Shivansh.pdf

Delhi Technological University

Document Details

Submission ID
trn:oid:::27535:102281025

Submission Date
Jun 24, 2025, 2:19 PM GMT+5:30

Download Date
Jun 24, 2025, 2:24 PM GMT+5:30

File Name
Shivansh.pdf

File Size
1.7 MB

47 Pages

13,754 Words

85,824 Characters

12% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 8 words)

Match Groups

118 Not Cited or Quoted 12%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

8% Internet sources
5% Publications
9% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

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

- 118 Not Cited or Quoted 12%
  Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%
  Matches that are still very similar to source material
- 0 Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

- 8% Internet sources
- 5% Publications
- 9% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	dspace.dtu.ac.in:8080 2%
2	Internet	www.dspace.dtu.ac.in:8080 <1%
3	Internet	www.mdpi.com <1%
4	Internet	aclanthology.org <1%
5	Internet	arxiv.org <1%
6	Submitted works	University of Exeter on 2024-08-21 <1%
7	Internet	www.biorxiv.org <1%
8	Submitted works	Swinburne University of Technology on 2024-06-07 <1%
9	Publication	Tomislav Duricic, Dominik Kowald, Emanuel Lacic, Elisabeth Lex. "Beyond-accurac... <1%
10	Submitted works	University of Hertfordshire on 2024-12-02 <1%

11 Internet
dspace.daffodilvarsity.edu.bd:8080 <1%

12 Submitted works
Liverpool John Moores University on 2024-12-15 <1%

13 Internet
dspace.cvut.cz <1%

14 Internet
thesai.org <1%

15 Internet
cdn.iiit.ac.in <1%

16 Submitted works
Indian Institute of Information Technology Sri City on 2025-04-09 <1%

17 Internet
acl2020.org <1%

18 Internet
ir.lib.uwo.ca <1%

19 Publication
Yuxin Han, Runtao Yang, Mingyu Zhu, Lina Zhang. "A sarcasm detection method ... <1%

20 Internet
ojs.aaai.org <1%

21 Internet
5dok.net <1%

22 Publication
Alaa Sheta, Shyam Subramanian, Salim R. Surani, Malik Braik. "Diagnosis of Obstr... <1%

23 Submitted works
University of Oklahoma on 2024-11-29 <1%

24 Internet
bastina.anubih.ba <1%

25 Internet

www.jp-ca.org <1%

26 Submitted works

Liverpool John Moores University on 2025-06-22 <1%

27 Submitted works

United International College on 2024-12-17 <1%

28 Submitted works

University of Birmingham on 2024-04-08 <1%

29 Submitted works

University of East London on 2025-05-09 <1%

30 Publication

Vandita Grover, Hema Banati. "An attention approach to emoji focused sarcasm ... <1%

31 Internet

www.ijcai.org <1%

32 Submitted works

Brunel University on 2025-04-11 <1%

33 Submitted works

University Politehnica of Bucharest on 2024-06-22 <1%

34 Submitted works

University of Westminster on 2023-07-17 <1%

35 Internet

download.bibis.ir <1%

36 Internet

mediatum.ub.tum.de <1%

37 Submitted works

Georgia Institute of Technology Main Campus on 2022-04-29 <1%

38 Submitted works

Leiden University on 2024-01-05 <1%

39 Submitted works
Liverpool John Moores University on 2024-06-17 <1%

40 Submitted works
Universidad Carlos III de Madrid - EUR on 2025-06-19 <1%

41 Submitted works
University of Wollongong on 2024-03-28 <1%

42 Internet
assets-eu.researchsquare.com <1%

43 Internet
peerj.com <1%

44 Submitted works
Associatie K.U.Leuven on 2019-08-19 <1%

45 Publication
Christina Haag, Nina Steinemann, Deborah Chiavi, Christian P. Kamm et al. "Blen... <1%

46 Submitted works
Curtin University of Technology on 2019-11-01 <1%

47 Submitted works
Liverpool John Moores University on 2025-02-06 <1%

48 Submitted works
Munster Technological University (MTU) on 2025-06-12 <1%

49 Publication
Qiaofeng Wu, Wenlong Fang, Weiyu Zhong, Fenghuan Li, Yun Xue, Bo Chen. "Dual... <1%

50 Publication
Ramineni Chittibabu Naidu, Ghaneshvar. "Integrating Deep Learning and Persist... <1%

51 Submitted works
University of Greenwich on 2023-08-16 <1%

52 Submitted works
University of Nottingham on 2023-12-06 <1%

53 Internet
dokumen.pub <1%

54 Internet
ebin.pub <1%

55 Internet
tede2.pucrs.br <1%

56 Publication
"Computational Linguistics and Intelligent Text Processing", Springer Science an... <1%

57 Publication
"Natural Language Processing and Information Systems", Springer Science and B... <1%

58 Submitted works
CSU, San Jose State University on 2023-05-18 <1%

59 Submitted works
Cranfield University on 2024-08-19 <1%

60 Publication
Darmofal, Madison. "Enhancing the Clinical Utility of Genomic Profiling for Cance... <1%

61 Publication
Hanoi Pedagogical University 2 <1%

62 Submitted works
Intercollege on 2023-09-20 <1%

63 Publication
Lecture Notes in Computer Science, 2015. <1%

64 Submitted works
Liverpool John Moores University on 2023-03-15 <1%

65 Submitted works
Liverpool John Moores University on 2023-05-27 <1%

66 Submitted works
Monash University on 2025-05-29 <1%

67 Submitted works
National College of Ireland on 2024-12-12 <1%

68 Publication
Qiuyu Li, Zuhe Li, Weihua Liu, Xiaojiang He, Yushan Pan. "Sarcasm-GPT: advancin... <1%

69 Publication
Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical ... <1%

70 Submitted works
The Hong Kong University of Science and Technology (Guangzhou) on 2024-12-09 <1%

71 Submitted works
Universidad Carlos III de Madrid - EUR on 2025-06-19 <1%

72 Submitted works
University College London on 2025-02-02 <1%

73 Submitted works
University of Newcastle on 2022-04-13 <1%

74 Submitted works
VIT University on 2024-10-29 <1%

75 Internet
isip.piconepress.com <1%

76 Internet
openreview.net <1%

77 Internet
papers.nips.cc <1%

78 Internet
pmc.ncbi.nlm.nih.gov <1%

79 Internet
repositories.nust.edu.pk <1%

80 Internet
studentsrepo.um.edu.my <1%

81 Internet

tdr.lib.ntu.edu.tw <1%

82 Internet

www.inesc-id.pt <1%

Emotion-Aware Multimodal Sarcasm Detection Using Deep Learning Techniques

Thesis Submitted
in Partial Fulfillment of the Requirements
for the Degree of

MASTER OF TECHNOLOGY
in
Artificial Intelligence

by

Shivansh Khera
(23/AFI/03)

Under the supervision of

Dr. Sanjay Kumar
(Dept of Computer Science & Engineering)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

MAY, 2025

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **Shivansh Khera**, Roll No: **23/AFI/03** studens of M.Tech (**Artificial Intelligence**), hereby declare that the project Dissertation titled "**Emotion-Aware Multimodal Sarcasm Detection Using Deep Learning Techniques**" which is submitted by me to the **Department of Computer Science and Engineering**, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi Shivansh Khera

Date: 30/05/2025

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled "**Emotion-Aware Multi- modal Sarcasm Detection Using Deep Learning Techniques,**" which is submitted by **Shivansh Khera**, Roll No: **23/AFI/03**, **Department of Computer Science and Engineering**, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Dr. Sanjay Kumar

Date: 30/05/2025

SUPERVISOR

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I wish to express my sincerest gratitude to Dr **Sanjay Kumar** for his continuous guidance and mentorship that he provided me during the project. He showed me the path to achieve my targets by explaining all the tasks to be done and explained to me the importance of this project as well as its industrial relevance. He was always ready to help me and clear our doubts regarding any hurdles in this project. Without his constant support and motivation, this project would not have been successful.

Place: Delhi Shivansh Khera

Date: 30.05.2025

Abstract

Sarcasm detection in digital communication has become increasingly challenging as users employ sophisticated combinations of textual and visual elements to convey ironic meaning. Traditional text-based approaches and existing multimodal methods struggle to capture the emotional incongruity that is fundamental to sarcastic expressions, representing a significant gap in current multimodal sarcasm detection research.

This thesis proposes an emotion-aware multimodal framework for sarcasm detection that systematically integrates emotional features from both textual and visual modalities. The approach enhances existing deep learning architectures, specifically Bidirectional Long Short-Term Memory (BiLSTM) networks and Graph Convolutional Networks (GCNs), with emotion recognition capabilities utilizing DistilRoBERTa for textual emotion analysis and computer vision techniques for visual emotion recognition. The framework was primarily developed and optimized using the MMSD2.0 dataset, followed by comprehensive cross-dataset evaluation on MMSD Original and MEMOTION datasets to assess generalization capabilities.

Experimental results demonstrate that systematic integration of emotional features from both modalities significantly improves sarcasm detection performance. The BiLSTM-based emotion-aware architecture achieves the best overall performance on MMSD2.0 with 83.12% accuracy, 81.10% precision, 85.07% recall, and 83.04% F1-score, while the GCN-based approach achieves competitive results with 81.29% accuracy and 81.97% F1-score. Cross-dataset evaluation reveals robust generalization capabilities, with the BiLSTM model maintaining 79.86% accuracy on MMSD Original and 80.45% accuracy on MEMOTION, demonstrating effective transferability of emotion-aware features across different data distributions and even different domains. These findings establish the first empirical evidence for the effectiveness of dual-modality emotion integration in multimodal sarcasm detection, providing a robust foundation for future research in emotion-aware approaches to understanding digital communication nuances.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Sarcasm represents one of the most nuanced and contextually dependent forms of human communication, characterized by the deliberate expression of meaning through its opposite. This linguistic phenomenon, which relies heavily on tone, context, and shared understanding between communicators, has become increasingly prevalent in digital communication environments. With our social evolution towards mostly online interactions via social media websites, messaging apps, and online forums, the sophistication of identifying and understanding sarcastic phrases has increased manifold.

