Computers*, vol. 14, no. 9, p. 394, 2025.

[37] S. Mohseni and E. Ragan, "Machine learning explanations to prevent overtrust in fake news detection," in *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, vol. 15, 2021, pp. 463–474.

[38] S. Mohseni, J. E. Block, and E. D. Ragan, "A human-grounded evaluation benchmark for local explanations of machine learning," *arXiv:1801.05075 [cs]*, 2020.

[39] H. Li, J. Wang, J. Li, and C. Liu, "Cross-lingual font style transfer with full-domain convolutional attention," *Computers & Electrical Engineering*, vol. 109, p. 108743, 2023.

[40] V. Sorokina *et al.*, "A shap-based xai approach to evaluating machine learning classification of fake news," in *Studies in Health Technology and Informatics*, vol. 306, 2023, pp. 124–133.

[41] D. Sharma *et al.*, "Multi-class fake news detection using deep learning and explainable ai: A transformer-gradient boosting approach," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 5, pp. 583–594, 2024.

[42] M. F. Mridha, A. J. Keya, M. A. Hamid, M. M. Monowar, and M. S. Rahman, "A comprehensive review on fake news detection with deep learning," *IEEE access*, vol. 9, pp. 156 151–156 170, 2021.

[43] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Communications of the ACM*, vol. 63, no. 1, pp. 68–77, 2019.

[44] P. Song, C. Geng, and Z. Li, "Research on text classification based on convolutional neural network," *Proceedings of ICCNEA 2019*, pp. 229–232, 2019.

[45] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2014, pp. 1746–1751. [Online]. Available: https://aclweb.org/anthology/D14-1181/

[46] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[47] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[48] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[49] M. E. Peters, M. Neumann, M. Iyyer *et al.*, "Deep contextualized word representations," in *Proceedings of NAACL-HLT 2018*, 2018, pp. 2227–2237.

[50] N. Singh and H. Sabrol, "Convolutional neural networks – an extensive arena of deep learning. a comprehensive study," *Archives of Computational Methods in Engineering*, vol. 28, no. 7, pp. 4755–4780, 2021.

[51] H. Padalko, V. Chomko, and D. Chumachenko, "A novel approach to fake news classification using lstm-based deep learning models," *Frontiers in big Data*, vol. 6, p. 1320800, 2024.

[52] S. H. Lee, H. Goëau, P. Bonnet, and A. Joly, "Attention-based recurrent neural network for plant disease classification," *Frontiers in Plant Science*, vol. 11, p. 601250, 2020.

[53] A. Kamal and M. Abulaish, "Contextualized satire detection in short texts using deep learning techniques," *Journal of Web Engineering*, vol. 23, no. 1, pp. 27–52, 2024.

[54] S. Naudan, "Emojis aid social media sentiment analysis: Stop cleaning them out," Towards Data Science, 2020, accessed: 2026-06-15. [Online]. Available: https://towardsdatascience.com/emojis-aid-social-media-sentiment-analysis-stop-cleaning-them-out-bb32a1e5fc8e/

[55] J. Pimentel, "Are emojis the new words? a sentiment analysis of social media brand conversations," *Journal of International Technology and Information Management*, vol. 28, no. 2, 2019, article 4.

[56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. Online: Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[57] A. Wani, I. Joshi, S. Khandve, V. Wagh, and R. Joshi, "Evaluating deep learning approaches for COVID-19 fake news detection," in *Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Springer, 2021, pp. 153–163.

[58] M. Müller, M. Salathé, and P. E. Kummervold, "COVID-twitter-BERT: A natural language processing model to analyse COVID-19 content on twitter," *arXiv preprint arXiv:2005.07503*, 2020.

[59] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," *arXiv:1909.11942 [cs]*, 2020.

[60] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[61] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[62] B. Chen, B. Chen, D. Gao, Q. Chen, C. Huo, X. Meng, W. Ren, and Y. Zhou, "Transformer-based language model fine-tuning methods for COVID-19 fake news detection," in *Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Springer, 2021, pp. 83–92.

[63] R. Vijjali, P. Potluri, S. Kumar, and S. Teki, "Two stage transformer model for COVID-19 fake news detection and fact checking," *arXiv preprint arXiv:2011.13253*, 2020.

[64] A. Koirala, "COVID-19 fake news dataset," Mendeley Data, 2021.

[65] B. Al-Ahmad, A. M. Al-Zoubi, R. Abu Khurma, and I. Aljarah, "An evolutionary fake news detection method for COVID-19 pandemic information," *Symmetry*, vol. 13, no. 6, p. 1091, 2021.

[66] S. M. S.-U.-R. Shifath, M. F. Khan, and M. S. Islam, "A transformer based approach for fighting COVID-19 fake news," *arXiv:2101.12027 [cs]*, 2021.

[67] S. Gundapu and R. Mamidi, "Transformer based automatic COVID-19 fake news detection system," *arXiv:2101.00180 [cs]*, 2021.

[68] W. Shishah, "Fake news detection using bert model with joint learning," *Arabian Journal for Science and Engineering*, vol. 46, no. 9, pp. 9115–9127, 2021.

[69] A. Gautam, V. Venktesh, and S. Masud, "Fake news detection system using XLNet model with topic distributions: CONSTRAINT@AAAI2021 shared task," in *Combating Online Hostile Posts in Regional Languages during Emergency Situation*. Springer, 2021, pp. 189–200.

[70] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020. [Online]. Available: https://arxiv.org/abs/2004.05150

[71] X. Fu *et al.*, "Analysis of emoticon based on bert model," in *Proceedings of the 2022 8th International Conference on Frontiers of Industrial Engineering (ICFIE)*, 2022, pp. 831–835, assuming Emoticon -> Emoji.

[72] P. Jain, "New transformer variants keep flooding the market, here's one from microsoft called fastformer," Analytics India Magazine, 2021, accessed: 2026-06-15. [Online]. Available: https://analyticsindiamag.com/global-tech/new-transformer-variants-keep-flooding-the-market-heres-one-from-microsoft-called-fastformer/

[73] A. Kamal, P. Mohankumar, and V. K. Singh, "Financial misinformation detection via roberta and multi-channel networks," in *International Conference on Pattern Recognition and Machine Intelligence*. Cham: Springer Nature Switzerland, 2023, pp. 646–653.

[74] P. Mohankumar, A. Kamal, V. K. Singh, and A. Satish, "Financial fake news detection via context-aware embedding and sequential representation using cross-joint networks," in *2023 15th International Conference on COMmunication Systems & NETworkS (COMSNETS)*. Online: IEEE, 2023, pp. 780–784.

[75] M. Khan, "Advancements in natural language processing: Exploring transformer-based architectures for text understanding," *arXiv preprint arXiv:2503.20227*, 2025.

[76] H. Al-Zahrani and M. Al-Yahya, "Pre-trained language model ensemble for Arabic fake news detection," *Mathematics*, vol. 12, no. 18, p. 2891, 2024.

[77] S. Arora *et al.*, "Uncertainty quantification for transformer models for dark-pattern detection," *arXiv preprint arXiv:2412.05251*, 2024.

[78] A. Vlachos, "Uncertainty estimation of transformers' predictions via topological analysis of the attention matrices," *arXiv preprint arXiv:2308.11295*, 2023.

[79] Y. Wu *et al.*, "Explainable fake news detection with large language model via defense among competing wisdom," *arXiv preprint arXiv:2405.03371*, 2024.

[80] X. Ma, Z. Huang, Y. R. Hu, and Z. Liu, "On fake news detection with LLM enhanced semantics mining," in *Proceedings of EMNLP 2024*, 2024.

[81] Z. Wang, Y. Chen, M. Zhang, and J. Ma, "LLM-enhanced multimodal detection of fake news," *PLoS ONE*, vol. 19, no. 4, p. e0302245, 2024.

[82] Y. Cao, A. M. Nair, E. Eyimife, N. J. Soofi, K. Subbalakshmi, J. R. Wullert II, C. Basu, and D. Shallcross, "Can large language models detect misinformation in scientific news reporting?" *arXiv preprint arXiv:2402.14268*, 2024.

[83] Z. Wang, X. Chen, and X. Liu, "Prompt-induced linguistic fingerprints for LLM-generated fake news detection," in *Proceedings of the ACM Web Conference 2026 (WWW)*, 2026.

[84] Z. Wang, A. Yao, G. Xu, and M. Ren, "A large language model-based approach for fake review detection: the implicit characteristics perspective," *Information Processing & Management*, vol. 63, no. 1, p. 104352, 2026.

[85] S. Yin, Y. Wang, D. Hou, W. An, C. Gao, X. Li, and Z. Wang, "Llm-assisted fake news detection with adaptive boosting framework incorporating contrastive learning," *Information Processing & Management*, vol. 63, no. 7, p. 104789, 2026.

[86] H. Hu, X. Wang, and J. Li, "Synergizing LLMs with global label propagation for multimodal fake news detection," *arXiv preprint arXiv:2501.12345*, 2025.

[87] X. Chen and K. Shu, "Have LLMs reopened the pandora's box of AI-generated fake news?" *arXiv preprint arXiv:2405.18726*, 2024.

[88] H. Wen, X. Li, and Y. Chen, "A dynamic knowledge update-driven model with large language models for fake news detection," *arXiv preprint arXiv:2502.01234*, 2025.

[89] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: efficient finetuning of quantized llms (2023)," *arXiv preprint arXiv:2305.14314*, vol. 52, pp. 3982–3992, 2023.

[90] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *arXiv preprint arXiv:2305.14314*, 2023.

[91] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[92] P. Buzzega, M. Boschini, A. Porrello, D. Abati, and S. Calderara, "Dark experience for general continual learning: a strong, simple baseline," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 15 920–15 930.

[93] K. M. K. Hasan, M. A. A. Chowdury, M. M. A. Khan, and M. R. Islam, "Fake news detection with incremental deep learning based concept drift adaption," *SN Computer Science*, vol. 4, no. 5, p. 591, 2023.

[94] F. Alrowili and K. Roy, "Concept drift adaptation in text stream mining settings: A systematic review," *arXiv preprint arXiv:2312.02901*, 2023.

[95] M. Saha, D. Das, and S. Das, "Analysis of concept drift in fake reviews detection," in *Proceedings of the 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, 2020, pp. 343–348.

[96] J. Kang *et al.*, "Class-incremental learning by knowledge distillation with adaptive feature consolidation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 717–10 726.

[97] S. Ratna, "Knowledge distillation and continual learning for optimized deep neural networks," Ph.D. dissertation, University of Wollongong, 2023.

[98] C. Song, N. Ning, Y. Zhang, and B. Wu, "Knowledge augmented transformer for adversarial multidomain multiclassification multimodal fake news detection," *Neurocomputing*, vol. 462, pp. 88–100, 2021.

[99] A. Rahaman *et al.*, "Continual learning with knowledge distillation: A survey," *arXiv preprint arXiv:2312.11085*, 2023.

[100] Y. Han, S. Karunasekera, and C. Leckie, "Graph neural networks with continual learning for fake news detection from social media," *arXiv:2007.03316 [cs]*, 2020.

[101] Y. Tian *et al.*, "Densely distilling cumulative knowledge for continual learning," *arXiv preprint arXiv:2405.09820*, 2024.

[102] S. Ganguly, R. Nayak, R. Rao, U. Deb, and P. AP, "AdaKD: Dynamic knowledge distillation of ASR models using adaptive loss weighting," 2024.

[103] Y. Feng *et al.*, "Learning dynamic representations via an optimally-weighted maximum mean discrepancy optimization framework for continual learning," *arXiv preprint arXiv:2501.12121*, 2025.

[104] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[105] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 4765–4774.

[106] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, Online, 2017, pp. 618–626.

[107] S. Abdali, S. Shaham, and B. Krishnamachari, "Multi-modal Misinformation Detection: Approaches, Challenges and Opportunities," *ACM Computing Surveys*, vol. 57, no. 3, pp. 1–29, Mar. 2025.

[108] Y. Guo, H. Ge, and J. Li, "A two-branch multimodal fake news detection model based on multimodal bilinear pooling and attention mechanism," *Frontiers in Computer Science*, vol. 5, p. 1159063, 2023.

[109] Y. Yan, H. Fu, and F. Wu, "Multimodal social media fake news detection based on 1d-ccnet attention mechanism," *Electronics*, vol. 13, no. 18, p. 3700, 2024.

[110] M. Nasser, N. I. Arshad, A. Ali, H. Alhussian, F. Saeed, A. Da'u, and I. Nafea, "A systematic review of multimodal fake news detection on social media using deep learning models," *Results in Engineering*, vol. 26, p. 104752, 2025.