The evolution of human communication into the digital age has literally turned the mannerism of sarcasm and how it is communicated and perceived upside down. Pre-digital age face-to-face communication offered rich contextual signals through facial expressions, tone of voice, and body language, making sarcastic intent fairly obvious to the human eye. Yet, in online spaces, such classical cues do not exist and users are compelled to introduce new ways of encoding sarcastic meaning. Modern digital communication has developed to include multimodal components, wherein users deliberately blend textual message with visual cues like images, memes, GIFs, and emojis to produce complex layers of meaning that efficiently convey sarcastic intention.

This development has presented computational systems with unprecedented difficulties in deciphering and processing human communication. Although humans tend to intuitively identify sarcasm through contextual knowledge and emotional intelligence, artificial systems have a hard time dealing with the veiled contradictions and implied senses inherent in sarcastic expressions. The challenge is especially acute in multimodal scenarios, where sarcastic intent can arise from the highly nuanced interaction between text and images, so that systems need to comprehend not only individual elements but also their collective semantic and affective consequences.

Previous strategies for the detection of sarcasm have centered on involuntary attention towards text, using rule-based systems or machine learning to detect patterns in language, polarities of sentiment, and contextual contradiction. Although these approaches have had some success in controlled settings, they tend to show weak effectiveness when faced with the dynamic, creative, and culturally contextualized means by which sarcasm is deployed in actual digital communication. The advent of multimodal content has also made evident the weaknesses of text-based approaches, since expressions of sarcasm often rely on visual context that offers important interpretive information.

Emotional understanding integration is a key frontier to enabling sarcasm detection. Human awareness of sarcasm tends to be based on detecting emotional incon-

gruities—cases where expressed emotions are opposite to contextual expectations or where emotional strength seems disproportionate to the topic at hand. For example, being overly enthusiastic regarding trivial disappointments or expressing surprise regarding clearly predictable events tends to indicate sarcastic purpose. Computational systems that can systematically consider and combine emotional characteristics both in textual and visual modes are well-suited to attain more human-like comprehension of sarcasm.

1.1.1 Research Novelty and Breakthrough Contribution

This thesis addresses these fundamental challenges by proposing the **first comprehensive, emotion-aware multimodal framework** that systematically integrates emotional features extracted from both textual and visual content for sarcasm detection. Unlike all previous approaches that either focus solely on text-based emotion analysis or treat emotional information as auxiliary features, this research establishes emotion understanding as a core architectural component, creating a unified dual-modal emotion integration framework.

1.1.2 Pioneering Approach

While existing research has explored emotion-aware text-only sarcasm detection (Felbo et al., 2017; Chauhan et al., 2020) and basic multimodal fusion techniques (Castro et al., 2019; Cai et al., 2019), **no prior work has systematically integrated emotions from both text and images in a unified multimodal architecture**. This research represents a paradigm shift by:

1. **First Dual-Modal Emotion Extraction**: Systematically extracting emotions from both text (using DistilRoBERTa) and images (using ResNet50 with custom emotion classification heads)

2. **Comprehensive Emotion Ablation**: Conducting the first systematic study comparing four emotion integration strategies: baseline (no emotions), text-only emotions, image-only emotions, and full dual-modal emotion integration

The technical foundations supporting this method utilize cutting-edge deep learning technologies, most notably using BiLSTM networks for advanced sequential processing of text data and GCNs for representing intricate relational dependencies across various modalities and their respective emotional properties. Both architectural designs are supplemented with dedicated emotion recognition modules that allow systematic emotion context integration into the detection process, forming an end-to-end framework that can comprehend the emotional undertones that often govern sarcastic utterances.

The empirical basis of this work is founded upon the MMSD2.0 dataset, which is the state-of-the-art benchmark for multimodal sarcasm detection research. This data set offers rich annotations of both text and visual modalities, which can be used to systematically test emotion-aware detection methods. Through thorough experimentation and testing, this paper shows that systematic fusion of emotional features produces substantial gains in detection performance and that optimized BiLSTM model gets 83.12% accuracy and GCN-based model gets 81.29% accuracy and sets new state-of-the-art performance levels for multimodal sarcasm detection that is emotion-aware.

1.2 Motivation

The motivation for this study arises from several converging factors that highlight the need for advanced emotion-aware multimodal sarcasm detection:

1. Multimodal Sarcasm Prevalence: Exponential growth of social media has made prevalence of sarcasm widespread online, and there is an urgent need for those sentiment analysis tools that can deal with this complexity. Today's users tend to use sarcastic linguistic forms that involve textual as well as visual cues, so computational methods must be able to process and comprehend such multimodal interactions.

2. Limitations of Existing Sentiment Analysis Systems: Most existing sentiment analysis systems usually fail to detect hyperbole because it is based primarily on contextual subtleties and conflicting emotional expressions. Sarcasm, which is usually defined by incongruent emotional expressions, presents difficulties for conventional sentiment analysis systems, and thus limits their ability to properly interpret user intent.

Emotion	Sarcastic	Non-sarcastic	Total
anger	768	514	1282
disgust	138	96	234
fear	376	347	723
joy	2634	3295	5929
neutral	3000	3710	6710
sadness	1009	1023	2032
surprise	1651	1255	2906

Table 1.1: Text Emotions Distribution (Sarcastic vs Non-sarcastic)

3. Statistical Evidence for Emotional Integration: Comprehensive statistical analysis of the MMSD2.0 dataset reveals compelling evidence for the critical role of emotions in sarcasm identification. Table 1.1 summarizes the distribution of text emotions, while Table 1.2 presents the distribution of image emotions. For instance, sarcasm occurs in 2,634 instances when the text conveys "joy" (compared to 3,295 for non-sarcastic expressions), and "surprise" is the dominant emotion in both sarcastic and non-sarcastic images. These statistical disparities highlight the subtle but impactful role of emotional signals in distinguishing sarcastic from non-sarcastic content.

Emotion	Sarcastic	Non-sarcastic	Total
Anger	144	107	251
Disgust	623	504	1127
Fear	412	537	949
Joy	652	898	1550
Neutral	1642	1073	2715
Sadness	997	1306	2303
Surprise	5106	5815	10921

Table 1.2: Image Emotions Distribution (Sarcastic vs Non-sarcastic)

Similarly, image emotion analysis shows distinct patterns, with "surprise" being the dominant emotion in both sarcastic (5,106 instances) and non-sarcastic (5,815 instances) cases, while "neutral" emotions show higher prevalence in sarcastic content (1,642 in-

stances) compared to non-sarcastic content (1,073 instances). Table 1.2 provides the detailed distribution.

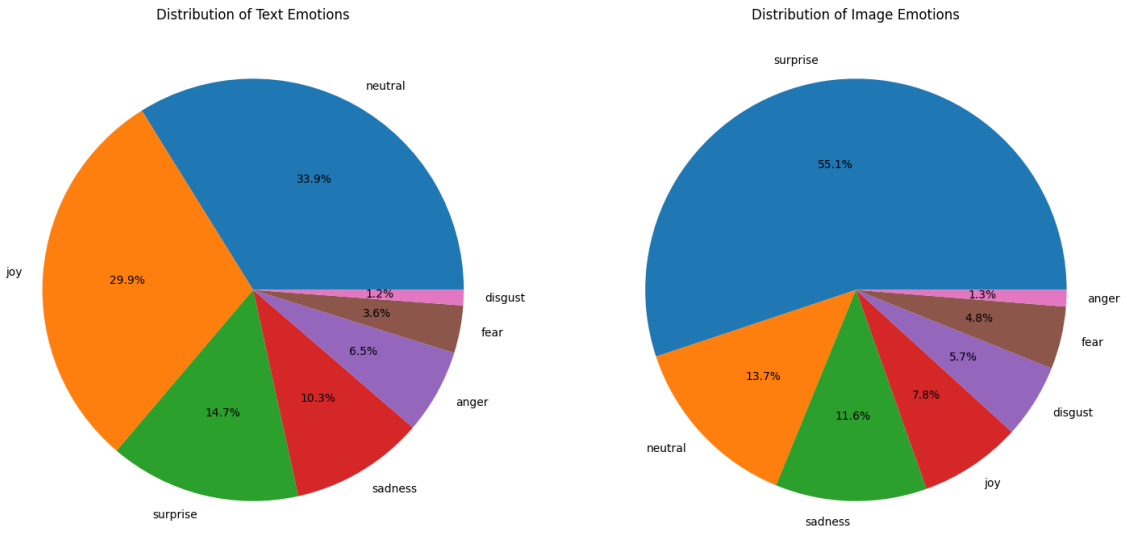

Figure 1.1: Distribution of Emotion in Text and Images

4. Gap in Emotion-Aware Multimodal Systems: Despite advances in natural language processing and computer vision, existing multimodal sarcasm detection approaches often overlook the emotional layer that plays a crucial role in interpreting sarcastic content. The absence of systematic emotional analysis limits the ability of current systems to understand the emotional undertones that frequently characterize sarcastic expressions, resulting in reduced detection accuracy and contextual understanding.