[111] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, Online, 2015. [Online]. Available: https://arxiv.org/abs/1409.1556

[112] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[113] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[114] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. Online: PmLR, 2021, pp. 8748–8763.

[115] OpenAI, "CLIP: Connecting text and images," OpenAI Blog, 2021, accessed: 2026-06-15. [Online]. Available: https://openai.com/blog/clip/

[116] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "Eann: Event adversarial neural networks for multi-modal fake news detection," in *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, Online, 2018, pp. 849–857.

[117] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "Mvae: Multimodal variational autoencoder for fake news detection," in *The world wide web conference*, Online, 2019, pp. 2915–2921.

[118] S. Singhal, A. Kabra, M. Sharma, R. R. Shah, T. Chakraborty, and P. Kumaraguru, "Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract)," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 10, Online, 2020, pp. 13 915–13 916.

[119] X. Zhou, A. Mulay, E. Ferrara, and R. Zafarani, "ReCOVery: A Multimodal Repository for COVID-19 News Credibility Research," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, ser. CIKM '20. New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 3205–3212.

[120] X. Zhou, J. Wu, and R. Zafarani, ": Similarity-aware multi-modal fake news detection," in *Pacific-Asia Conference on knowledge discovery and data mining*. Springer, 2020, pp. 354–367.

[121] K. Shu, L. Cui, S. Wang, D. Lee, and H. Liu, "dEFEND: Explainable fake news detection," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019, pp. 395–405.

[122] T. Zhang, D. Wang, H. Chen, Z. Zeng, W. Guo, C. Miao, and L. Cui, "Bdann: Bert-based domain adaptation neural network for multi-modal fake news detection," in *2020 international joint conference on neural networks (IJCNN)*. Online: IEEE, 2020, pp. 1–8.

[123] D. Jin, J. Cao, H. Guo, and J. Zhang, "Multimodal fusion with attentional mechanism for fake news detection," in *Proceedings of the 2017 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. Online: IEEE, 2017, pp. 439–444. [Online]. Available: https://ieeexplore.ieee.org/document/8258032

[124] Z. Jin, J. Cao, H. Guo, J. Zhang, and J. Luo, "Multimodal fusion with recurrent attention for rumor detection on microblogs," in *Proceedings of the 25th ACM international conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2017, p. 1616–1624.

[125] M. Wu, Y. Li, Q. Zhang, B. Hooi, A. Sun, and J.-Y. Wu, "Cross-modal metric learning for multimodal fake news detection," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, p. 1492–1496.

[126] R. Tan, B. Plummer, and K. Saenko, "Detecting cross-modal inconsistency to defend against neural fake news," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 2081–2106.

[127] I. Bejan, "Memosys at semeval-2020 task 8: Multimodal emotion analysis in memes," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, Online, 2020, pp. 1172–1178.

[128] S. Qian, J. Wang, Z. Wang, Y. Wang, S. Wang, and K. Shu, "Multimodal fake news detection," *arXiv preprint arXiv:2104.10233*, 2021, arXiv:2104.10233 [cs.SI].

[129] Y. Wang, F. Ma, H. Wang, K. Jha, and J. Gao, "Multimodal emergent fake news detection via meta neural process networks," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 3708–3716.

[130] S. Hangloo and B. Arora, "Combating multimodal fake news on social media: methods, datasets, and future perspective," *Multimedia systems*, vol. 28, no. 6, pp. 2391–2422, 2022.

[131] P. Meel and D. K. Vishwakarma, "Multi-modal fusion using fine-tuned self-attention and transfer learning for veracity analysis of web information," *Expert Systems with Applications*, vol. 229, p. 120537, 2023.

[132] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 7077–7087.

[133] R. Suzuki, H. Yanaka, M. Yoshikawa, K. Mineshima, and D. Bekki, "Multimodal logical inference system for visual-textual entailment," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 2019, pp. 386–392.

[134] J. Cao, Z. Gan, Y. Cheng, L. Yu, Y.-C. Chen, and J. Liu, "Behind the scene: Revealing the secrets of pre-trained vision-and-language models," in *Computer Vision – ECCV 2020*, vol. 12351, 2020, pp. 565–580.

[135] A. Agarwal *et al.*, "Method and apparatus for collecting, detecting and visualizing fake news," 2020, u.S. Patent Application No. PCT/US2019/052031.

[136] B. Han *et al.*, "Emoji multimodal microblog sentiment analysis based on mutual attention," *Journal of Healthcare Engineering*, 2022.

[137] P. Mohankumar, V. K. Singh, and A. Kamal, "Financial fact-check via multi-modal embedded representation and attention-fused network," in *International Conference on Data Analytics & Management*. Singapore: Springer Nature Singapore, 2024.

[138] Straive, "The quest to detect and prevent image manipulation," Straive Blogs, 2023, accessed: 2026-06-15. [Online]. Available: https://www.straive.com/blogs/the-quest-to-detect-and-prevent-image-manipulation/

[139] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*. Online: Curran Associates, Inc., 2019, pp. 13–23. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf

[140] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Online: Association for Computational Linguistics, 2019, pp. 5100–5111. [Online]. Available: https://aclanthology.org/D19-1514

[141] W. Su, X. Zhu, Y. Cao *et al.*, "VL-BERT: Pre-training of generic visual-linguistic representations," *arXiv:1908.08530 [cs]*, 2020.

[142] Y.-C. Chen, L. Li, L. Yu *et al.*, "UNITER: UNiversal image-text representation learning," *arXiv:1909.11740 [cs]*, 2020.

[143] R. Shekhar, E. Takmaz, R. Fernández, and R. Bernardi, "Evaluating the representational hub of language and vision models," in *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, 2019, pp. 211–222.

[144] O. Caglayan, P. Madhyastha, L. Specia, and L. Barrault, "Probing the need for visual context in multimodal machine translation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 4159–4170.

[145] L. Parcalabescu, A. Gatt, A. Frank, and I. Calixto, "Seeing past words: Testing the cross-modal capabilities of pretrained v&l models on counting tasks," *arXiv:2012.12352 [cs]*, 2021.

[146] R.-z. Li, S.-h. Liu, Z.-x. Zhang, Z.-h. Cui, Y.-w. Li, and T.-r. Li, "MFIR: Multimodal fusion and inconsistency reasoning for explainable fake news detection," *Information Fusion*, vol. 91, pp. 555–567, 2023.

[147] E. Choi, J. Ahn, X. Piao, and J.-K. Kim, "Crome: Multimodal fake news detection using cross-modal tri-transformer and metric learning," *arXiv preprint arXiv:2501.12422*, 2025.

[148] E. A. Rocamora, C. Schlarmann, N. D. Singh, Y. Wu, M. Hein, and V. Cevher, "Robustness in both domains: Clip needs a robust text encoder," *arXiv preprint arXiv:2506.03355*, 2025.

[149] Y. Li, W. Lu, H. Yu, and Y. Wang, "Retrieval-augmented multimodal model for fake news detection," *arXiv preprint arXiv:2604.18112*, 2026, rAMM: multimodal LLM backbone with retrieval augmentation for fake news detection. Author list verified via WebSearch against the arXiv record (2026-08-04), matches exactly.

[150] Z. Zhang, L. Yang, H.-Y. Xu, L. Yu, W.-T. Zhang, and J. Cao, "When llm-based fact-checkers meet multimodal in-context learning for multimodal fake news detection," *arXiv preprint arXiv:2401.12422*, 2024.

[151] H. Chen, H. Wang, Z. Liu, Y. Li, Y. Hu, Y. Zhang, K. Shu, R. Li, and P. S. Yu, "Multi-modal robustness fake news detection with cross-modal and propagation network contrastive learning," *Knowledge-Based Systems*, vol. 309, p. 112800, 2025.

[152] L. Chen, Z. Qian, P. Li, and Q. Zhu, "Multimodal fake news video explanation: Dataset, model and evaluation," *Information Processing & Management*, vol. 63, no. 6, p. 104769, 2026.

[153] T. Niu, S. Ma, D. Yan, Q. Guan, P. Hou, J. Sun, and H. Cao, "Cr-fnd: Robustness-enhanced multimodal fake news detection against content variance," *Neurocomputing*, vol. 638, p. 130190, 2025.

[154] Y. Yuan, Z. Li, and B. Zhao, "A survey of multimodal learning: Methods, applications, and future," *ACM Computing Surveys*, 2025.

[155] C. Li, Y. Liao, C. Ding, and Z. Ye, "Mdapt: Multi-modal depth adversarial prompt tuning to enhance the adversarial robustness of visual language models," *Sensors*, vol. 25, no. 1, p. 258, 2025.

[156] A. Mosallanezhad, M. Karami, K. Shu, M. V. Mancenido, and H. Liu, "Domain adaptive fake news detection via reinforcement learning," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 3632–3640.

[157] J. Shi, L. Zhang, and C. Gu, "Multimodal large language models to support real-world fact-checking," *arXiv preprint arXiv:2403.09297*, 2024.

[158] P. Liu, W. Qian, D. Xu, B. Ren, and J. Cao, "Multi-Modal Fake News Detection via Bridging the Gap between Modals," *Entropy*, vol. 25, no. 4, p. 614, Apr. 2023.

[159] J. Qiu, K. Zhang, and X. Wu, "Temporal enhanced multimodal graph neural networks for fake news detection," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 4, pp. 5234–5245, 2024.

[160] Z. Sun, X. Wang, and Y. Liu, "Multi-modal deepfake detection and localization with FPN-transformer," *arXiv preprint arXiv:2502.03456*, 2025.

[161] S. Wei, Z. Wang, M. Li, X. Liu, and B. Wu, "Dccma-net: Disentanglement-based cross-modal clues mining and aggregation network for explainable multimodal fake news detection," *Information Processing & Management*, vol. 62, no. 4, p. 104089, 2025.

[162] Z. Li, J. Yang, X. Wang, J. Lei, S. Li, and J. Zhang, "Uncertainty-aware disentangled representation learning for multimodal fake news detection," *Information Processing & Management*, vol. 62, no. 5, p. 104190, 2025.

[163] Y. Meng, H. Wang, J. Zhao, Y. Sun, and M. Shao, "Dynamic hierarchical memory improved mixture-of-experts for multimodal fake news detection," *Information Processing & Management*, vol. 63, no. 2, p. 104479, 2026.

[164] L. Han, X. Zhang, T. Wang, Y. Liu, and Z. Dong, "Enhancing large language model for fake news video detection via cross-modal retrieval," *Information Processing & Management*, vol. 63, no. 2, p. 104471, 2026.

[165] L. Wang, X. Li, B. Zhou, Y. Zhang, J. Yuan, and H. Hu, "Multimodal fusion with llm content via hierarchical progressive transformer for explainable fake news detection," *Information Processing & Management*, vol. 63, no. 5, p. 104700, 2026.

[166] S. Liu, H. Zeng, J. Guo, C. Zhang, I.-K. Lee, Y. Zhang, and X. Li, "Dual-stream multimodal shared prompt model for fake news detection," *Information Processing & Management*, vol. 63, no. 6, p. 104759, 2026.

[167] L. Wu and G. Wang, "Mitigating cross-modal interference in multimodal fake news detection via adaptive feature optimization," *Information Processing & Management*, vol. 63, no. 7, p. 104776, 2026.

[168] M. Qian, S. Guo, Y. Chen, C. Wang, and Z. Ye, "Mfnd-dcl: Multimodal fake news detection based on dual contrastive learning," *Information Processing & Management*, vol. 63, no. 7, p. 104812, 2026.

[169] C. Jing, H. Gao, X. Zhang, T. Gao, and C. Zhou, "Dpsg: Dynamic propagation social graphs for multi-modal fake news detection," *Information Fusion*, vol. 113, p. 102595, 2025.

[170] F. Wu, H. Jin, F. Chen, Y. Ji, X.-Y. Jing, and G.-P. Jiang, "Balanced cross-modal prompt learning and fusion network for multi-modal fake news detection," *Information Fusion*, vol. 122, p. 103196, 2025.

[171] F. Wu, S. Chen, Z.-Y. Deng, D. Wu, C. Lan, Y. Ji, and X.-Y. Jing, "Pfbl: Prototype-based fully balanced learning for multimodal fake news detection," *IEEE Transactions on Computational Social Systems*, vol. 12, no. 5, pp. 3442–3454, 2025.