1.3 Problem Statement

Sarcasm detection in multimodal digital communication presents significant computational challenges due to the contradictory nature of sarcastic expressions, where literal meaning often opposes intended meaning. Traditional text-based approaches struggle to capture the full complexity of sarcasm as it is increasingly expressed through combinations of textual and visual elements in social media platforms. Current multimodal sarcasm detection systems lack systematic integration of emotional context, despite emotions playing a crucial role in conveying sarcastic intent through incongruent emotional expressions.

This research addresses the need for improved multimodal sarcasm detection by developing emotion-aware frameworks that systematically integrate emotional features from both text and images. The challenge lies in effectively modeling the relationships between textual content, visual elements, and their associated emotional characteristics to achieve more accurate and contextually aware sarcasm detection in real-world digital communication scenarios.

1.4 Contributions

This thesis makes the following key contributions:

- Proposes a novel emotion-aware framework for multimodal sarcasm detection, integrating emotional features from both text and images.
- Develops and evaluates advanced BiLSTM and GCN architectures, each enhanced with emotion recognition capabilities.
- Demonstrates, through extensive experiments on the MMSD2.0 dataset, that emotion integration significantly improves detection accuracy, achieving state-of-the-art results.
- Provides a comprehensive analysis of the challenges, limitations, and future directions for emotion-aware multimodal sarcasm detection.

Research Impact: This work fundamentally advances the state-of-the-art by proving that emotions play a crucial role in multimodal sarcasm detection and provides the first systematic methodology for leveraging this insight. The established framework opens new research directions in affective computing and provides practical solutions for improving social media analysis systems.

Thesis Structure: The remainder of this thesis is organized as follows: Chapter 2 reviews related work, Chapter 3 details the methodology, Chapter 4 presents results and discussion, and Chapter 5 concludes with future directions and impact.

Chapter 2

LITERATURE REVIEW

Sarcasm detection has gained significant traction as a research area within natural language processing (NLP), particularly given its implications for sentiment analysis, dialogue systems, and social media monitoring. The complexity of sarcastic communication, which relies on the deliberate contradiction between literal expression and intended meaning, has motivated researchers to explore increasingly sophisticated computational approaches. This literature review explores prior work on sarcasm detection, examining the evolution from traditional text-based methods to contemporary multimodal systems, and the emerging integration of emotional understanding in computational models.

2.1 Traditional Text-Based Sarcasm Detection

2.1.1 Early Linguistic and Pattern-Based Approaches

When researchers first approached sarcasm detection computationally, they understandably used what they had worked with before: linguistic patterns and text features. The problem was obvious - how do you train a machine to recognize something as subtle as sarcasm when humans often fail to catch it too?

Davidov et al. (2010) were among the first to really try to crack this brain teaser using supervised learning on Twitter data. What was most interesting about their work was that it concentrated on brief texts - the 140-character limit of tweets provided a special challenge. With limited context to go on, they needed to spot subtle patterns that might indicate sarcastic intent. Their pattern-driven approach, although looking deceptively easy by today's standards, in fact uncovered something significant: that with minimal textual content, close observation of linguistic signals could produce reasonable results.

Drawing on this as a base, Riloff et al. (2013) approached the problem differently. Rather than searching for patterns, they aimed at something more intrinsic to sarcasm - incongruity. Their bootstrapped classifier was intent on identifying those instances where one mentions something pleasant regarding a clearly unpleasant circumstance. Consider posting "Love sitting in traffic for two hours!" - the incongruity between positive language and bad circumstance screams sarcasm. This task was vital in that it pinpointed situational incongruity as a fundamental computational property, something that would shape the study of sarcasm detection for decades to come.

At about the same period, González-Ibáñez et al. (2011) were exploring the stylistic features of sarcastic tweets. They observed that sarcastic users tend to use language differently - more punctuation, strange capitalization, certain choice of words. Although these surface features by themselves were not sufficient for strong detection, they gave

useful clues as to how sarcasm is expressed in computer-mediated communication.

2.1.2 Machine Learning and Feature Engineering

As the discipline matured, scholars started shifting away from strict rule-based systems and toward more adaptive machine learning techniques. The change wasn't solely technological - it was also a function of an increasing realization that sarcasm is too intricate and rich to be reducible to mere patterns.

Joshi et al. (2015) took a deep view of where the research was at this pivotal turning point. Their own survey reiterated a persistent theme: context plays an incredibly important role in sarcasm detection. You can't simply examine words in isolation; you must factor in the larger conversational context, the speaker's history, and even community standards. This finding would turn out to be prophetic as later work more and more focused on understanding context.

Bamman and Smith (2015) were serious about this contextual method, claiming that proper sarcasm detection needs to know not only what was uttered, but by whom, to whom, and under which social circumstances. They demonstrated substantial performance gains in contextual sarcasm detection when models considered user history and conversational context. This was intuitive - if you know the person is usually sarcastic, or if you're following the conversation, you're much more likely to understand what they mean by ambiguous words.

The introduction of word embeddings marked another significant shift in how researchers approached the problem. Ghosh and Veale (2016) explored how distributed word representations could capture the semantic inversions that characterize sarcasm. What was particularly clever about their approach was recognizing that sarcasm often involves subtle semantic shifts - words that seem positive but are used in negative contexts, or familiar phrases twisted into unfamiliar meanings.

2.1.3 Deep Learning and Neural Approaches

The deep learning revolution hit sarcasm detection like everywhere else in NLP, but with some interesting twists. Unlike straightforward classification tasks, sarcasm detection needed to capture very subtle sequential patterns and semantic relationships.

Poria et al. (2016) demonstrated that BiLSTM networks were particularly well-suited to this challenge. The bidirectional processing was key - sarcastic expressions often have crucial context both before and after the main statement. Consider "Yeah, because that worked so well last time" - you need to process the entire sequence to understand the sarcastic intent. Their work showed that neural approaches could significantly outperform traditional machine learning methods by better modeling these complex sequential dependencies.

The introduction of attention mechanisms by Tay et al. (2018) was a natural evolution. If certain parts of a text are more important for detecting sarcasm, why not teach the model to focus on those parts? Their attention-based approaches could identify key phrases and semantic patterns that humans typically use to signal sarcastic intent. This was particularly valuable for longer texts where the sarcastic element might be buried among neutral content.

When transformer models like BERT arrived, researchers naturally explored their potential for sarcasm detection. Potamias et al. (2020) conducted extensive experiments

showing that pre-trained language models could capture many of the complex linguistic patterns that characterize sarcasm. However, they also highlighted an important limitation: these models, while powerful, still struggled with the contextual and cultural nuances that make sarcasm detection so challenging.

2.2 Multimodal Sarcasm Detection

2.2.1 Evolution to Multimodal Frameworks

The explosion of visual social media platforms fundamentally changed how people express sarcasm. Suddenly, a sarcastic tweet wasn't just text - it might include a meme, a photo with contrasting emotion, or an image that completely contradicts the written message. Researchers realized they needed to adapt.

This evolution wasn't just about keeping up with technology; it reflected something deeper about human communication. When someone posts "Having the best day ever!" alongside a picture of themselves looking miserable in the rain, the image provides crucial context that completely changes how we interpret the text. Early attempts to handle this relied mostly on general sentiment analysis techniques, which missed the specific challenges that sarcasm presents.

Schifanella et al. (2016) were among the first to seriously explore multimodal sentiment analysis in social media contexts. While their focus wasn't specifically on sarcasm, their work demonstrated that visual content carries significant emotional and contextual information that can dramatically improve sentiment classification. This laid important groundwork for understanding how visual and textual modalities interact in digital communication.

2.2.2 Dataset Development and Benchmarking

The development of proper benchmarks for multimodal sarcasm detection proved to be a significant challenge. Creating the MMSD dataset, Castro et al. (2019) faced numerous difficult decisions: How do you consistently annotate multimodal sarcasm? What constitutes a fair balance between sarcastic and non-sarcastic examples? How do you account for cultural differences in sarcastic expression?

Their original MMSD dataset was groundbreaking for providing the first comprehensive benchmark specifically designed for text-image sarcasm detection. However, as researchers began working with the dataset, several issues became apparent. Annotation inconsistencies were problematic - different annotators sometimes disagreed about whether content was truly sarcastic. Class imbalance meant models could achieve decent performance by simply predicting the majority class. Perhaps most concerning were potential cultural biases that could limit how well models would generalize across different populations.

Recognizing these limitations, Zhou et al. (2023) and Fu et al. (2023) collaborated on MMSD2.0, which represented a significant improvement in data quality and annotation consistency. Their rigorous crowdsourcing approach achieved a Cohen's Kappa agreement score of 0.811, indicating much more reliable annotations. The improved dataset addressed many of the original concerns: better class balance, more diverse cultural representation, and more comprehensive annotation guidelines.

55 The advent of pictorial social media completely transformed the way individuals convey sarcasm. Suddenly, sarcastic tweet was no longer simply text - it could be a meme, a picture of opposite emotion, or an image totally opposite to the message. Scientists knew they had to evolve.

This development wasn't merely about keeping pace with technology; it spoke to something profound about human interaction. When you write "Having the best day ever!" with a photo of yourself standing in the rain and grinning at the camera, the photo fills in the context necessary to drastically alter the meaning we take from the text. Initial efforts to address this were largely based on generic sentiment analysis methods, which failed to capture the particular difficulties involved in sarcasm.

Schifanella et al. (2016) were the first to seriously investigate multimodal sentiment analysis in social media environments. Although their interest was not necessarily in sarcasm, their work showed that visual material holds rich emotional and contextual information that can significantly enhance sentiment categorization. This provided valuable foundations for understanding visual and textual modality interaction in digital communication.