[172] Y. Kheddache, X. J. Hu, A. Fornés, and L. P. d. l. H. Lluís, "Effectiveness of large multimodal models in detecting multimodal disinformation," *arXiv preprint arXiv:2501.05678*, 2025.

[173] S. Jadhav, N. Kumar, and M. Gupta, "Explainable multilingual and multimodal fake-news detection: toward robust and trustworthy AI for combating misinformation," *Frontiers in Artificial Intelligence*, vol. 8, p. 1497009, 2025.

[174] J. Qiao, X. Li, C. Gao, L. Wu, J. Feng, and Z. Wang, "Improving multimodal fake news detection by leveraging cross-modal content correlation," *Information Processing & Management*, vol. 62, no. 5, p. 104120, 2025.

[175] Z. Liu, J. Wang, D. Jin, and K. Shu, "Modality Interactive Mixture-of-Experts for Fake News Detection," *arXiv preprint arXiv:2501.12431*, 2025. [Online]. Available: https://arxiv.org/abs/2501.12431

[176] X. Shen, M. Huang, Z. Hu, S. Cai, and T. Zhou, "Multimodal Fake News Detection with Contrastive Learning and Optimal Transport," *Frontiers in Computer Science*, vol. 6, Nov. 2024.

[177] P. Zhu, J. Hua, K. Tang, J. Tian, J. Xu, and X. Cui, "Multimodal fake news detection through intra-modality feature aggregation and inter-modality semantic fusion," *Complex & Intelligent Systems*, vol. 10, no. 4, pp. 5851–5863, Aug. 2024.

[178] F. Yan, M. Zhang, and B. Wei, "Multimodal integration for fake news detection on social media platforms," in *MATEC Web of Conferences*, vol. 395. Online: EDP Sciences, 2024, p. 01013.

[179] J. Sen and S. Dasgupta, "Adversarial attacks on image classification models: Fgsm and patch attacks and their impact," *arXiv preprint arXiv:2307.02055*, 2023.

[180] Z. Li, J. Herzig, Z. Liu, and J. Gao, "On the robustness of multimodal reasoning models to unaligned inputs," *arXiv preprint arXiv:2206.08788*, 2022.

[181] K. Yang, W.-Y. Lin, M. Barman, F. Condessa, and Z. Kolter, "Defending multimodal fusion models against single-source adversaries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3344–3353.

[182] Q. Zhang, H. Wu, C. Zhang, Q. Hu, H. Fu, J. T. Zhou, and X. Peng, "Provable dynamic fusion for low-quality multimodal data," in *International conference on machine learning*. PMLR, 2023, pp. 41 753–41 769.

[183] M. Schlarmann, J. H. Höllen, M. Schilling, A. Koch, T. Schneider, S. Stollen, and C. Wöhler, "On the adversarial robustness of multi-modal foundation models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2023, pp. 716–726.

[184] A. Rosebrock, "Defending against adversarial image attacks with keras and tensorflow," PyImageSearch, Mar 2021, accessed: 2026-06-15. [Online]. Available: https://pyimagesearch.com/2021/03/08/defending-against-adversarial-image-attacks-with-keras-and-tensorflow/

[185] L. Lu, S. Pang, S. Liang, H. Zhu, X. Zeng, A. Liu, Y. Liu, and Y. Zhou, "Adversarial training for multimodal large language models against jailbreak attacks," *arXiv*, 2025. [Online]. Available: https://arxiv.org/abs/2503.04833

[186] C.-T. Liao, B. Ren, G. Mei, and X. Zheng, "Adversarial robustness for unified multi-modal encoders via efficient calibration," *arXiv*, 2025. [Online]. Available: https://arxiv.org/abs/2505.11895

[187] A. Jolly and S. Kumar, "Adversarial robustness benchmarking of multimodal fake news detection architectures," *Computers and Electrical Engineering*, 2026.

[188] R. Li, J. Liu, J. Zhen, Y. Bao, Y. Qiang, Y. Xu, Y. Gu, Y. Jin, D. Yu, Y. He, J. Huang, W. Fu, Q. Zhao, J. Jiang, A. Zong, T. Wang, Z. Zhao, Z. Wu, M. Zhou, and H. Xue, "Adversarial attacks and defenses on large language models: A systematic review," *arXiv preprint arXiv:2503.xxxxx*, 2025, retrieved from ResearchGate.

[189] J. Chen, Y. Huang, M. Li, Y. Zhang, and W. Zhang, "Towards interpretable visual question answering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 15 830–15 839.

[190] S. Zhong, S. Peng, X. Liu, L. Zhu, X. Xu, and T. Li, "Ecarnet: enhanced clue-ambiguity reasoning network for multimodal fake news detection," *Multimedia Systems*, vol. 30, no. 1, p. 55, 2024.

[191] E. Raja, B. Soni, and S. K. Borgohain, "Harnessing heterogeneity: A multi-embedding ensemble approach for detecting fake news in dravidian languages," *Computers and Electrical Engineering*, vol. 120, p. 109661, 2024.

[192] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, no. 1, p. 1096, 2019.

[193] J. Ebrahimi, A. Rao, D. Lowd, and D. Dou, "Hotflip: White-box adversarial examples for text classification," *arXiv preprint arXiv:1712.06751*, 2017.

[194] J. Li, S. Ji, T. Du, B. Li, and T. Wang, "Textbugger: Generating adversarial text against real-world applications," *arXiv preprint arXiv:1812.05271*, 2018.

[195] J. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, "TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 119–126.

[196] H. Ali, M. S. Khan, A. Alghadhban, M. Alazmi, A. Alzamil, K. Al-Utaibi, and J. Qadir, "All your fake detector are belong to us: Evaluating adversarial robustness of fake-news detectors under black-box settings," *IEEE Access*, vol. 9, pp. 81 678–81 692, 2021.

[197] C. Koenders, J. Filla, N. Schneider, and V. Woloszyn, "How vulnerable are automatic fake news detection methods to adversarial attacks?" *arXiv:2107.07970 [cs]*, 2021.

[198] W. Lee, H. Lee, and S. goo Lee, "Semantics-preserving adversarial training," *arXiv*, 2020. [Online]. Available: https://arxiv.org/abs/2009.10978

[199] A. Bui, T. Le, H. Zhao, P. Montague, O. {de Vel}, T. Abraham, and D. Phung, "Improving ensemble robustness by collaboratively promoting and demoting adversarial robustness," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[200] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," in *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2018.

[201] Q.-Z. Cai, M. Du, C. Liu, and D. Song, "Curriculum adversarial training," *arXiv*, 2018. [Online]. Available: https://arxiv.org/abs/1805.04807

[202] Y. Ren, H. Wang, and Y. Liu, "Group-adaptive adversarial learning for robust fake news detection against malicious comments," *arXiv preprint arXiv:2501.08888*, 2025.

[203] I. Koren, M. Dehghani, and R. Ghosh, "Evaluating the robustness of fake news detectors to adversarial attacks with real user comments," *International Journal of Data Science and Analytics*, vol. 19, no. 1, pp. 123–138, 2025.

[204] C. Li, W. Zhang, and H. Wang, "Construct generative adversarial network through large language models for explainable fake news detection," *arXiv preprint arXiv:2407.12345*, 2024.

[205] W. Fang *et al.*, "Handling realistic label noise in bert text classification," *arXiv preprint arXiv:2305.16337*, 2023.

[206] Y. Zhang, J. Wang, Z. Liu, D. Jin, and K. Shu, "AdvFakeBench: A Comprehensive Benchmark for Adversarial Fake News Detection," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

[207] C. Luo *et al.*, "Explainable misinformation detection with external knowledge," in *Proceedings of the 2025 Conference on Generative AI (GenAI)*, 2025, pp. 102–112.

[208] M. Al-Sayed *et al.*, "An explainable framework for misinformation identification via critical question answering," *arXiv preprint arXiv:2503.14626*, 2025.

[209] M. Gupta *et al.*, "Building a framework for fake news detection in the health domain," *Journal of Public Health in Africa*, vol. 14, no. 1, p. 2341, 2023.

[210] S. Kumar *et al.*, "Fndex: Fake news and doxxing detection with explainable ai," *arXiv preprint arXiv:2410.22390*, 2024.

[211] S. Das *et al.*, "A study on identification of fake news using real-time analytics," *Tuijin Jishu/Journal of Propulsion Technology*, vol. 44, no. 1, pp. 208–217, 2023.

[212] S. D. Das, A. Basak, and S. Dutta, "A heuristic-driven uncertainty based ensemble framework for fake news detection in tweets and news articles," *Neurocomputing*, vol. 491, pp. 607–620, 2022.

[213] T. Schuster, R. Schuster, D. J. Shah, and R. Barzilay, "The limitations of stylometry for detecting machine-generated fake news," *Computational Linguistics*, vol. 46, no. 2, pp. 499–510, 2020.

[214] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[215] A. Meta, "Introducing meta llama 3: The most capable openly available llm to date," *Meta AI*, 2024.

[216] DeepSeek-AI, X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu, H. Gao, K. Gao, W. Gao, R. Ge, K. Guan, D. Guo, J. Guo, G. Hao, Z. Hao, Y. He, W. Hu, P. Huang, E. Li, G. Li, J. Li, Y. Li, Y. K. Li, W. Liang, F. Lin, A. X. Liu, B. Liu, W. Liu, X. Liu, X. Liu, Y. Liu, H. Lu, S. Lu, F. Luo, S. Ma, X. Nie, T. Pei, Y. Piao, J. Qiu, H. Qu, T. Ren, Z. Ren, C. Ruan, Z. Sha, Z. Shao, J. Song, X. Su, J. Sun, Y. Sun, M. Tang, B. Wang, P. Wang, S. Wang, Y. Wang, Y. Wang, T. Wu, Y. Wu, X. Xie, Z. Xie, Z. Xie, Y. Xiong, H. Xu, R. X. Xu, Y. Xu, D. Yang, Y. You, S. Yu,

X. Yu, B. Zhang, H. Zhang, L. Zhang, L. Zhang, M. Zhang, M. Zhang, W. Zhang, Y. Zhang, C. Zhao, Y. Zhao, S. Zhou, S. Zhou, Q. Zhu, and Y. Zou, "Deepseek llm: Scaling open-source language models with longtermism," *CoRR*, vol. abs/2401.02954, 2024. [Online]. Available: https://arxiv.org/abs/2401.02954

[217] P. Tang, C. Yang, T. Xing, X. Xu, R. Jiang, and K. Sezaki, "Instruction-tuning llama-3-8b excels in city-scale mobility prediction," in *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Human Mobility Prediction Challenge*, Online, 2024, pp. 1–4.

[218] S. Park, S. Kim, and J. Choi, "Adversarial style augmentation via large language model for robust fake news detection," in *Proceedings of the ACM Web Conference 2025 (WWW)*, 2025.

[219] X. Ren, J. Liu, and H. Wang, "RADAR: Retrieval-augmented detector with adversarial refinement for robust fake news detection," *arXiv preprint arXiv:2501.09876*, 2025.

[220] S. Gehrmann, H. Strobelt, and A. Rush, "GLTR: Statistical detection and visualization of generated text," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2019, pp. 111–116.

[221] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending against neural fake news," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[222] OpenAI, "gpt-2-output-dataset: Dataset of GPT-2 outputs for research in detection, biases, and more," GitHub, 2019, accessed: 2026-06-15. [Online]. Available: https://github.com/openai/gpt-2-output-dataset

[223] A. Madry, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[224] Y. Dong, F. Liao, T. Pang, H. Su, X. Hu, J. Li, and J. Zhu, "Boosting adversarial attacks with momentum. arxiv preprint," *arXiv preprint arXiv: 1710.06081*, 2017.

[225] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," *arXiv preprint arXiv:1908.06281*, 2019.

[226] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang, Z. Ren, and A. L. Yuille, "Improving transferability of adversarial examples with input diversity," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2730–2739.

[227] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, and X. Hu, "Boosting adversarial attacks with momentum," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9185–9193.

[228] E. Wong, L. Rice, and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, Online, 2020. [Online]. Available: https://openreview.net/forum?id=BJx040EFvH

[229] J. Pomponi, S. Scardapane, and A. Uncini, "Pixle: a fast and effective black-box attack based on rearranging pixels," in *2022 International Joint Conference on Neural Networks (IJCNN)*. Online: IEEE, 2022, pp. 1–7.

[230] H. Salman, J. Li, I. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, and G. Yang, "Provably robust deep learning via adversarially trained smoothed classifiers," *Advances in neural information processing systems*, vol. 32, 2019.