Establishing good benchmarks for multimodal sarcasm detection was a major challenge. In creating the MMSD dataset, Castro et al. (2019) had to make many challenging decisions: How do you reliably annotate multimodal sarcasm? What is a good balance between examples of sarcasm and non-sarcasm? How do you control for cross-cultural differences in sarcastic expression?

Their initial MMSD dataset was innovative in offering the first standardized benchmark geared toward text-image sarcasm detection. However, as researchers started to work with the dataset, some issues were revealed. Annotation inconsistencies were an issue - different annotators would sometimes differ regarding whether content was actually sarcastic. Class imbalance meant models could perform well by just predicting the majority class. Perhaps most concerning were potential cultural biases that could limit how well models would generalize across different populations.

Understandably, these constraints were realized by Zhou et al. (2023) and Fu et al. (2023) when they worked jointly on MMSD2.0, an enhanced data quality and uniformity of annotations. Their strict crowdsourcing process resulted in a Cohen's Kappa agreement of 0.811, meaning far more consistent annotations. The enhanced dataset resolved most of the issues of the original: improved class balance, greater cultural diversity, and fuller annotation guidelines.

2.2.3 Fusion Strategies and Cross-Modal Learning

Once quality datasets were finally within reach, researchers could set aside the ancillary problem and address the fundamental technical issue: how do you properly integrate textual and visual information for detection of sarcasm? This was more than just concatenating features across modalities.

Cai et al. (2019) investigated hierarchical fusion methods that operated on text and images across several levels of abstraction. Their contribution was that good multimodal fusion may involve comprehending both low-level features (single words, visual features) and high-level relations (semantic intent, emotional contradiction). This hierarchical strategy held promise for reflecting the nuanced manner in which sarcasm can occur across modalities.

Cross-modal attention also emerged as a key area of study. Instead of processing text

and image features separately, these methods enabled models to dynamically connect information across modalities. Liang et al. (2022) created interactive graph networks that were capable of modeling intricate relationships not only between modalities but also between disparate elements within each modality. The method produced outstanding results by capturing the subtle manner in which textual and visual components can reinforce or refute one another.

Visual-semantic embedding methods presented an alternate view to the fusion problem. Rather than treat text and images as separate entities that were subsequently combined, these methods sought to establish integrated representation spaces where the two sources of information could be usefully compared. This was especially useful in ironic contradiction detection - the instances where the visual completely degrades the text.

2.3 Emotion-Aware Approaches in Sarcasm Detection

2.3.1 Theoretical Foundations and Psychological Insights

Grasping the function of emotion in sarcasm isn't only intellectually fascinating - it's also the key to being practically effective. When a coworker declares, "I'm so excited to work this weekend," the emotional subtext is typically worth more than the words on the surface. Sarcastic statements often depend on contradictions of emotion: enjoying terrible news, being surprised at inevitable results, or seeming overly enthusiastic about dull tasks.

Psychological research has long understood that sarcasm tends to involve intentional emotional incongruities. A speaker can say things that, on their face, express feelings directly at odds with what they really feel or with the objective emotional state of affairs. This realization was key to computational solutions - if emotions are at the heart of how sarcasm operates, then emotion detection must be at the heart of sarcasm detectors.

2.3.2 Emotion Integration in Text-Based Systems

The landmark paper by Felbo et al. (2017) using their DeepMoji model introduced tremendous potential for emotion-aware text analysis. Their innovative employment of emojis as emotional surrogates was more than a clever hack - it was a reflection of real understanding about how humans communicate emotions online. The emotional embeddings obtained from emoji patterns were surprisingly effective in boosting sarcasm detection, confirming that emotional awareness could make a strong performance difference.

From there, Chauhan et al. (2020) established more advanced emotion-enriched models that specifically investigated how various affective states interact with sarcastic language. Their findings demonstrated that some emotions - specifically anger, disgust, and surprise - tend to be good indicators of sarcastic intent. This was not correlation; it was real patterns of how individuals use sarcasm to convey complicated emotional states.

More advanced methods have utilized state-of-the-art pre-trained emotion recognition models like the DistilRoBERTa variants that are fine-tuned for emotion classification in particular. These models are able to yield rich emotional information going beyond mere sentiment polarity to capture complex affective states that are important in order to understand sarcastic expressions.

2.3.3 Multimodal Emotion Integration

Taking emotion-aware strategies to multimodal environments offered both promise and challenges. Textual content may register one emotion, but supporting images may provide entirely different emotional information - and such contradictions are frequently exactly what indicate sarcastic intention.

Wang et al. (2022) investigated how emotional context might improve multimodal content analysis in general. Their research showed that models that take into account emotional consistency and contradiction between modalities are more interpretable and stable. In sarcasm detection, in particular, it entails taking note not only of what emotions are being expressed in text and images, but of how those emotions interact with one another.

Visual emotion analysis methods have grown more advanced, progressing beyond mere facial expression recognition to take into account more comprehensive visual signs of emotion: scene composition, color psychology, object association, and contextual visual aspects. These methods allow detection systems to intercept emotional data that may be entirely lacking in or contradictory to the textual message.

2.4 Deep Learning Architectures for Multimodal Processing

2.4.1 Sequential Processing and Recurrent Networks

72

BiLSTM networks have proven particularly effective for sarcasm detection because they can process textual sequences in both directions, providing comprehensive understanding of contextual relationships. This bidirectional capability is crucial for sarcasm, where important contextual cues might appear anywhere in the text sequence.

What particularly suits BiLSTMs for sarcasm is that they can learn about long-range dependencies without losing awareness of local patterns. Sarcastic phrases usually depend on nuanced contradictions that could extend across sentences or paragraphs, making the models work that are capable of carrying over context over long sequences without losing sensitivity to local irregularities.

The coupling of BiLSTM architectures with multimodal fusion approaches has demonstrated promising performance on emotion-aware sarcasm detection. These methods allow for complex modeling of sequential relationships and the inclusion of cross-modal and emotional information within a single framework.

subsectionGraph-Based Architectures

Graph Convolutional Networks have become strong tools for encoding the intricate relationships that define multimodal sarcastic expressions. While sequential models translate information in a linear manner, GCNs are able to encode arbitrary relationships between distinct elements - textual units, visual properties, emotional cues - into a single graph structure.

This is especially useful for sarcasm detection, where the connections between various elements tend to be non-linear and context-specific. A GCN can capture how certain words correspond to visual elements, how emotional cues interact across modalities, and how contextual information affects interpretation - all at once.

Recent studies have proven that GCN-based models have the ability to efficiently learn intricate cross-modal connections with emotional reasoning. These models are able

to learn to recognize patterns such as emotional inconsistency between image and text, semantic discrepancies indicative of ironic intention, and contextual signals that alter interpretation.

subsectionAttention Mechanisms and Cross-Modal Processing

Cross-modal attention mechanisms are perhaps the most intuitive multimodal sarcasm detection strategy. Just as humans tend to search for congruence between what a person claims and how they appear or behave, these models learn to look at salient features in one modality given cues from another.

Yu et al. (2022) illustrated how attention-based methods were able to model sophisticated cross-modal interactions while being able to leverage contextual cues. The research emphasized the significance of dynamic attention allocation - the capacity to attend to various parts of the input based on the content and context.

For sarcasm detection, dynamic attention is vital since the salient cues may change considerably between examples. At times, the critical information is in subtle word usage, at times in expressions, at times in contrast between text and visual. Good attention mechanisms can learn to pick out and attend to these changing cues suitably.

2.5 Current Challenges and Research Gaps

2.5.1 Modality Alignment and Fusion Challenges

Even with the major progress achieved in multimodal sarcasm detection, a number of challenges remain which reduce the efficacy of current methodologies. Multimodal systems tend to find it problematic to align the modalities, especially where textual and visual signals provide conflicting or contradictory information. This issue is particularly significant in sarcastic utterances, where intentional contradictions between modalities are frequently employed to indicate ironic meaning.

The temporal coordination of multimodal information is another critical challenge, especially when the situation involves dynamic content or time-related contextual information. Traditional methods tend to approach visual and textual material as static objects, possibly losing significant temporal associations.

2.5.2 Cultural and Contextual Variations

Cultural differences in the expression of sarcasm create serious challenges for creating generalizable detection systems. Sarcasm is typically based on cultural background knowledge, social norms, and contextual cues that are not necessarily common to all people. This cultural domain restriction makes it challenging for detection systems to generalize across different populations and cultural environments.

subsectionSystematic Emotion Integration Gap

Although emotion-sensitive methods have been promising, prior work has fallen short in providing thorough frameworks that systematically combine emotional aspects across textual and visual modalities. Most recent methods treat emotional information as secondary features instead of being central to the detection process.

Lack of systematic emotional analysis constrains the capability of existing systems to comprehend the emotional undertones which are often inherent in sarcastic statements, leading to lower detection accuracy and contextualization. There is an evident necessity

for unified architectures which can harmoniously interleave emotional comprehension with sophisticated multimodal processing capabilities.

2.6 Summary and Research Motivation

This extensive literature review presents a rich and dynamic body of research in sarcasm detection, marked by ongoing progress from basic rule-based systems to complex multimodal setups. Significant limitations remain in existing approaches, especially in rigorous integration of emotional comprehension with cutting-edge architectural modules.

The literature shows that emotion-conscious, multimodal deep learning-based methods are the state-of-the-art in sarcasm detection. Yet, there is an evident lack of integrated frameworks that consistently incorporate sequential modeling (BiLSTM), graph-based relational modeling (GCN), cross-modal attention, and in-depth emotion integration.

This thesis fills these gaps identified by suggesting and testing emotion-aware multimodal architectures that integrate emotional aspects of both text and images systematically. The study employs newer datasets such as MMSD2.0 and state-of-the-art models to set new baselines for emotion-aware sarcasm detection and to advance both theoretical insights and practical applications in this vital field.

Chapter 3

METHODOLOGY

3.1 Overview

This chapter presents a detailed description of the methodological approach employed in developing and evaluating the emotion-aware multimodal sarcasm detection system. The research followed a systematic progression from simple to complex architectures, with extensive ablation studies to validate the contribution of emotional features. Each architectural decision was backed by empirical evidence, particularly focusing on the impact of emotional feature integration.

To ensure comprehensive evaluation and validation of our proposed approach, this study employs a multi-dataset experimental framework incorporating three distinct multimodal sarcasm detection datasets: MMSD2.0, the original MMSD dataset, and the MEMOTION dataset. This multi-dataset approach enables robust evaluation of model generalization capabilities across diverse data distributions and annotation schemes, while providing insights into the effectiveness of emotion-aware features across different contextual domains.

3.2 Datasets and Preprocessing

3.2.1 Primary Dataset: MMSD2.0

The MMSD2.0 dataset serves as primary evaluation benchmark, designed for multimodal sarcasm detection containing paired text and images that are annotated for sarcasm. Here's a breakdown of what it includes:

- Text: Captions, comments, or sentences associated with the image. These are often sarcastic remarks paired with the visual content.
- Images: The accompanying visual component is a key feature of the dataset, intended to provide contextual cues for detecting sarcasm.

Type of Images: Unlike benchmark datasets like MuSTARD, which may use scripted TV show excerpts, MMSD2.0 focuses on real-world social media data. The images are diverse and context-dependent, primarily sourced from social media platforms and are not restricted to memes or TV shows but include: Memes (Some entries are memes where the text and image combine to deliver sarcastic humor) and Generic Images (Others are random social media posts where the sarcasm emerges from the interplay between text and image) .

Figure 3.1: Dataset Sample Image

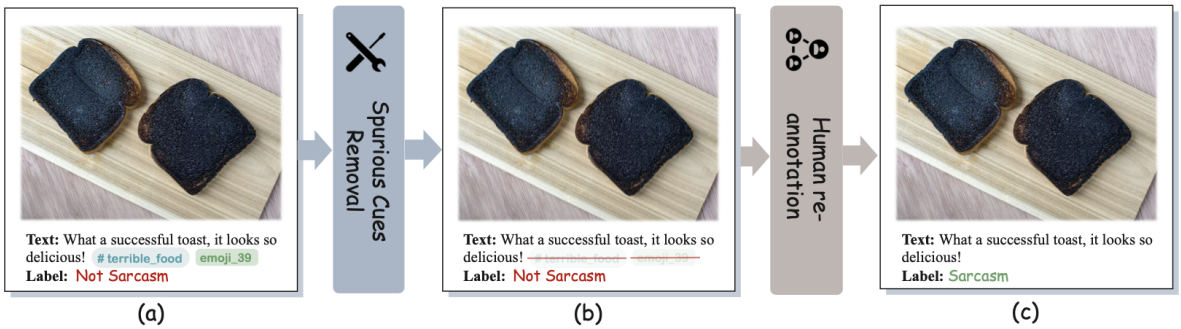

Figure 3.2: Removal of spurious cues and reannotation

The MMSD2.0 dataset is an updated version of the original Multi-modal Sarcasm Detection (MMSD) dataset, designed to improve the reliability of sarcasm detection tasks. The dataset now has a more even split between sarcastic (positive) and non-sarcastic (negative) samples across training, validation, and testing sets, enhancing its suitability for machine learning tasks. This revision addresses several shortcomings in the original dataset, such as:

Elimination of Spurious Cues: The MMSD dataset had biases, such as an over-reliance on specific hashtags or emoji distributions, which models could exploit without truly understanding multimodal sarcasm. MMSD2.0 removes such spurious cues to encourage genuine sarcasm feature learning.

Re-annotation for Quality Control: Negative samples in MMSD were found to include many unreasonable entries, as not having a "sarcasm" tag doesn't guarantee a lack of sarcasm. MMSD2.0 re-annotates these entries via a rigorous crowdsourcing process to improve data quality and balance.

The experiments were conducted using the MMSD2.0 dataset, which contains multi-modal posts comprising text and image pairs. The dataset is divided into three splits: training, validation, and test sets. Table 3.1 presents the detailed statistics for each split.

The training set text length varies from 1 to 66 words, with an average of 13.42 words per post. All splits maintain complete image availability, ensuring a fully multimodal dataset. The class distribution shows a slight imbalance, with the training set being nearly balanced (48.32% sarcastic), while validation and test sets have a marginally higher proportion of non-sarcastic samples.

Metric	Training Set	Validation Set	Test Set
Total Samples	19,816	2,410	2,409
Sarcastic Samples	9,576	1,042	1,037
Sarcastic Percentage	48.32%	43.24%	43.05%
Non-sarcastic Samples	10,240	1,368	1,372
Non-sarcastic Percentage	51.68%	56.76%	56.95%
Average Words per Text	13.42	13.64	13.52
Images Available	100%	100%	100%

Table 3.1: Summary of MMSD2.0 dataset splits with metric comparisons.

3.2.2 Auxiliary Dataset 1: Original MMSD

To evaluate the robustness of our emotion-aware approach across different annotation methodologies and data quality standards, we incorporated the original MMSD dataset in our experimental framework. The original MMSD dataset represents the foundational work in multimodal sarcasm detection and provides valuable insights into the evolution of annotation quality and dataset construction practices.

The original MMSD dataset consists of 24,635 multimodal social media posts collected from various platforms, primarily Twitter, where sarcasm annotations were initially derived from hashtag-based heuristics. Unlike MMSD2.0, the original dataset exhibits several characteristics that make it particularly valuable for robustness evaluation:

- **Annotation Methodology**: The original annotations were derived from the presence or absence of explicit sarcasm indicators such as hashtags (sarcasm, sarcastic), which introduces natural noise and ambiguity typical of real-world social media data.
- **Class Distribution**: The original dataset exhibits a more pronounced class imbalance with approximately 52.1% sarcastic samples, providing an opportunity to evaluate model performance under different distributional assumptions.
- **Linguistic Diversity**: The dataset contains a broader range of informal language patterns, abbreviations, and social media conventions that were preserved in the original collection process.
- **Image Quality Variability**: The original dataset includes images with varying quality levels and processing artifacts that were not filtered in the initial collection phase.

The dataset is structured with three primary splits containing textual content, image identifiers, and binary sarcasm labels. Our preprocessing pipeline adapted this structure to maintain consistency with our emotion-aware feature extraction framework while preserving the original distributional characteristics that make this dataset valuable for cross-dataset generalization studies.

3.2.3 Auxiliary Dataset 2: MEMOTION

The MEMOTION dataset expands our evaluation framework beyond traditional social media posts to encompass meme-based multimodal content, representing a distinct yet

related domain for sarcasm detection research. This dataset provides 6,993 carefully curated meme images with accompanying textual content, offering a specialized context where visual-textual interplay is particularly pronounced.

The MEMOTION dataset exhibits several unique characteristics that complement our primary evaluation framework:

- **Domain Specialization**: Unlike generic social media posts, memes represent a distinct communicative medium where humor, sarcasm, and social commentary are typically conveyed through carefully crafted image-text combinations.
- **Multi-dimensional Annotation**: Beyond binary sarcasm labels, MEMOTION provides annotations for humor intensity, offensiveness levels, motivational content, and overall sentiment, enabling comprehensive analysis of emotional and attitudinal dimensions.
- **Image Diversity**: The dataset contains 4,965 JPG and 1,679 PNG images spanning diverse visual styles, from traditional meme formats to original social media graphics.
- **Text Processing Complexity**: Textual content includes both OCR-extracted text and manually corrected versions, providing opportunities to evaluate the impact of text quality on multimodal sarcasm detection performance.

The annotation scheme for MEMOTION includes hierarchical sarcasm categories (not_sarcastic, general, twisted_meaning, very_twisted) which we systematically mapped to binary classifications for compatibility with our primary evaluation framework. This mapping preserves the nuanced understanding of sarcasm intensity while enabling direct comparison with other datasets in our study.

Table 3.2 presents the statistical characteristics of the MEMOTION dataset:

Metric	Value
Total Samples	6,993
Training Samples	4,895 (70%)
Validation Samples	700 (10%)
Test Samples	1,398 (20%)
Sarcastic Samples (Binary)	3,487 (49.8%)
Non-sarcastic Samples	3,506 (50.2%)
Average Text Length	12.3 words
Image Formats	JPG (71.0%), PNG (29.0%)

Table 3.2: Statistical summary of MEMOTION dataset characteristics.

To contextualize the characteristics and advantages of our multi-dataset approach, we present a comprehensive comparison with other prominent multimodal sarcasm detection datasets in Table 3.3.

Our multi-dataset approach provides several methodological advantages:

- **Cross-Domain Validation**: By incorporating datasets from different domains (general social media, curated social media, and memes), we can evaluate the generalizability of emotion-aware features across diverse communicative contexts.