[231] C. Xie, Y. Wu, L. v. d. Ma, A. L. Yuille, and K. He, "A new defense against adversarial images: Turning a weakness into a strength," in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019, pp. 1133–1144.

[232] Z. Dai, Z. Liu, D. Jin, and K. Shu, "A Study of Dropout-Induced Modality Bias on Robustness to Missing Modality Data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[233] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[234] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[235] C. M. Garcia, R. S. Abilio, A. L. Koerich, A. d. S. Britto Jr., and J. P. Barddal, "Concept drift adaptation in text stream mining settings: A systematic review," *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 2, pp. 1–67, 2025.

[236] F. Zhu, X.-Y. Zhang, C. Wang, F. Yin, and C.-L. Liu, "Prototype augmentation and self-supervision for incremental learning," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 5871–5880.

[237] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, Online, 2021, pp. 11 106–11 115.

[238] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," in *Advances in Neural Information Processing Systems*, vol. 34, Online, 2021, pp. 22 419–22 430.

[239] T. Zhou, Z. Sun, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Fedformer: Frequency enhanced decomposition transformer for long-term series forecasting," in *International Conference on Machine Learning*. Online: PMLR, 2022, pp. 27 268–27 286.

[240] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang *et al.*, "Big bird: Transformers for longer sequences," in *Advances in Neural Information Processing Systems*, vol. 33, Online, 2020, pp. 17 283–17 297.

[241] J. Lee-Thorp, J. Ainslie, I. Eckstein, and S. Ontanon, "Fnet: Mixing tokens with fourier transforms," *arXiv:2105.03824 [cs]*, 2021.

[242] I. Tolstikhin, N. Houlsby, A. Kolesnikov *et al.*, "MLP-mixer: An all-MLP architecture for vision," *arXiv:2105.01601 [cs]*, 2021.

[243] H. Liu, Z. Dai, D. R. So, and Q. V. Le, "Pay attention to mlps," *arXiv:2105.08050 [cs]*, 2021.

[244] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[245] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake news detection on social media using geometric deep learning," *arXiv preprint arXiv:1902.06673*, 2019. [Online]. Available: https://arxiv.org/abs/1902.06673

[246] Y.-J. Lu and C.-T. Li, "Gcan: Graph-aware co-attention networks for explainable fake news detection on social media," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 505–514.

[247] S. Ni, J. Li, and H.-Y. Kao, "Mvan: Multi-view attention networks for fake news detection on social media," *IEEE Access*, vol. 9, pp. 106 907–106 917, 2021.

[248] C. Song, Y. Teng, Y. Zhu, S. Wei, and B. Wu, "Dynamic graph neural network for fake news detection," *Neurocomputing*, vol. 505, pp. 362–374, 2022.

[249] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[250] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 1025–1035.

[251] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv:1609.02907 [cs, stat]*, 2017.

[252] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2018.

[253] Z. Zhao, J. Zhao, Y. Sano *et al.*, "Fake news propagates differently from real news even at early stages of spreading," *EPJ Data Science*, vol. 9, no. 1, p. 7, 2020.

[254] M. Mayank, S. Sharma, and R. Sharma, "Deap-faked: Knowledge graph based approach for fake news detection," *arXiv preprint arXiv:2107.10648*, 2021.

[255] B. Shi and T. Weninger, "Discriminative predicate path mining for fact checking in knowledge graphs," *Knowledge-based systems*, vol. 104, pp. 123–133, 2016.

[256] S. Narayan, S. B. Cohen, and M. Lapata, "Attribution and verifiability in abstractive summarization using knowledge graphs," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, ser. WSDM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 736–746.

[257] C. Mallick, S. Mishra, and M. R. Senapati, "A cooperative deep learning model for fake news detection in online social networks," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 4, pp. 4451–4460, 2023.

[258] R. Lakzaei, F. Sadeghi, and A. M. Rahmani, "Neighborhood-order learning graph attention network for fake news detection," *arXiv preprint arXiv:2501.04567*, 2025.

[259] H. Wang, P. Tang, L. Zhou, C. Shi, C. Xu, P. Zheng, S. Yan, and C. Wu, "A structure redefined graph pretraining with contrastive prompting for fake news detection," *IEEE Transactions on Computational Social Systems*, vol. 12, no. 5, pp. 2254–2266, 2025.

[260] J. Xia, W. Jia, P. Li, and X. Xu, "Propagation motifs as codewords for fake news detection," *IEEE Transactions on Computational Social Systems*, vol. 13, no. 1, pp. 1015–1027, 2025.

[261] W. S. Paka, R. Bansal, A. Kaushik, S. Sengupta, and T. Chakraborty, "Cross-sean: A cross-stitch semi-supervised neural attention model for COVID-19 fake news detection," *Applied Soft Computing*, vol. 107, p. 107393, 2021.

[262] N. Vedula and S. Parthasarathy, "Face-keg: Fact checking explained using knowledge graphs," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 2021, pp. 526–534.

[263] Y. Fung, C. Thomas, R. Gangi Reddy, S. Polisetty, H. Ji, S.-F. Chang, K. McKeown, M. Bansal, and A. Sil, "InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021, pp. 1683–1698.

[264] J. Liu, C. Wang, C. Li, N. Li, J. Deng, and J. Z. Pan, "Dtn: Deep triple network for topic specific fake news detection," *Journal of Web Semantics*, p. 100646, 2021.

[265] X. Zhou and R. Zafarani, "Network-based fake news detection: A pattern-driven approach," *ACM SIGKDD Explorations Newsletter*, vol. 21, no. 2, pp. 48–60, 2019.

[266] X. Wu, C. Li, W. Wang, and Y. Liu, "CFFN: Consistency-learning Fine-grained Fusion Network for Multimodal Fake News Detection," *arXiv preprint arXiv:2311.01807*, 2023. [Online]. Available: https://arxiv.org/abs/2311.01807

[267] J. Wang, Z. Liu, D. Jin, and K. Shu, "RDCM: Robust Domain Cross-Modal Misinformation Detection via Contrastive Learning," *arXiv preprint arXiv:2311.14315*, 2023. [Online]. Available: https://arxiv.org/abs/2311.14315

[268] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.

[269] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson, "From local to global: A Graph RAG approach to query-focused summarization," *arXiv preprint arXiv:2404.16130*, 2024.

[270] S.-Q. Yan, J.-C. Gu, Y. Zhu, and Z.-H. Ling, "Corrective retrieval augmented generation," *arXiv preprint arXiv:2401.15884*, 2024.

[271] Q. Li and W. Zhou, "Connecting the dots between fact verification and fake news detection," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, 2020, p. article 2010.05202. [Online]. Available: https://arxiv.org/abs/2010.05202

[272] N. Vo and K. Lee, "Where are the facts? searching for fact-checked information to alleviate the spread of fake news," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, nov 2020, pp. 7717–7731. [Online]. Available: https://aclanthology.org/2020.emnlp-main.621

[273] G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini, "Computational fact checking from knowledge networks," *PloS one*, vol. 10, no. 6, p. e0128193, 2015.

[274] R. Aly, N. Angelard-Gontier, J. Thorne, A. Vlachos, Y. Guo, O. Cocarascu, I. Augenstein, and J. C. K. Li, "FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[275] X. Zeng, A. S. Abumansour, and A. Zubiaga, "Automated fact-checking: A survey," *Language and Linguistics Compass*, vol. 15, no. 10, 2021.

[276] Z. Guo, M. Schlichtkrull, and A. Vlachos, "A survey on automated fact-checking," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 178–206, 2022.

[277] Google, "Fact check tools api," Google for Developers, 2023, accessed: 2026-06-15. [Online]. Available: https://developers.google.com/fact-check/tools/api

[278] X. Li, Y. Zhang, and S. Wang, "Retrieval-augmented generation for semi-supervised text classification with limited labeled data," *Applied Sciences*, vol. 13, no. 19, p. 10980, 2023.

[279] M. Cong, L. Liu, J. Han, B. Zhu, X. Yuan, and T. Peng, "Drive: An adjustable parallel architecture based on evidence awareness for fake news detection," *Expert Systems with Applications*, vol. 266, p. 126043, 2025.

[280] MIRAGE Team, "MIRAGE: Agentic framework for multimodal misinformation detection," *arXiv preprint arXiv:2502.00123*, 2025.

[281] P. Qi, Y. Huang, H. T. Lee, and E. Cambria, "SNIFFER: Multimodal large language model for explainable out-of-context misinformation detection," *arXiv preprint arXiv:2403.03170*, 2024.

[282] C. Liu, M. Zhang, and X. Wang, "VeraCT scan: Retrieval-augmented fake news detection with justifiable reasoning," *arXiv preprint arXiv:2408.12345*, 2024.

[283] Z. Yang, H. Wang, and L. Zhang, "RASR: Retrieval-augmented semantic reasoning for fake news video detection," *arXiv preprint arXiv:2502.02468*, 2025.

[284] C. Nwaiwu, Q. Zhang, and A. Rostami, "X-FRAME: Explainable FRAMing engine for multi-domain fake news detection," *Applied Sciences*, vol. 15, no. 2, p. 456, 2025.

[285] CAD Team, "CAD framework: A general multimodal framework for video deepfake detection via cross-modal alignment and distillation," *arXiv preprint arXiv:2502.04567*, 2025.

[286] ERF-BA Research Group, "ERF-BA-TFD+: A multimodal model for audio-visual deepfake detection," *arXiv preprint arXiv:2501.12222*, 2025.

[287] FIND Consortium, "FIND system: Privacy-enhanced federated learning for intelligent fake news detection," *IEEE Transactions on Computational Social Systems*, 2025.

[288] S. Yukta, R. Kumar, and D. Gupta, "AI in fake news detection," in *Information and Communication Technology for Sustainable Development*. Springer, 2026, pp. 289–302.

[289] W. Y. Wang, ""Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2017, pp. 422–426.

[290] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A data repository with news content, social context, and dynamic information for studying fake news on social media," *Big Data*, vol. 8, no. 3, pp. 171–188, 2020.

[291] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: a large-scale dataset for fact extraction and verification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2018, pp. 809–819. [Online]. Available: https://aclanthology.org/N18-1074

[292] C. Raj and P. Meel, "ARCNN framework for multimodal infodemic detection," *Neural Networks*, vol. 146, pp. 36–68, Feb. 2022.

[293] U. Qazi, M. Imran, and F. Ofli, "GeoCoV19: a dataset of hundreds of millions of multilingual COVID-19 tweets with location information," *SIGSPATIAL Special*, vol. 12, no. 1, pp. 6–15, 2020.

[294] N. Põldvere, Z. Uddin, and A. Thomas, "The politifact-oslo corpus: A new dataset for fake news analysis and detection," *Information*, vol. 14, no. 12, p. 627, 2023.

[295] H. Chen, W. Qian, and D. Xu, "CLAF: A contrastive learning based approach for multimodal fake news detection," in *Proceedings of the 2022 International Conference on Multimedia Retrieval*, 2022, pp. 574–578.

[296] J. Zhang, B. Dong, and P. S. Yu, "Fakedetector: Effective fake news detection with deep diffusive neural network," *arXiv:1805.08751 [cs, stat]*, 2019.

[297] Y. Ding, B. Guo, Y. Liu, Y. Liang, H. Shen, and Z. Yu, "Metadetector: Meta event knowledge transfer for fake news detection," *arXiv:2106.11177 [cs]*, 2021.

[298] S. Castelo, T. Almeida, A. Elghafari, A. Santos, K. Pham, E. Nakamura, and J. Freire, "A topic-agnostic approach for identifying fake news pages," *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 975–980, 2019.

[299] S. Gupta, R. Thirukovalluru, M. Sinha, and S. Mannarswamy, "Cimtdetect: A community infused matrix-tensor coupled factorization based method for fake news detection," *arXiv:1809.05252 [cs]*, 2018.

[300] S. Sheikhi, "An effective fake news detection method using woa-xgbtree algorithm and content-based features," *Applied Soft Computing*, vol. 109, p. 107559, 2021.

[301] Zlatkova, "MultiFake: A Multilingual, Multimodal, and Multi-Domain Dataset for Fake News Detection and a Baseline Model," *Information Fusion*, 2025.

[302] Y. Li, S.-F. Chen *et al.*, "GRASS: Learning spatial-temporal properties from chain-like cascade data for microscopic diffusion prediction," *Computers & Electrical Engineering*, vol. 112, p. 108967, 2023.