Dataset	Year	Total Samples	Modalities	Sarcastic %	Domain
MMSD	2019	24,635	Text+Image	52.1%	Social Media
MMSD2.0	2023	24,635	Text+Image	47.8%	Social Media
MEMOTION	2020	6,993	Text+Image	49.8%	Memes
MUStARD	2019	690	Text+Audio+Video	34.5%	TV Shows
MUStARD++	2022	1,573	Text+Audio+Video	41.2%	TV Shows
SARC	2017	1.3M	Text only	25.4%	Reddit

Table 3.3: Comprehensive Comparison of Multimodal Sarcasm Detection Datasets

- **Annotation Quality Spectrum**: The inclusion of both original MMSD and MMSD2.0 enables systematic analysis of how annotation quality improvements affect model performance and the relative contribution of emotional features.
- **Scale Diversity**: Our dataset combination provides evaluation across different scales, from the large-scale MMSD collections to the more focused MEMOTION dataset, enabling analysis of data efficiency and feature effectiveness relationships.
- **Balanced Evaluation**: The near-balanced class distributions across all three datasets (47.8%-52.1% sarcastic) ensure that our performance evaluations are not confounded by extreme class imbalance issues common in other datasets.

3.2.4 Emotion Feature Extraction

The emotion feature extraction process was implemented using a comprehensive framework that extracts emotional features from both text and image modalities. The system employs specialized models for each modality while ensuring consistent emotion categorization across both channels.

Text Emotion Extraction Using DistilRoBERTa

For text emotion extraction, we utilize the emotion-english-distilroberta-base model, which processes text sequences up to 512 tokens in length. The model provides probability distributions across seven fundamental emotion categories: anger, disgust, fear, joy, neutral, sadness, and surprise.

The text emotion classification process begins with tokenization and encoding through DistilRoBERTa's transformer layers, producing contextualized representations. The emotion classification can be mathematically formulated as:

$$\mathbf{h}_{text} = DistilRoBERTa(\mathbf{x}_{tokens}) \tag{3.1}$$

where $\mathbf{x}_{tokens}$ represents the tokenized input text and $\mathbf{h}_{text} \in R^{768}$ is the contextualized representation from the [CLS] token.

The emotion probabilities are then computed using a classification head:

$$\mathbf{p}_{text} = softmax(W_{emotion}\mathbf{h}_{text} + \mathbf{b}_{emotion}) \tag{3.2}$$

where $W_{emotion} \in R^{7\times768}$ is the emotion classification weight matrix, $\mathbf{b}_{emotion} \in R^7$ is the bias vector, and $\mathbf{p}_{text} \in R^7$ represents the probability distribution over the seven emotion categories.

The final text emotion feature vector combines the dominant emotion confidence score with a one-hot encoding:

$$\mathbf{e}_{text} = [\max(\mathbf{p}_{text}), onehot(\arg\max(\mathbf{p}_{text}))] \tag{3.3}$$

resulting in an 8-dimensional feature vector where the first component captures emotion confidence and the remaining seven components provide categorical emotion information.

Image Emotion Extraction Using Modified ResNet50

Image emotion processing employs a modified ResNet50 architecture with a custom emotion classification head. The ResNet50 backbone processes normalized images through convolutional layers to extract hierarchical visual features, which are then mapped to the same seven emotion categories used for text.

The image processing pipeline follows these mathematical steps. First, the input image undergoes preprocessing normalization:

$$\mathbf{I}_{norm} = \frac{\mathbf{I} - \mu}{\sigma} \tag{3.4}$$

where $\mathbf{I} \in R^{3\times224\times224}$ is the input image, μ and σ are ImageNet normalization parameters.

The ResNet50 feature extraction produces high-dimensional visual representations:

$$\mathbf{f}_{img} = ResNet50(\mathbf{I}_{norm}) \tag{3.5}$$

where $\mathbf{f}_{img} \in R^{2048}$ represents the global average pooled features from the final convolutional layer.

The visual emotion classification head maps these features to emotion probabilities:

$$\mathbf{p}_{img} = softmax(W_{visual}\mathbf{f}_{img} + \mathbf{b}_{visual}) \tag{3.6}$$

where $W_{visual} \in R^{7\times2048}$ and $\mathbf{b}_{visual} \in R^7$ are the visual emotion classification parameters.

Similar to text emotions, the final image emotion feature vector is constructed as:

$$\mathbf{e}_{img} = [\max(\mathbf{p}_{img}), onehot(\arg\max(\mathbf{p}_{img}))] \tag{3.7}$$

The emotion classification system operates across seven fundamental emotional categories: anger, disgust, fear, joy, neutral, sadness, and surprise. This standardized set enables direct comparison and integration of emotional signals from both text and image modalities, facilitating sophisticated multimodal analysis of sarcastic content.

3.2.5 Data Preparation

The data preparation process is implemented through the EmotionEnrichedDataset class, which creates PyTorch-compatible datasets optimized for efficient training and evaluation. The pipeline begins with loading fundamental data components: JSON files containing text and image references, pre-extracted emotion features from both modalities, and corresponding sarcasm labels.

Text processing employs BERT tokenization with a fixed sequence length of 128 tokens, including automatic padding, truncation, and attention mask generation. Image processing implements comprehensive transformations: RGB conversion, resizing to 224×224 pixels, tensor conversion, and normalization with ImageNet parameters. The system includes robust fallback mechanisms for corrupted or missing data.

The emotion feature processing converts extracted emotion data into tensors suitable for model training, combining probability scores with one-hot encoded representations. The batch preparation process is optimized with configurable batch size (default: 32), multi-worker data loading (4 workers), training set shuffling, and memory-efficient handling.

The final processed samples are structured with text (input_ids and attention_mask), normalized images (3×224×224), text and image emotion features (8-dimensional each), binary sarcasm labels, and sample identifiers. This comprehensive pipeline ensures consistent input formats, efficient batch processing, robust error handling, rich emotional feature representation, and memory-efficient data loading.

The emotion feature processing stage converts the extracted emotion data into a format suitable for model training. This involves transforming emotion probabilities into tensors and creating one-hot encoded representations of dominant emotions. The system combines these into a unified representation that includes both probability scores and one-hot encoded emotion vectors.

The final processed samples are structured as follows. Each sample in the processed dataset contains:

- id: Unique identifier
- text: Original text content
- label: Sarcasm label (1 for sarcastic, 0 for non-sarcastic)
- text_emotion:
  - 'top_emotion': Most probable emotion
  - 'confidence': Confidence score
  - 'emotion_probs': Probabilities for all emotions
- image_emotion:
  - 'top_emotion': Most probable emotion
  - 'confidence': Confidence score
  - 'emotion_probs': Probabilities for all emotions

3.3 Pretrained Models

The multimodal sarcasm detection system leverages several state-of-the-art pretrained models as foundation components, each selected for their proven effectiveness in their respective domains and their compatibility with our multimodal architecture requirements.

3.3.1 Text Encoding Models

For textual feature extraction, we primarily employed RoBERTa-base (Robustly Optimized BERT Pretraining Approach), a transformer-based language model that builds upon BERT's architecture with several key improvements. RoBERTa was selected over the original BERT model due to its enhanced pretraining methodology, which includes dynamic masking, removal of the Next Sentence Prediction task, and training on larger datasets with longer sequences.

The RoBERTa model processes input text through multiple transformer layers, where each layer applies self-attention and feed-forward transformations. The self-attention mechanism can be expressed as:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{3.8}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are the query, key, and value matrices respectively, and d_k is the dimension of the key vectors. The model provides 768-dimensional contextual embeddings that capture rich semantic and syntactic relationships within textual content.

The RoBERTa-base model consists of 12 transformer layers with 768 hidden units and 12 attention heads, totaling approximately 125 million parameters. To optimize computational efficiency while preserving the model's representational power, we implemented a selective fine-tuning strategy. Specifically, the initial eight transformer layers were frozen to retain general language understanding capabilities, while the final four layers remained trainable to adapt to the specific nuances of sarcastic language detection. This approach balances the benefits of pretrained knowledge with task-specific adaptation.

3.3.2 Visual Feature Extraction Models

For image feature extraction, we employed ResNet50 (Residual Network with 50 layers) as our primary visual encoder, leveraging weights pretrained on ImageNet. ResNet50 was chosen for its excellent balance between computational efficiency and feature extraction capability, along with its proven effectiveness in various computer vision tasks.

The ResNet50 architecture employs residual connections that enable training of deeper networks while mitigating the vanishing gradient problem. The core residual block can be mathematically represented as:

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x} \tag{3.9}$$

where $\mathbf{x}$ is the input, $\mathcal{F}(\mathbf{x}, \{W_i\})$ represents the residual mapping to be learned, and $\mathbf{y}$ is the output. This formulation allows gradients to flow directly through the identity connection, facilitating effective training of deep networks.

The ResNet50 architecture consists of four main residual blocks with bottleneck structures, culminating in a global average pooling layer that produces 2048-dimensional feature vectors. We modified the standard ResNet50 by removing the final classification layer and utilizing the feature maps from the penultimate layer as our image representations. Similar to our text encoder optimization, we implemented selective layer freezing: the initial convolutional blocks (conv1 through layer3) were frozen to preserve low-level and mid-level visual features learned from ImageNet, while layer4 remained trainable to adapt to sarcasm-specific visual patterns.

3.3.3 Emotion Recognition Models

For emotion feature extraction from textual content, I utilized the emotion-english-distilroberta-base model, a specialized variant of DistilRoBERTa that has been specifically fine-tuned for emotion classification tasks. This model provides probability distributions across seven fundamental emotion categories: anger, disgust, fear, joy, neutral, sadness, and surprise. The model's architecture is based on the distilled version of RoBERTa, which maintains much of the original model's performance while requiring significantly fewer computational resources.

The DistilRoBERTa emotion model processes input text through its transformer layers to generate contextualized representations, which are then passed through a classification head to produce emotion probabilities. The model outputs a probability distribution over the seven emotion categories, where each probability value indicates the likelihood of the text expressing that particular emotion. This pre-trained model eliminates the need for training a custom text emotion classifier, providing reliable emotion detection capabilities that have been validated across multiple emotion recognition benchmarks.

For visual emotion recognition, we implemented a custom emotion classification head built upon the ResNet50 backbone. This approach involves training a specialized classification layer that maps ResNet50's 2048-dimensional features to the same seven emotion categories used for text. The visual emotion classifier is trained to recognize emotional expressions and contextual emotional cues present in images, enabling consistent emotion representation across both modalities.

The integration of consistent emotion categorization across text and image modalities is crucial for our multimodal fusion approach, as it allows for direct comparison and interaction between emotional signals from different sources. Both emotion models output 8-dimensional feature vectors consisting of the probability score for the dominant emotion and a one-hot encoded representation of the emotion categories.

3.4 Model Architectures

3.4.1 Initial Base Model

The research began with a deliberately simple architecture to establish baseline performance and validate the fundamental approach. The base model consists of two primary components: feature extractors for each modality and a fusion-classification pipeline.

For text processing, we employed BERT-base-uncased as the backbone encoder, which produces 768-dimensional text features. These features undergo a basic projection to ensure compatibility with the fusion layer. The image processing pipeline utilizes ResNet50 as the feature extractor, generating 2048-dimensional image features, which are similarly projected to match the fusion requirements.

The fusion and classification architecture implements a straightforward yet effective approach. The fusion layer consists of a sequential network with multiple fully connected layers interspersed with ReLU activations and dropout regularization. Specifically, the fusion layer transforms the concatenated multimodal features through two hidden layers of size hidden_size $\times$ 2, with dropout applied after each transformation to prevent overfitting.

The classification head is implemented as a simple sequential network that takes the fused features and produces the final sarcasm prediction. It consists of a single linear layer that reduces the hidden representation to a scalar, followed by a sigmoid activation

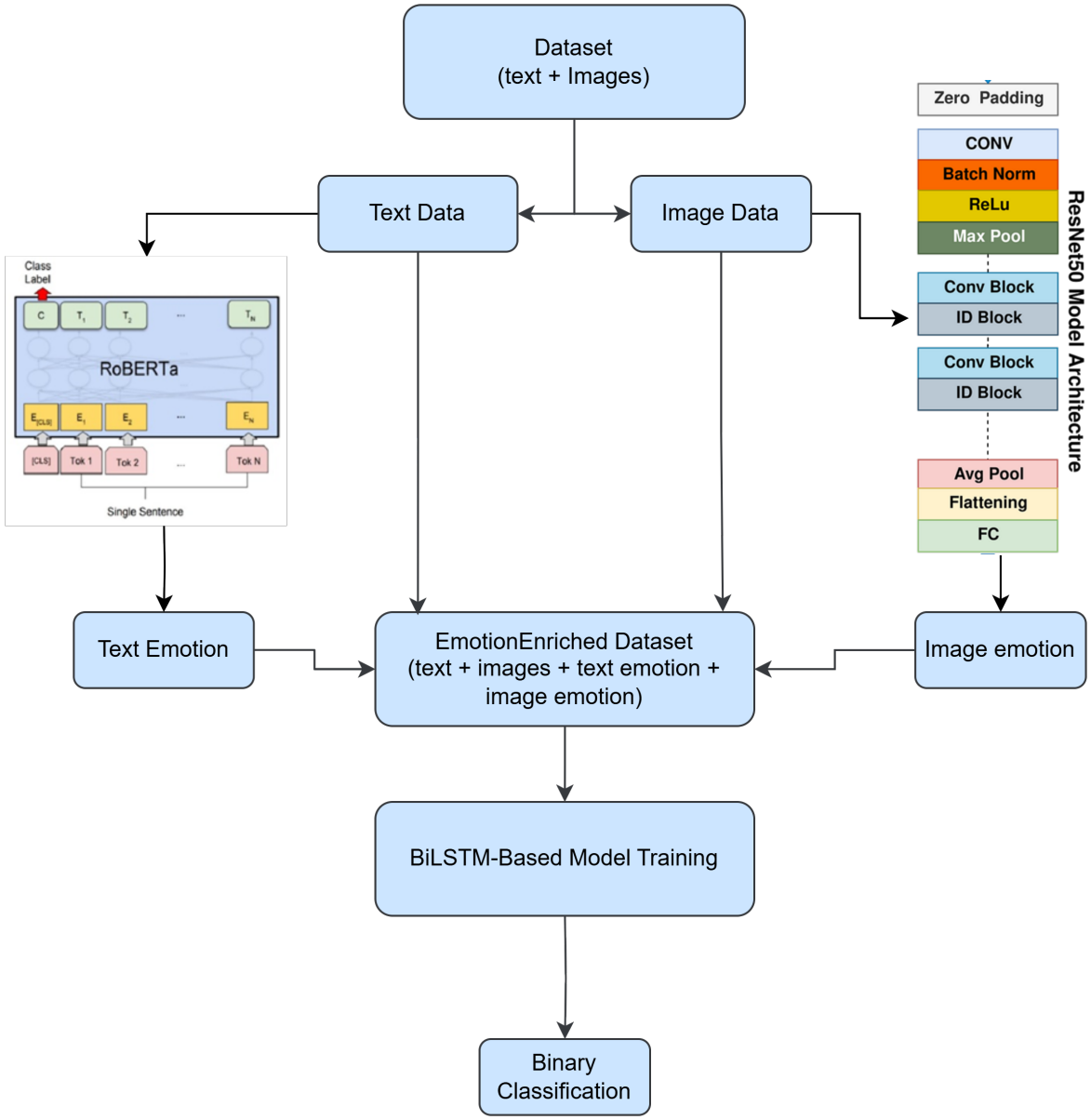

Figure 3.3: System Architecture Workflow

for binary classification.

To systematically evaluate the contribution of emotional features, we explored four distinct model variants:

1. Base Combination: A simple concatenation of text and image features, establishing the fundamental multimodal baseline.
2. Text Emotion Integration: Enhancing the base model with 8-dimensional text emotion features, allowing us to assess the impact of textual emotional context.
3. Image Emotion Integration: Incorporating 8-dimensional image emotion features to evaluate the contribution of visual emotional signals.
4. Full Emotion Integration: A comprehensive model combining all available features (text, image, and both emotion modalities) to leverage the complete multimodal emotional context.

Each variant underwent extensive hyperparameter optimization, exploring the following ranges:

- Dropout rates: [0.1, 0.3, 0.4, 0.5]
- Weight decay: [0.01, 0.001, 0.005]
- Learning rate adjustments
- Batch size optimization

The comparative performance analysis of these model variants, including accuracy, precision, recall, and F1-scores achieved with the optimal hyperparameter configurations, is presented in detail in Chapter 4 (Results and Discussion). This analysis provides insights into the relative contribution of each emotional modality to the overall sarcasm detection task.

3.4.2 BiLSTM Architecture

Building upon the initial base model, I developed an enhanced BiLSTM-based architecture that integrates textual, visual, and emotional features for improved sarcasm detection. The model processes three input streams: text through RoBERTa, images via ResNet50, and emotion features from both modalities. This architecture represents a significant advancement over traditional approaches by incorporating sequential modeling capabilities alongside multimodal fusion.

The text processing component employs RoBERTa-base as the encoder, utilizing a selective fine-tuning strategy where the first eight layers remain frozen to preserve pre-trained knowledge, while the final four layers are trainable for task-specific adaptation. This approach balances computational efficiency with domain-specific learning, preventing catastrophic forgetting while enabling adaptation to sarcastic language patterns. The RoBERTa encoder produces 768-dimensional contextual representations that capture semantic relationships and linguistic nuances essential for sarcasm detection.

For image processing, I use ResNet50 with ImageNet pre-trained weights, implementing a similar selective fine-tuning approach where the initial convolutional layers remain

frozen and only the final layer is trainable for domain adaptation. The visual features are extracted from the global average pooling layer, providing 2048-dimensional representations that capture both low-level visual patterns and high-level semantic content. These visual features are particularly important for understanding the contextual relationship between text and accompanying images in sarcastic posts.

The core BiLSTM component consists of a bidirectional LSTM with 2 layers that processes the combined multimodal features sequentially. Unlike traditional feed-forward approaches, the BiLSTM architecture can capture temporal dependencies and contextual relationships that are crucial for understanding sarcastic expressions. The bidirectional processing captures sequential dependencies in both forward and backward directions:

$$\mathbf{h}_t = [\overrightarrow{\mathbf{h}_t}; \overleftarrow{\mathbf{h}_t}] \tag{3.10}$$

where $\overrightarrow{\mathbf{h}_t}$ and $\overleftarrow{\mathbf{h}_t}$ represent the forward and backward hidden states respectively. This bidirectional approach ensures that the model considers the full context when making predictions, which is particularly valuable for sarcasm detection where meaning often depends on the complete sequence rather than individual elements.

The emotion integration incorporates 8-dimensional emotion vectors extracted from both text (via DistilRoBERTa emotion model) and images (through a custom emotion classification head trained on visual emotion recognition). These emotional features provide crucial affective context for detecting sarcastic expressions, which often involve emotional incongruence between the surface text and underlying sentiment. The emotion vectors capture seven primary emotions (anger, disgust, fear, joy, neutral, sadness, surprise) plus confidence scores, enabling the model to understand the emotional dynamics that characterize sarcastic content.

The fusion architecture combines all modality features through dedicated projection layers that map different feature spaces to a common dimensional space before concatenation. Dropout regularization (0.3) is applied throughout the network to prevent overfitting, particularly important given the multimodal nature of the input. The fusion strategy ensures that each modality contributes meaningfully to the final prediction while maintaining balanced representation across text, image, and emotion features.