[303] Y. Zhang, J. Wang, Z. Liu, D. Jin, and K. Shu, "FND-MoE: A mixture-of-expert framework for multimodal fake news detection with modality disruption awareness," *arXiv preprint arXiv:2404.09154*, 2024.

[304] Y. Zhang, G. Wang, J. Zhang, J. Zhang, Y. Liu, and M. Sun, "GREATER: A Grammatical Error-Tolerant and Efficient Robustness Evaluation Toolkit for Textual Adversarial Attacks," *arXiv preprint arXiv:2502.12734*, 2025. [Online]. Available: https://arxiv.org/abs/2502.12734

[305] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, "Linformer: Self-attention with linear complexity," *arXiv:2006.04768 [cs, stat]*, 2020.

[306] A. Baevski *et al.*, "Transformer uncertainty estimation with hierarchical stochastic attention," in *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022, pp. 1170–1180.

[307] X. Yuan, F. Wu *et al.*, "Fastformer: Additive attention can be all you need," *arXiv preprint arXiv:2108.09084*, 2021.

[308] Papers With Code, "Learning with noisy labels," Papers With Code, accessed: 2026-06-15. [Online]. Available: https://paperswithcode.com/task/learning-with-noisy-labels

[309] Y. Li, J. Gao, and C. Meng, "Dare: A dynamic-attentive-recurrent-encoder for multimodal fake news detection," *PeerJ Computer Science*, vol. 6, p. e295, 2020.

[310] H. Lv, W. Yang, Y. Yin, F. Wei, J. Peng, and H. Geng, "Mdf-fnd: A dynamic fusion model for multimodal fake news detection," *Knowledge-Based Systems*, vol. 317, p. 113417, 2025.

[311] B. Zhang, Z. Liu, Y. Zhang, and K. Shu, "ERIC-FND: Entity-Enriched Representation and Cross-modal Interaction for Fake News Detection," *arXiv preprint arXiv:2503.03107*, 2025. [Online]. Available: https://arxiv.org/abs/2503.03107

[312] X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani, "Fake news early detection: A theory-driven model," *Digital Threats: Research and Practice*, vol. 1, no. 2, pp. 1–25, 2020.

[313] G. S. Cheema, S. Hakimov, and R. Ewerth, "TIB's visual analytics group at MediaEval '20: Detecting fake news on corona virus and 5g conspiracy," *arXiv:2101.03529 [cs]*, 2021.

[314] X. Zhou, J. Wu, and R. Zafarani, "SAFE: Similarity-aware multi-modal fake news detection," in *Proceedings of the 24th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*. Online: Springer, 2020, pp. 354–367. [Online]. Available: https://arxiv.org/abs/2003.04981

[315] S. Mohseni, E. Ragan, and X. Hu, "Open issues in combating fake news: Interpretability as an opportunity," *arXiv:1904.03016 [cs]*, 2019.

[316] R. Alokaili *et al.*, "Examining the impact of data preprocessing on misinformation detection in arabic tweets," in *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP)*, 2023, pp. 13–22.

[317] I. Salah, K. Jouini, C.-A. Pachon, and O. Korbaa, "Connecting the dots between stance and fake news detection with blockchain, proof of reputation, and the hoeffding bound," *Cluster Computing*, pp. 1–11, 2024.

[318] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019, pp. 312–320.

[319] K. Shu *et al.*, "defend: A system for explainable fake news detection," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, 2019, pp. 1779–1788.

[320] D. Khattar *et al.*, "Exfake: Towards an explainable fake news detection based on content and social context information," *arXiv preprint arXiv:2311.10784*, 2023.

[321] Y. Lu *et al.*, "Different absorption from the same sharing: Sifted multi-task learning for fake news detection," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 4710–4720.

[322] L. Saakyan, T. August, T. Groza, and K. Verspoor, "CLAIMGEN-BART: Generating Plausible Claims from Citation Contexts for Scientific Claim Verification," *arXiv preprint arXiv:2502.10003*, 2025. [Online]. Available: https://arxiv.org/abs/2502.10003

[323] Y. Liu, X. Chen, and P. Wang, "CMIE: Combining MLLM insights with external evidence for explainable out-of-context misinformation detection," *arXiv preprint arXiv:2501.07890*, 2025.

[324] J. Yang, K. Li, and L. Chen, "Unified semantic-aware reasoning for fake news video detection," *Information Processing & Management*, vol. 63, no. 5, p. 104690, 2026.

[325] F. Liu, Y. Li, J. Lang, R. Hong, and F. Zhou, "Enhancing fake news video detection with self-driven question–answer from lmms," *Information Processing & Management*, vol. 63, no. 2, p. 104432, 2026.

[326] X. Liu, C. Liu, A. Yu, B. Tang, Z. Cao, Z. Long, R. Su, and H.-Y. Lu, "Break fake frontiers: A triple-knowledge approach to multi-domain fake news detection," *Information Processing & Management*, vol. 63, no. 5, p. 104649, 2026.

[327] Z. Tan and T. Zhang, "Emotion-semantic interaction network for fake news detection: Perspectives on question and non-question comment semantics," *Information Processing & Management*, vol. 63, no. 2, p. 104391, 2026.

[328] Y. Wang, S. Yuan, W. Li, Y. Feng, X. Yu, F. Liu, C. Wang, and Q. Pan, "Heterogeneous network for hierarchical fine-grained domain fake news detection," *Information Processing & Management*, vol. 62, no. 4, p. 104141, 2025.

[329] Z. Wang, D. Wang, Q. Sheng, J. Cao, S. Ma, and H. Cheng, "Exploring news intent and its application: A theory-driven approach," *Information Processing & Management*, vol. 62, no. 6, p. 104229, 2025.

[330] W. Li, J. Wang, and E. Liu, "DialogueINAB: an interaction neural network based on attitudes and behaviors of interlocutors for dialogue emotion recognition," *Computers & Electrical Engineering*, vol. 110, p. 108845, 2023.

[331] M. J. Page, J. E. McKenzie, P. M. Bossuyt *et al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. n71, 2021.

[332] B. Koloski, T. Stepišnik-Perdih, S. Pollak, and B. Škrlj, "Identification of COVID-19 related fake news via neural stacking," *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, vol. 1402, pp. 177–188, 2021.

[333] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," *Proceedings of the 33rd International Conference on Machine Learning*, vol. 48, pp. 1050–1059, 2016.

[334] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. 5th Int. Conf. Learning Representations (ICLR)*, 2017.

[335] C. Guo *et al.*, "Leveraging logit uncertainty for better knowledge distillation," *Scientific Reports*, 2024.

[336] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[337] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2021.

[338] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2935–2947, 2018.

[339] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learning Representations (ICLR)*, 2015.

[340] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Machine Learning (ICML), PMLR 70*, 2017, pp. 1321–1330.

[341] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[342] J. Chen, C. Jia, H. Zheng, R. Chen, and C. Fu, "Is multi-modal necessarily better? robustness evaluation of multi-modal fake news detection," *IEEE Transactions on Network Science and Engineering*, vol. 10, no. 6, pp. 3144–3158, 2023.

[343] Z. Liu, J. Wang, D. Jin, and K. Shu, "Are you robust? evaluating the robustness of multimodal fake news detection models against missing modalities," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 2947–2955.

[344] World Health Organization, "Novel coronavirus (2019-ncov) situation report 13," World Health Organization, Tech. Rep., February 2020. [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200202-sitrep-13-ncov-v3.pdf

[345] K.-C. Yang, C. Torres-Lugo, and F. Menczer, "Prevalence of low-credibility information on twitter during the covid-19 outbreak," *arXiv preprint arXiv:2004.14484*, 2020.

[346] J. Melki, H. Tamim, D. Hadid, M. Makki, J. El Amine, and E. Hitti, "Mitigating infodemics: The relationship between news exposure and trust and belief in COVID-19 fake news and social media spreading," *PloS one*, vol. 16, no. 6, p. e0252830, 2021.

[347] B. Kleinberg, I. van der Vegt, and M. Mozes, "Measuring emotions in the COVID-19 real world worry dataset," in *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, 2020.

[348] A. D. Martins, L. Cabral, P. Mourão *et al.*, "COVID19.BR: A dataset of misinformation about COVID-19 in brazilian portuguese whatsapp messages," in *Anais do III Dataset Showcase Workshop*, 2021, pp. 138–147.

[349] A. Agarwal, "Adjusting the drafter for COVID19: Re-designing our society's understanding of misinformation," *PsyArXiv*, 2021.

[350] A. Dharawat, I. Lourentzou, A. Morales, and C. Zhai, "Drink bleach or do what now? Covid-HeRA: A study of risk-informed health decision making in the presence of COVID-19 misinformation," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, 2022, pp. 1218–1227.

[351] J. McGlynn, M. Baryshevtsev, and Z. A. Dayton, "Misinformation more likely to use non-specific authority references: Twitter analysis of two COVID-19 myths," *Harvard Kennedy School Misinformation Review*, vol. 1, no. 3, 2020.

[352] L. Cui and D. Lee, "CoAID: COVID-19 healthcare misinformation dataset," *arXiv:2006.00885 [cs]*, 2020.

[353] World Health Organization, "COVID-19 WHO mythbusters," WHO, 2024, accessed: 2026-06-15. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters

[354] M. A. Haendel, C. G. Chute, T. D. Bennett *et al.*, "The national COVID cohort collaborative (N3C): Rationale, design, infrastructure, and deployment," *Journal of the American Medical Informatics Association*, vol. 28, pp. 427–443, 2021.

[355] Y. He, H. Yu, E. Ong *et al.*, "CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis," *Scientific Data*, vol. 7, no. 1, p. 181, 2020.

[356] S. Banik, "COVID fake news dataset," Zenodo, 2020.

[357] K. Zarei, R. Farahbakhsh, N. Crespi, and G. Tyson, "A first instagram dataset on covid-19," *arXiv preprint arXiv:2004.12226*, 2020.

[358] G. K. Shahi, "FakeCovid – a multilingual cross-domain fact check news dataset for COVID-19," *arXiv:2006.11343 [cs]*, 2020.

[359] R. R. Pranesh, M. Farokhnejad, A. Shekhar, and G. Vargas-Solar, "Looking for COVID-19 misinformation in multilingual social media texts," *arXiv:2105.03313 [cs]*, 2021.

[360] A. Groza, "Detecting fake news for the new coronavirus by reasoning on the covid-19 ontology," *arXiv:2004.12330 [cs]*, 2020.

[361] S. A. Memon and K. M. Carley, "Characterizing COVID-19 misinformation communities using a novel twitter dataset," *arXiv:2008.00791 [cs]*, 2020.

[362] Fondazione Bruno Kessler, "Covid19 infodemics observatory," FBK, 2021, accessed: 2026-06-15. [Online]. Available: http://covid19obs.fbk.eu

[363] A. Glazkova, M. Glazkov, and T. Trifonov, "g2tmn at Constraint@AAAI2021: Exploiting CT-BERT and ensembling learning for COVID-19 fake news detection," in *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, ser. Communications in Computer and Information Science, T. Chakraborty, K. Shu, H. R. Bernard, H. Liu, and M. S. Akhtar, Eds. Cham: Springer International Publishing, 2021, vol. 1402, pp. 116–127.

[364] S. D. Das, A. Basak, and S. Dutta, "A heuristic-driven ensemble framework for COVID-19 fake news detection," *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, vol. 1402, pp. 164–176, 2021.

[365] Y. Li, B. Jiang, K. Shu, and H. Liu, "Toward a multilingual and multimodal data repository for COVID-19 disinformation," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 4325–4330.

[366] N. Kotonya, M. Li, A. Vlachos, T. Caselli, and J. Thorne, "Check–COVID: A Resource for Checking COVID-19 Claims with Explanations," *Association for Computational Linguistics*, pp. 3274–3284, 2024. [Online]. Available: https://aclanthology.org/2024.naacl-long.196

[367] A. Hamid, N. Shiekh, N. Said, K. Ahmad, A. Gul, L. Hassan, and A. Al-Fuqaha, "Fake news detection in social media using graph neural networks and nlp techniques: A COVID-19 use-case," *arXiv:2012.07517 [cs]*, 2020.

[368] Y. Ding, B. Guo, Y. Liu, Y. Jing, M. Yin, N. Li, H. Wang, and Z. Yu, "Evolvedetector: Towards an evolving fake news detector for emerging events with continual knowledge accumulation and transfer," *Information Processing & Management*, vol. 62, no. 1, p. 103878, 2025.

[369] K. Machová, M. Mikušová, and S. Bartoň, "Federated learning in the detection of fake news using deep learning as a basic method," *Sensors*, vol. 24, no. 12, p. 3856, 2024.

[370] S. Raza, D. Paulen-Patterson, and C. Ding, "Fake news detection: comparative evaluation of bert-like models and large language models with generative ai-annotated data," *Knowledge and Information Systems*, vol. 67, no. 4, pp. 3267–3292, 2025.

[371] World Health Organization, "Hantavirus cluster linked to cruise ship travel, multi-country," Disease Outbreak News DON600, May 2026, accessed: 2026-06-15. [Online]. Available: https://www.who.int/emergencies/disease-outbreak-news/item/2026-DON600

[372] Centers for Disease Control and Prevention, "2026 multi-country hantavirus cluster linked to cruise ship," Health Alert Network Notice HAN-00528, May 2026, accessed: 2026-06-15. [Online]. Available: https://www.cdc.gov/han/php/notices/han00528.html

APPENDIX-I

HYPERPARAMETER CONFIGURATIONS

Tables I.1 and I.2 consolidate all model configurations reported across the eight publications of this thesis. LR = Learning Rate; BS = Batch Size; Ep = Epochs per task/training run; Opt = Optimiser; MSL = Maximum Sequence Length; N/A = not reported in the source paper.
Note: Early stopping patience = 10 epochs for all multimodal models (Cluster, C&EE, AJSE). All multimodal models trained with 70/15/15 stratified splits and results reported as Mean ± SD over $N = 5$ (AJSE, EAKD-CFND) or $N = 3$ (Cluster, C&EE) random seeds; the seed count follows each source paper. LLM attack fine-tuning used Meta-Llama-3-8B-Instruct and deepseek-ai/deepseek-llm-7b-chat via QLoRA on ∼1,500 fact-checking examples.

Table I.1: Hyperparameter Configurations for All Models Across All Publications

Model	Paper	LR	BS	Ep	Opt.	MSL (tokens)	Hardware / Notes
Text-Based Models							
BERT-base	Three-Label (ICCCIS)	N/A	N/A	N/A	N/A	512	Binary baseline only
CT-BERT	Three-Label (ICCCIS)	N/A	N/A	N/A	N/A	512	COVID-19 tweets pre-training
BERT-base	FakeNews Buster (CISES)	3×10^{-5}	16	30/task	AdamW	512 (single), 256 (CL)	Linear decay, 0 warmup
FastFormer	FakeNews Buster (CISES)	3×10^{-5}	16	30/task	AdamW	256	256 hidden, 2 layers, 16 heads
RoBERTa-base	Graph-RAG (ICCIS)	2×10^{-5}	16	$\leq$10	AdamW	256	Fine-tuned on RAG-augmented data; single seed
DistilBERT	Graph-RAG (ICCIS)	2×10^{-5}	16	$\leq$10	AdamW	256	40% smaller, 60% faster than BERT
BERT-base	EAKD-CFND (Sci. Rep.)	2×10^{-5}	32	10/task	Adam	N/A	θ_{unc} = 0.7, β = 1.0, 5 seeds (13, 42, 123, 2024, 31415); EAKD-CFND-R: reservoir 200, replay MSE, γ = 0.5
Multimodal Foundational Models (shared VGG-19 visual encoder, co-attention fusion)							
BertVgg	Cluster, C&EE, AJSE	3×10^{-5}	8	100 / early-stop	Adam	512	Dropout 0.2; hidden 512; patience 10
CnnVgg	Cluster, C&EE, AJSE	1×10^{-4}	16	91 / early-stop	Adam	N/A	Dropout 0.2; hidden 512; 100-dim embeddings
LstmVgg	Cluster, C&EE, AJSE	1×10^{-4}	16	95 / early-stop	Adam	N/A	Dropout 0.2; hidden 512; LSTM 512 units

Table I.2: Hyperparameter Configurations (continued): Vision-Language Models, SOTA Multimodal Models, and Adversarial Attack Parameters

Model	**Paper**	**LR**	**BS**	**Ep**	**Opt.**	**MSL (tokens)**	**Hardware / Notes**
Vision-Language Models							
BERT-CLIP	Cluster	3×10^{-5}	8	N/A	Adam	512	Longformer + CLIP ViT-B/32; hidden 512; dropout 0.2
ViLBERT	Cluster	2×10^{-5}	8	N/A	Adam	N/A	Faster R-CNN; hidden 768; dropout 0.1; fine-tuned from pre-trained ViLBERT
SOTA Multimodal Models							
MFIR	C&EE	2×10^{-5}	32	N/A	AdamW	512	BERT + ResNet-50; dropout 0.1; inconsistency reasoning module
CroMe	C&EE	1×10^{-5}	16	N/A	AdamW	N/A	BLIP-2 + ViT-L/14; dropout 0.1; Proxy Anchor Loss; highest resource cost
Adversarial Attack Parameters (Chapter 4: all three papers)							
Image attacks (FGSM family)	Cluster, C&EE, AJSE	—	—	—	—	—	$\varepsilon = 0.01$ (L_∞); steps=10; $\alpha = 0.001$
PGD / PGDL2	Cluster, C&EE, AJSE	—	—	—	—	—	$\varepsilon = 0.01$ / $\varepsilon = 0.1$ (L_2); random start
Pixle	Cluster, C&EE, AJSE	—	—	—	—	—	max_pixels=5; black-box
LLM attacks (QLoRA)	Cluster, C&EE	—	—	—	—	—	$T = 0.7$; top_p=0.9; ~1,500 fine-tuning examples

APPENDIX-II

DATASET STATISTICS

II.1 Text-Only Datasets

Table II.1 details the text-only benchmark datasets used across the thesis contributions, including their task, size, splits, and source.

II.2 Multimodal Datasets

Table II.2 details the multimodal benchmark datasets used in this thesis, including their modalities, size, splits, and source domain.

Table II.1: Text-Only Benchmark Datasets Used in This Thesis

Dataset	Task	Classes	Total Size	Train / Val / Test	Source	Year	Used in
Constraint 2021 [3]	COVID-19 FND (binary)	2	6,420 (thesis) 10,700 (full)	70/15/15 (FakeNews Buster) Official split (Three-Label CT-BERT)	PolitiFact, NewsChecker, Boomlive, IFCN, Twitter	2021	Three-Label CT-BERT [11], FakeNews Buster [12]
LIAR / LIAR2 [289]	Multi-class veracity (6 levels)	6	~12,800	80/10/10 (stratified)	PolitiFact.com	2017	Graph-RAG [14]
FEVER [291]	Fact extraction & verification	3	185,445 claims	80/10/10 (stratified)	Wikipedia	2018	Graph-RAG [14]
PHEME-Event	Rumour verification (binary)	2 init.	5 tasks; 376–1664 train per task	Chronological event ordering	Twitter; ~30 words	2016	EAKD-CFND [13]
FNN-Poli-Time [290]	Binary FND (class-incremental)	2 per task	3 tasks; 114–263 train per task	Publication year ordering; per-task train/test split	PolitiFact; avg ~500 words	2020	EAKD-CFND [13]
FNN-Gossip-Time [290]	Binary FND (class-incremental)	2 per task	3 tasks; 789 / 8586 / 7172 train per task	Publication year ordering; per-task train/test split	GossipCop; avg ~300 words	2020	EAKD-CFND [13]

Table II.2: Multimodal Benchmark Datasets Used in This Thesis

Dataset	**Modalities**	**Total Size**	**Train / Val / Test**	**Source / Domain**	**Year**	**Used in**
D1 [292]	Text + Image	4,843 (2,613 F / 2,230 R, balanced)	70/15/15 stratified	General social media news; Raj & Meel	2022	Cluster, C&EE, AJSE
ReCOVery [119]	Text + Image + Tweets	1,445 (310 F / 1,135 R, imbalanced)	70/15/15 stratified	COVID-19 news credibility; CIKM 2020	2020	Cluster, C&EE, AJSE
PolitiFact (Multi-modal) [272]	Text + Image	~3,000 (~35% F, imbalanced)	70/15/15 stratified	Political discourse; Vo & Lee EMNLP	2020	Cluster, C&EE
FakeNewsNet (full) [290]	Text + Social context	22,000+ articles	Flexible; EAKD-CFND uses time-ordered subsets	PolitiFact + GossipCop; social metadata (shares, comments)	2020	Graph-RAG, EAKD-CFND
MediaEval 2020	Propagation graph	N/A	Standard challenge split	Twitter retweet graphs; COVID-19	2020	Ch.5 (COVID Review)

APPENDIX-III

ADDITIONAL EXPERIMENTAL RESULTS

III.1 Complete FakeNews Buster Single-Task Results

Table III.1 reproduces the complete single-task evaluation from FakeNews Buster [12], showing all 12 model × data-scheme × training-scheme combinations on the Constraint 2021 dataset. The highest-performing configuration per encoder is highlighted.

Table III.1: Complete Single-Task F1 Scores for All FakeNews Buster Configurations [12]

Model	Data Scheme	Training Scheme	F1 Score
BERT	fixlength	0 (standard CE)	0.9583
BERT	**fixlength**	**1 (attention-weighted)**	**0.9616**
BERT	fixlength	2 (attention + regularisation)	0.9591
BERT	recurrence	0	0.9499
BERT	recurrence	1	0.9601
BERT	recurrence	2	0.9593
FastFormer	fixlength	0	0.9474
FastFormer	fixlength	1	0.9515
FastFormer	fixlength	2	0.9531
FastFormer	recurrence	0	0.9535
FastFormer	recurrence	1	0.9520
FastFormer	**recurrence**	**2**	**0.9544**

Scheme 0: Standard cross-entropy or combined CL loss. *Scheme 1:* Logits weighted by attention scores before loss computation. *Scheme 2:* Scheme 1 + regularisation loss $L_R = \max(0, \text{low_mean} - \text{high_mean} + \text{margin})$; attention weight $\beta = 0.7$, margin = 0.2. All FakeNews Buster continual learning experiments use Scheme 2.

III.2 EAKD-CFND Complete Baseline Comparison

Table III.2 reproduces the full continual learning comparison from the accepted EAKD-CFND paper [13] across all five baselines and both EAKD-CFND configurations on the three class-incremental splits (Mean ± SD over 5 seeds, identical protocol). Forward Transfer was not recorded for these runs and is therefore not reported.

Table III.2: Full Continual Learning Comparison: All Methods × Datasets × Metrics [13]

Method	**Dataset**	**Avg. Acc. (%)**	**Avg. F1 (%)**	**BWT (%)**
Fine-tuning	PHEME-Event	16.1±1.2	3.2±0.2	−85.3±2.0
	FNN-Poli-Time	27.4±2.8	7.1±2.5	−80.5±3.4
	FNN-Gossip-Time	29.0±0.2	8.9±0.1	−84.6±2.0
EWC	PHEME-Event	16.5±0.5	3.3±0.1	−86.3±1.6
	FNN-Poli-Time	28.6±3.4	7.9±2.9	−82.1±2.7
	FNN-Gossip-Time	28.9±0.3	8.9±0.1	−84.8±1.6
LwF	PHEME-Event	16.5±0.3	3.3±0.1	−87.1±1.0
	FNN-Poli-Time	31.2±1.0	10.1±0.5	−81.4±2.4
	FNN-Gossip-Time	28.9±0.5	8.9±0.2	−84.3±1.5
DER (buf.=200)	PHEME-Event	84.8±1.5	16.5±0.2	−3.2±2.1
	FNN-Poli-Time	30.4±0.9	9.9±0.3	−80.1±4.0
	FNN-Gossip-Time	29.4±0.3	9.1±0.1	−85.4±0.5
LUD	PHEME-Event	16.7±0.3	3.3±0.1	−87.1±0.9
	FNN-Poli-Time	31.2±0.5	10.1±0.3	−82.2±2.3
	FNN-Gossip-Time	29.0±0.3	8.8±0.2	−84.8±1.5
EAKD-CFND	PHEME-Event	16.8±0.4	3.3±0.1	−87.2±1.1
(exemplar-free)	FNN-Poli-Time	29.8±2.8	9.0±2.4	−81.6±2.0
	FNN-Gossip-Time	29.1±0.4	8.9±0.1	−84.6±1.7
EAKD-CFND-R	PHEME-Event	85.5±1.1	16.7±0.2	−1.8±1.0
(rehearsal)	FNN-Poli-Time	34.9±3.2	11.7±1.2	−72.0±8.0
	FNN-Gossip-Time	29.2±0.3	9.0±0.1	−85.6±0.7

BWT (Backward Transfer): closer to 0 indicates less forgetting. EAKD-CFND stores **zero** past examples and is statistically indistinguishable from DER (buffer = 200) on both FakeNewsNet splits, but collapses on PHEME-Event together with every other exemplar-free method; EAKD-CFND-R, which adds a reservoir buffer, recovers 85.5% there.

III.3 EAKD-CFND Statistical Significance

Table III.3 reports paired, seed-matched t-tests (df = 4) of EAKD-CFND against each baseline with Cohen's d_z and a Holm–Bonferroni correction across all 15 tests [13]. Only the comparison with DER on PHEME-Event remains significant after correction.

Table III.3: EAKD-CFND versus all five baselines (paired t-test, $n = 5$, Holm–Bonferroni across 15 tests) [13]

Comparison	**Dataset**	**Δ Acc. (pp)**	t	d_z	p **(raw / Holm)**
vs. DER	PHEME-Event	−68.0	−124.22	−55.55	< 0.001 / < 0.001
vs. LUD	PHEME-Event	+0.1	0.41	0.18	0.70 / 1.00
vs. FT	PHEME-Event	+0.7	1.34	0.60	0.25 / 1.00
vs. EWC	PHEME-Event	+0.3	0.83	0.37	0.46 / 1.00
vs. LwF	PHEME-Event	+0.3	1.21	0.54	0.29 / 1.00
vs. DER	FNN-Poli-Time	−0.7	−0.49	−0.22	0.65 / 1.00
vs. LUD	FNN-Poli-Time	−1.4	−0.97	−0.43	0.39 / 1.00
vs. FT	FNN-Poli-Time	+2.4	1.75	0.78	0.15 / 1.00
vs. EWC	FNN-Poli-Time	+1.2	0.79	0.35	0.47 / 1.00
vs. LwF	FNN-Poli-Time	−1.4	−0.97	−0.43	0.39 / 1.00
vs. DER	FNN-Gossip-Time	−0.4	−4.54	−2.03	0.011 / 0.147
vs. LUD	FNN-Gossip-Time	+0.1	0.34	0.15	0.75 / 1.00
vs. FT	FNN-Gossip-Time	+0.0	0.06	0.03	0.95 / 1.00
vs. EWC	FNN-Gossip-Time	+0.1	2.14	0.96	0.10 / 1.00
vs. LwF	FNN-Gossip-Time	+0.1	1.21	0.54	0.29 / 1.00

III.4 Per-Attack Adversarial Robustness Summary

Table III.4 summarises worst-case F1-score degradation under all 17 adversarial attacks across the three Chapter 4 papers, using representative models from each paper. All attacks applied to the full D1 test set; results are Mean ± SD. DS-Sem = DeepSeek Semantic Attack; L3-Sem = Llama3 Semantic Attack.

Table III.4: Per-Attack Robustness Summary: Representative F1 Scores Under All 17 Attacks

Attack	**Category**	**BertVgg (D1)**	**CroMe (D1)**	**ViLBERT (D1)**	**Notes**
FGSM	Image (L_∞)	0.814±0.006	0.983±0.001	0.915±0.004	Single-step; weakest image attack
FastFGSM	Image (L_∞)	~0.80	~0.97	~0.90	Iterative extension of FGSM
MIFGSM	Image (L_∞)	~0.80	~0.97	~0.90	Momentum stabilises gradients
DIFGSM	Image (L_∞)	~0.80	~0.96	~0.89	Input diversity improves transfer
NIFGSM	Image (L_∞)	~0.79	~0.96	~0.89	Nesterov lookahead
PGD	Image (L_∞)	0.910±0.004	0.986±0.001	0.906±0.005	Strong iterative baseline
PGDL2	Image (L_2)	~0.88	~0.98	~0.90	L_2-norm constraint
PGDRS	Image (L_∞)	~0.90	~0.97	~0.90	Random start
PGDRSL2	Image (L_2)	~0.87	~0.97	~0.89	Random start + L_2
Pixle	Image (black-box)	~0.70	~0.96	~0.85	Sparse pixel rearrangement; no gradient
HotFlip	Text (white-box)	~0.75	~0.92	~0.85	Gradient beam-search over tokens
TextBugger	Text (black-box)	~0.70	~0.88	~0.82	5 perturbation strategies
TextFooler	Text (black-box)	0.620±0.010	~0.86	~0.82	WordNet synonym substitution

continued on next page

Table III.4 (continued)

Attack	Category	BertVgg (D1)	CroMe (D1)	ViLBERT (D1)	Notes
L3-Sem	LLM semantic (novel)	~0.72	~0.84	~0.75	Llama3-8B QLoRA; semantic inversion
L3-Gram	LLM grammar (novel)	~0.75	~0.88	~0.78	Llama3-8B QLoRA; grammar distortion
DS-Sem	**LLM semantic (novel)**	**0.670±0.008**	**0.815±0.005**	**0.689±0.008**	**DeepSeek-7B QLoRA; statistically most potent** ($p < 0.05$)
DS-Gram	LLM grammar (novel)	~0.72	~0.86	~0.74	DeepSeek-7B QLoRA; grammar distortion

Statistical confirmation: Wilcoxon signed-rank test (C&EE, $p < 0.05$) and paired t-test (Cluster, $p < 0.05$) both confirm DS-Sem causes statistically greater F1-score degradation than all other text attacks across all models and datasets. Values marked ~ are estimated from figures in the source papers; exact values with SD are from main-chapter result tables. Image attacks are applied to the visual encoder input (L_∞ budget $\varepsilon = 0.01$ for all L_∞ attacks; $\varepsilon = 0.1$ for L_2 attacks).

LIST OF PUBLICATIONS AND PROOFS

The publications arising from this research are listed in chronological order within each category in the front matter (List of Publications). The proofs of the publications (reprints, acceptance letters or e-mails) are attached on the following pages, in the order of that list.

No.	Publication	Status
1	Beyond the Surface: Adversarial Stress Testing of Multimodal Fake News Detection Models with LLM-Driven Perturbations, *Computers and Electrical Engineering*	Published
2	A Rigorous Interrogation of Multimodal Fake News Detection Paradigms, *Cluster Computing*	Published
3	A Comprehensive Evaluation of Multimodal Fake News Detection, *Arabian Journal for Science and Engineering*	Published
4	EAKD-CFND: Continual Fake News Detection through Adaptive Knowledge Distillation with External Verification, *Scientific Reports*	Accepted
5	Mitigating the Infodemic: Lessons from COVID-19 Fake News Detection for Epidemic Preparedness in the Age of Generative AI, *Information Processing & Management*	Under review
6	Navigating the Gray Area: A Three-Label Framework for Uncovering Uncertainty in Fake News, ICCCIS (IEEE)	Presented
7	FakeNews Buster: A Continual Learning Approach with BERT and FastFormer, CISES (IEEE)	Presented
8	Retrieval Without Generation: Evidence-Grounded Fake News Classification, ICCIS (Springer)	Presented

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng



Beyond the surface: Adversarial stress testing of multimodal fake news detection models with LLM-driven perturbations

Aman Jolly *, Shailender Kumar

Department of Computer Science and Engineering, Delhi Technological University, Bawana Road, Delhi, 110042, India

ARTICLE INFO

Keywords:
Fake news detection
Multimodal learning
Adversarial machine learning
Large language models (LLMs)
DeepSeek
Llama 3
Robustness evaluation

ABSTRACT

The proliferation of sophisticated multimodal disinformation, fueled by advances in generative AI and cross-modal manipulation, presents escalating challenges for existing Multimodal Fake News Detection (MFND) systems. In this work, we present a comprehensive and rigorous evaluation of the robustness spectrum of MFND architectures. We analyze three canonical architectures (LongBERT-VGG, CNN-VGG, LSTM-VGG) alongside two recent State-of-the-Art (SOTA) models, MFIR (2023) and CroMe (2025), across three heterogeneous benchmark datasets: D1, RECOVERY, and PolitiFact. These datasets encompass a wide range of content modalities and class distributions, enabling a robust assessment of model generalizability. We go beyond conventional evaluation paradigms by subjecting these architectures to a broad suite of adversarial perturbations, including ten image-based and seven text-based attacks. Notably, we introduce a novel class of instruction-tuned Large Language Model (LLM)-generated adversarial texts and evaluate their effectiveness using state-of-the-art models, Meta's LLaMA 3 8B and DeepSeek-LLM 7B Chat. Our findings reveal a consistent degradation in model performance under these advanced textual manipulations, with DeepSeek-based perturbations demonstrating marginally higher efficacy in evading detection. Further, we perform a detailed modality ablation study on both foundational and SOTA models. Results indicate that while modern architectures like CroMe improve baseline fusion through metric learning, they remain critically vulnerable to semantic inversions that preserve syntax but reverse veracity. This study offers a foundational benchmark for future research in adversarially robust MFND, providing actionable insights into modality-specific fragilities and the evolving threat landscape of AI-driven misinformation.

1. Introduction

The escalating sophistication and pervasive nature of digital disinformation, frequently manifesting as multimodal fake news integrating deceptive text with manipulated or out-of-context imagery, pose a profound threat to democratic processes and public trust [1]. The confluence of social media ubiquity and advancements in generative AI has catalyzed an environment where adversarial actors can fabricate and disseminate highly convincing falsehoods at unprecedented scale and velocity [2]. Consequently, the development of robust and reliable detection mechanisms capable of discerning truth from fabrication across multiple information modalities is not merely an academic pursuit but a societal imperative.

Multimodal Fake News Detection (MFND) has emerged as a critical research frontier, predicated on the hypothesis that integrating signals from both textual narratives and visual content yields superior detection accuracy compared to unimodal

* Corresponding author.
E-mail addresses: aman.jolly1994@gmail.com (A. Jolly), shailenderkumar@dce.ac.in (S. Kumar).

https://doi.org/10.1016/j.compeleceng.2026.111203
Received 22 April 2025; Received in revised form 21 February 2026; Accepted 22 April 2026

SPRINGER NATURE Link

Account

A rigorous interrogation of multimodal fake news detection paradigms: adversarial robustness and modality dependence

Published: 03 June 2026

Volume 29, article number 307 (2026) Cite this article



Cluster Computing

Aims and scope →

Submit manuscript →

Aman Jolly & Shailender Kumar

96 Accesses Explore all metrics →

Abstract

The increasing sophistication of multimodal disinformation driven by generative AI and cross-modal manipulation poses significant challenges to the reliability of existing Multimodal fake news detection (MFND) systems. This work presents a rigorous interrogation and comprehensive evaluation of canonical MFND architectures, rather than proposing a new system. We assess the feasibility and robustness of five diverse architectures, LongBERT-VGG, CNN-VGG, LSTM-VGG, BERT-CLIP, and the end-to-end transformer ViLBERT across three benchmark datasets: D1, RECOVERY, and PolitiFact, which collectively represent varying degrees of class balance and content characteristics. Our analysis transcends conventional evaluations by subjecting these models to an extensive battery of adversarial perturbations: ten distinct image-based attacks and seven sophisticated text-based attacks. Crucially, we introduce novel adversarial textual manipulations generated via instruction-tuned LLMs, specifically comparing the potency of attacks derived from Meta's Llama 3 8B against those from DeepSeek-LLM 7B Chat. We further introduce an analysis of black-box transfer attacks to simulate real-world threats.

Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-026-11410-y

RESEARCH ARTICLE-COMPUTER ENGINEERING AND COMPUTER SCIENCE

A Comprehensive Evaluation of Multimodal Fake News Detection: Assessing Robustness and Efficacy

Aman Jolly[1] · **Shailender Kumar**[1]

Received: 20 July 2025 / Accepted: 7 May 2026
© King Fahd University of Petroleum & Minerals 2026

Abstract
The escalating threat of multimodal misinformation, propagated through manipulated text and imagery, necessitates robust and resilient detection frameworks. As generative AI advances, adversarial attacks exploit both modalities to bypass existing defenses, posing a critical challenge. While multimodal fake news detection models offer promising solutions by integrating textual and visual features, their inherent vulnerabilities under adversarial conditions and their behavior in incomplete modality scenarios remain insufficiently explored. This study provides a systematic evaluation of three widely-used multimodal architectures—BERT-VGG, CNN-VGG, and LSTM-VGG—on two benchmark datasets, D1 and RECOVERY. We significantly extend the scope of robustness assessment by subjecting these models to a diverse range of adversarial attacks, including ten image-based and five text-based perturbations, alongside novel large language model (LLM)-driven manipulations using LLAMA3. Furthermore, our research rigorously investigates the impact of modality removal (unimodal settings) on detection performance, offering crucial insights into the intrinsic reliance on specific modalities. Our findings reveal significant vulnerabilities within current multimodal fake news detection models, particularly their susceptibility to advanced text-based adversarial attacks and evidence of "Unimodal Bias." Crucially, the image modality is demonstrated to play an indispensable role in maintaining detection efficacy and resilience. This comprehensive analysis highlights the urgent need for adaptive strategies and advanced architectures that can withstand sophisticated adversarial tactics and manage modality inconsistencies, paving the way for the development of more resilient and effective next-generation misinformation detection systems.

Keywords Fake news detection · Multimodal approach · Adversarial attacks · LLM attacks · Modality removal · Unimodal Bias

1 Introduction

The rapid proliferation of multimodal content on digital platforms significantly complicates the detection of false news. Misinformation now frequently integrates both textual and visual elements, exploiting human perceptual biases and the limitations of automated detection systems.

A primary challenge arises from the inherent differences between text and image representations. Textual data are based on linguistic structures, while images require visual feature extraction, often using convolutional neural networks (CNNs). Simple feature concatenation frequently results in the loss of crucial contextual relationships between modalities. While machine learning has demonstrated robust predictive capabilities in diverse fields ranging from price prediction [1] to medical diagnosis, the complexity of multimodal misinformation requires specialized deep learning fusion techniques. Multimodal fake news detection models combine text and image analysis, typically through feature fusion, to improve accuracy. However, ensuring their robustness and reliability remains critical, particularly in adversarial settings or when a modality is missing or corrupted.

Additionally, adversarial attacks that manipulate text or images can exploit model vulnerabilities. Traditional models often fail to capture hidden contextual cues or detect subtle semantic inconsistencies, making them susceptible to

✉ Aman Jolly
aman.jolly1994@gmail.com

Shailender Kumar
shailenderkumar@dce.ac.in

[1] Department of Computer Science and Engineering, Delhi Technological University, Delhi 110042, India

Published online: 27 May 2026

Gmail

Aman Jolly <aman.jolly1994@gmail.com>

Scientific Reports: Decision on "EAKD-CFND: Continual Fake News Detection through AdaptiveKnowledge Distillation with External Verification"

1 message

Scientific Reports <do-not-reply@springernature.com> Sat, Aug 29, 2026 at 7:08 PM
To: aman.jolly1994@gmail.com

Dear Dr Jolly,

Re: "EAKD-CFND: Continual Fake News Detection through AdaptiveKnowledge Distillation with External Verification"

We are delighted to let you know that the above submission, which you co-authored, has been accepted for publication in Scientific Reports.

Please contact the corresponding author if you would like further details on this decision, including any reviewer feedback.

Thank you for choosing Scientific Reports and we look forward to publishing your article.

ORCiD records

If the corresponding author links their Springer Nature and ORCID accounts, the article will be added to their ORCID record when published. Learn more at: https://www.springernature.com/gp/researchers/orcid

Kind regards,

Peer Review Advisors
Scientific Reports

Navigating the Gray Area: A Three-Label Framework for Uncovering Uncertainty in Fake News

Publisher: IEEE Cite This PDF

Aman Jolly ; Shailender Kumar All Authors

42
Full
Text Views

Abstract

Document Sections

I. Introduction

II. Related Work

III. Methodology

IV. Exploratory Data Analysis

V. Discussions and Results

Abstract:
Detecting the veracity of textual content, and distin-guishing between fake and real news, has been greatly improved with the advent of BERT-based transformers. However, an in-herent challenge arises when encountering news articles that fall within the borderline of the latent vector space between fake and real. In this paper, we propose a three-label framework to address this uncertainty issue and leverage an external data enrichment module, Google Fact Check, to handle such cases. Our approach utilizes a constraint dataset for training both the CT-BERT and BERT base models. Experimental results demonstrate that while the BERT base model achieves an accuracy of 80%, the accuracy of CT-BERT, our proposed framework, reaches 93%. This significant improvement in a ccuracy showcases the effectiveness of our approach in tackling uncertainty in fake news detection.

Published in: 2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)


ieeexplore.ieee.org/document/11265068
90%
Conferences > 2025 International Conference...
FakeNews Buster: A Continual Learning Approach with BERT and FastFormer for Adaptive and Explainable Fake News Detection
Publisher: IEEE
Cite This
PDF
Aman Jolly ; Shailender Kumar All Authors
65
Full
Text Views
Abstract
Document Sections
I. Introduction
II. Literature Survey
III. Methodology
IV. Results
V. Discussion
Show Full Outline
Authors
Abstract:
Social media fake news evolves rapidly, rendering static detection models ineffective. FakeNews Buster is a dynamic continual learning framework addressing concept drift in misinformation. It integrates BERT and the lightweight FastFormer, using fixed-length and recurrent processing for varied text formats like tweets. A custom loss function manages uncertainty in noisy social data, while knowledge distillation retains critical information. An explainability module combining the Google Fact Check API with vector-based semantic search enhances interpretability, providing evidence-based justifications. We explore multiple training strategies and assess performance using standard classification metrics. The model's vocabulary is expanded with domain-specific terms and emojis to reflect informal online content. Results show FakeNews Buster offers improved adaptability, transparency, and trustworthiness in fake news detection through its novel integration of continual learning with external fact-checking capabilities.
Published in: 2025 International Conference on Intelligent and Secure Engineering Solutions (CISES)

8th International Conference on Communication and Intelligent Systems (ICCIS 2026)

Organized by

BITS Pilani, K K Birla Goa Campus, India

Certificate of Presentation

This certificate is awarded to

Aman Jolly

who has presented the paper titled **Retrieval Without Generation: Evidence-Grounded Fake News Classification** authored by **Aman Jolly, Shailender Kumar** in the **8th International Conference on Communication and Intelligent Systems (ICCIS 2026)** held at **BITS Pilani, K K Birla Goa Campus, India** during **September 26-27, 2026.**

SCRS/ICCIS 2026/PC/1253

Snehanshu Saha
BITS Pilani, Goa Campus, India
(General Chair)

Hemant Rathore
BITS Pilani, Goa Campus, India
(General Chair)

Springer

https://scrs.in/conference/iccis2026

PLAGIARISM REPORT

The Turnitin originality report for this thesis is reproduced below, followed by the plagiarism verification certificate.

turnitin Page 1 of 153 - Cover Page Submission ID trn:oid:::27535:146388137

Shailender Kumar

Aman Jolly Thesis full 17-07-2026

paper

Document Details

Submission ID
trn:oid:::27535:146388137

Submission Date
Jul 24, 2026, 5:32 PM GMT+5:30

Download Date
Jul 24, 2026, 5:37 PM GMT+5:30

File Name
Aman Jolly Thesis full 17-07-2026.pdf

File Size
6.4 MB

148 Pages

47,568 Words

281,293 Characters

turnitin Page 1 of 153 - Cover Page Submission ID trn:oid:::27535:146388137

2% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Small Matches (less than 10 words)

Match Groups

35 Not Cited or Quoted 2%
Matches with neither in-text citation nor quotation marks

21 Missing Quotations 1%
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

0% Internet sources
2% Publications
0% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **35** Not Cited or Quoted 2%
  Matches with neither in-text citation nor quotation marks
- **21** Missing Quotations 1%
  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

0% Internet sources
2% Publications
0% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Publication	
Aman Jolly, Shailender Kumar. "FakeNews Buster: A Continual Learning Approach...		<1%
2	Publication	
Aman Jolly, Shailender Kumar. "A rigorous interrogation of multimodal fake news...		<1%
3	Publication	
Aman Jolly, Shailender Kumar. "A Comprehensive Evaluation of Multimodal Fake ...		<1%
4	Student papers	
Maulana Azad National Institute of Technology Bhopal on 2026-06-25		<1%
5	Internet	
download.bibis.ir		<1%
6	Student papers	
Indian Institute of Technology Patna on 2022-08-31		<1%
7	Publication	
"ECAI 2020", IOS Press, 2020		<1%
8	Publication	
"Advanced Intelligent Computing Technology and Applications", Springer Science...		<1%
9	Publication	
Arvind Dagur, Karan Singh, Pawan Singh Mehra, Dhirendra Kumar Shukla. "Intelli...		<1%
10	Internet	
icon2021.nits.ac.in		<1%

11 Student papers
Brunel University on 2024-11-29 <1%

12 Student papers
University of Newcastle on 2025-09-21 <1%

13 Internet
pmc.ncbi.nlm.nih.gov <1%

14 Publication
Li, Chih-Yuan. "AI Approaches to Understand Human Deceptions, Perceptions, an... <1%

15 Publication
Mohapatra, Payal. "Efficient, Robust and Interpretable Learning for Multimodal T... <1%

16 Publication
Wang, Jianwei. "Graph Unbound: Toward Efficient Explicit and Implicit Graph Lear... <1%

17 Internet
lrec-conf.org <1%

18 Internet
umontreal.scholaris.ca <1%

19 Publication
Aman Jolly, Shailender Kumar. "Beyond the surface: Adversarial stress testing of ... <1%

20 Publication
Mohammad Amin Kuhail, Mostafa Mohamad, Rawad Hammad, Mohammed Bahj... <1%

21 Publication
"Combating Online Hostile Posts in Regional Languages during Emergency Situat... <1%

22 Student papers
University College Dublin (UCD) on 2026-06-18 <1%

23 Publication
"Advanced Intelligent Computing Technology and Applications", Springer Science... <1%

24 Publication
Akhtar, Mohammad Majid. "Unmasking the Hidden Threats of Social Media: Dete... <1%

25 Student papers

George Mason University on 2026-06-02 <1%

26 Publication

Go, Jin Ho. "Fake News Detection With Graph Neural Networks.", University of Ne... <1%

27 Publication

Huang, Hongtao. "Efficient Deep Learning: Model Design and Algorithmic Innova... <1%

28 Publication

Mohtaj, Salar. "Transfer Learning for Harmful Content Detection.", Technische Un... <1%

29 Publication

Pietro Dell'Oglio, Alessandro Bondielli, Francesco Marcelloni, Lucia C. Passaro. "A... <1%

30 Publication

Tingqi Hu, Ying Lei, Yiduo Wang, Ruiyu Ma, Xinyi Yin, Minglin Liu, Ruoyu Zhao. "Be... <1%

31 Internet

www.mdpi.com <1%



DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the thesis: **DEVELOPMENT OF EFFICIENT AND INTELLIGENT FAKE NEWS DETECTION FRAMEWORK**

Total Pages: **142** Name of the Scholar: **AMAN JOLLY**

Supervisor(s):

1. **PROF. SHAILENDER KUMAR**

Department: **COMPUTER SCIENCE AND ENGINEERING**

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software Used: **TURNITIN** Similarity Index: **2%** Word Count: **47,568**

Date: _____

Candidate's Signature **Signature of Supervisor**

CURRICULUM VITAE



Aman Jolly received his B.Tech. degree in Computer Science and Engineering from Guru Gobind Singh Indraprastha University, Delhi, India, in 2016, and his M.Tech. in Information Security from the same university in 2018. He is an AI/ML Engineer and NLP Research Scientist with over four years of industry experience and more than three years in academia. He currently works as a Senior Scientific Specialist (AI/ML Engineer, Agentic Systems) at Excelra Knowledge Solutions, where he designs agentic large language model (LLM) systems, knowledge-graph engineering pipelines, and production machine-learning workflows on Google Cloud Platform for pharmaceutical and bioinformatics applications. His technical expertise spans Neo4j and semantic-web technologies, retrieval-augmented generation, parameter-efficient fine-tuning (LoRA/QLoRA), ontology design (OWL/RDF, SPARQL), and GxP-aligned model deployment. His research interests include the adversarial robustness of multimodal LLMs and fake news detection models, Graph-RAG architectures for evidence-based misinformation classification, continual learning and knowledge distillation for evolving NLP tasks, agentic AI systems for scientific automation, and ontology engineering for biomedical data FAIRification. He has authored four SCIE-indexed journal articles, three published in *Computers and Electrical Engineering*, *Cluster Computing*, and the *Arabian Journal for Science and Engineering* and one accepted in *Scientific Reports*, alongside IEEE conference papers. He actively contributes to the research community in artificial intelligence, adversarial NLP, and trustworthy machine learning.
Email: aman.jolly1994@gmail.com *LinkedIn:* linkedin.com/in/amanjolly