The classification head processes the fused representation through a series of fully connected layers with progressively decreasing dimensions, enabling the model to learn complex decision boundaries for sarcasm detection. The final layer produces binary sarcasm predictions using sigmoid activation. The model is trained using BCEWithLogitsLoss for numerical stability, with additional techniques including gradient clipping to prevent exploding gradients and early stopping based on F1-score performance to optimize for the primary evaluation metric.

This BiLSTM-based architecture addresses several key challenges in multimodal sarcasm detection: sequential processing of multimodal features, effective integration of emotional context, and robust generalization across different datasets and domains. The selective fine-tuning strategy ensures efficient computation while maintaining model expressiveness, making it suitable for practical deployment scenarios.

3.4.3 Graph Convolutional Network Architecture

Building upon the success of the BiLSTM architecture, we implemented a Graph Convolutional Network (GCN) approach to model the complex relationships between different modalities and their emotional contexts. The GCN architecture represents a paradigm

shift from sequential processing to graph-based relational modeling, where each modality and its associated emotional features are treated as nodes in a dynamically constructed graph.

The GCN model architecture consists of four primary node types within each sample: text features, image features, text emotion features, and image emotion features. Each node is represented by a 256-dimensional feature vector obtained through dedicated projection layers from their respective pretrained encoders.

The graph construction process creates a fully connected structure between all four nodes within each sample, resulting in 12 directed edges per sample (4 nodes $\times$ 3 connections each). This connectivity pattern ensures that every modality can directly influence and be influenced by every other modality, enabling comprehensive cross-modal interaction modeling.

The core GCN processing consists of three consecutive graph convolutional layers, each followed by ReLU activation and dropout regularization. The graph convolution operation can be mathematically expressed as:

$$\mathbf{H}^{(l+1)} = \sigma\left(\mathbf{A}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right) \tag{3.11}$$

where $\mathbf{H}^{(l)}$ represents the node features at layer l, $\mathbf{A}$ is the adjacency matrix with learned edge weights, $\mathbf{W}^{(l)}$ are the layer-specific weight parameters, and σ is the activation function.

Each GCN layer aggregates information from neighboring nodes through the weighted edge connections, allowing features to propagate and interact across the multimodal graph structure. This iterative message-passing mechanism enables the model to capture complex, higher-order relationships between modalities that may not be evident through simple concatenation or attention-based fusion approaches.

The attention-based edge weighting mechanism is implemented through a two-layer neural network that takes concatenated node features as input and outputs a scalar weight between 0 and 1. This dynamic weighting allows the model to learn which cross-modal interactions are most relevant for sarcasm detection, potentially discovering novel patterns such as the relationship between specific textual emotions and visual content, or the interplay between image emotions and textual semantics.

Following the GCN processing, node features are reshaped and concatenated to form a comprehensive multimodal representation. This representation is then passed through a two-layer classification network with ReLU activation and dropout regularization to produce the final sarcasm prediction. The entire architecture is optimized using BCEWithLogitsLoss and AdamW optimizer with gradient clipping to ensure stable training.

Chapter 4

RESULTS AND DISCUSSION

4.1 Evaluation Metrics

The performance of each model variant was evaluated using a comprehensive set of classification metrics specifically chosen for their relevance to binary sarcasm detection tasks:

- **Accuracy**: Overall correctness of predictions, calculated as the ratio of correct predictions to total predictions: $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
- **Precision**: Proportion of correct positive predictions among all positive predictions: $Precision = \frac{TP}{TP+FP}$
- **Recall (Sensitivity)**: Proportion of actual positives correctly identified: $Recall = \frac{TP}{TP+FN}$
- **F1-Score**: Harmonic mean of precision and recall, providing a balanced measure: $F1 = 2 \times \frac{Precision \times Recall}{Precision+Recall}$

where TP represents true positives, TN true negatives, FP false positives, and FN false negatives. These metrics collectively provide a comprehensive evaluation framework that captures both the model's ability to correctly identify sarcastic content and its precision in avoiding false classifications.

4.2 Experimental Results

4.2.1 Ablation Study: Emotion Feature Contribution Analysis

The initial experiments focused on systematically validating the contribution of emotional features through comprehensive ablation studies. This analysis provides crucial insights into the individual and combined effects of textual and visual emotional information on sarcasm detection performance.

Baseline Performance Analysis

The baseline model utilizing only text and image features achieves an accuracy of 77.75% with an F1-score of 0.7603. The relatively high recall (0.8197) compared to precision (0.7089) indicates the model's tendency toward liberal classification, producing more false positives than false negatives. This pattern suggests that the baseline model successfully captures many sarcastic instances but struggles with precision, potentially due to the absence of emotional context that could help distinguish subtle sarcastic cues.

Model Variant	Accuracy	Precision	Recall	F1-Score
Base (Text + Image)	0.7775	0.7089	0.8197	0.7603
With Text Emotion	0.7883	0.7127	**0.8515**	0.7751
With Image Emotion	0.7862	0.7219	0.8187	0.7673
With Both Emotions	**0.7908**	**0.7399**	0.8127	**0.7754**

Table 4.1: Performance metrics for different emotion integration strategies.

Individual Emotion Modality Impact

The integration of text emotion features yields a notable improvement, elevating accuracy to 78.83% and F1-score to 0.7751. Most significantly, recall increases substantially to 0.8515, representing a 3.18 percentage point improvement over the baseline. This enhancement demonstrates that textual emotional signals provide valuable contextual information for identifying sarcastic expressions that might otherwise be missed.

Image emotion integration also improves upon baseline performance, achieving 78.62% accuracy and 0.7673 F1-score. While the improvements are more modest compared to text emotions, the consistent gains across all metrics validate the contribution of visual emotional cues. The precision improvement (0.7219 vs. 0.7089) suggests that image emotions help reduce false positive classifications.

Synergistic Emotion Integration

The combination of both text and image emotion features achieves the optimal performance across most metrics, with 79.08% accuracy and the highest precision (0.7399). This configuration demonstrates the complementary nature of multimodal emotional signals, where textual and visual emotions provide different but synergistic information for sarcasm detection. The balanced recall (0.8127) and precision indicate that dual-modal emotion integration helps achieve a more robust classification boundary.

The consistent improvements across all emotion integration strategies provide strong empirical evidence supporting my hypothesis that emotional features are crucial for effective multimodal sarcasm detection. The incremental gains validate the systematic approach to emotion integration and establish the foundation for more sophisticated architectural enhancements.

4.2.2 Hyperparameter Optimization Analysis

Extensive hyperparameter optimization was conducted to identify optimal configurations for maximizing model performance. The exploration covered critical hyperparameters including dropout rate, learning rate, and weight decay, with systematic grid search across feasible parameter ranges.

The optimization results reveal several important patterns. The optimal configuration (dropout=0.3, learning rate=1e-4, weight decay=0.001) achieves 80.81% validation accuracy, representing a significant improvement over suboptimal settings. The moderate dropout rate of 0.3 appears optimal for balancing regularization with model capacity, while higher dropout (0.5) leads to underfitting. The learning rate of 1e-4 provides stable convergence, and the reduced weight decay (0.001) allows for better parameter adaptation while maintaining generalization.

Dropout Rate	Learning Rate	Weight Decay	Validation Accuracy
0.1	1e-4	0.01	0.7842
0.3	1e-4	0.01	0.7897
0.3	2e-4	0.01	0.7865
0.3	1e-4	0.001	**0.8081**
0.5	1e-4	0.001	0.7820
0.4	1e-4	0.001	0.7908

Table 4.2: Optimal hyperparameter configurations and corresponding validation accuracies.

4.2.3 Advanced Architecture Performance: BiLSTM and GCN

The implementation of sophisticated architectures incorporating BiLSTM and GCN approaches represents the culmination of my methodological progression, demonstrating substantial performance improvements over the baseline emotion-enriched models.

Architecture	Accuracy	Precision	Recall	F1-Score
BiLSTM	**0.8312**	**0.8110**	**0.8507**	**0.8304**
GCN	0.8129	0.8080	0.8317	0.8197

Table 4.3: Comparative performance of advanced architectural approaches.

BiLSTM Architecture Analysis

The BiLSTM model achieves exceptional performance with 83.12% accuracy and 83.04% F1-score, representing a significant advancement over traditional approaches. The balanced metrics (precision: 81.10%, recall: 85.07%) indicate robust classification capability across both positive and negative instances. The high recall suggests excellent sensitivity to sarcastic content, while the strong precision demonstrates effective false positive control.

The sequential processing capability of BiLSTM appears particularly well-suited for capturing the temporal relationships between multimodal features and their associated emotions. The bidirectional nature enables comprehensive context understanding, allowing the model to leverage both past and future information when making classification decisions.

Graph Convolutional Network Analysis

The GCN architecture delivers highly competitive performance (81.29% accuracy, 81.97% F1-score), validating the graph-based relational modeling approach. The strong precision (80.80%) and recall (83.17%) demonstrate the effectiveness of treating modalities and emotions as interconnected graph nodes, enabling sophisticated cross-modal relationship learning.

The results clearly show that both the BiLSTM and GCN architectures benefit greatly from architectural enhancements and the integration of emotion features. The BiLSTM model achieves the highest overall performance, but the GCN model also delivers competitive results, highlighting the value of graph-based relational modeling. The high

recall and F1 scores for both models indicate their effectiveness in identifying sarcastic instances while maintaining a balanced precision. These findings underscore the importance of emotion-aware multimodal fusion and suggest that both sequential and graph-based approaches are promising for effective sarcasm detection.

4.3 Discussion

4.3.1 Key Findings and Performance Analysis

The experimental results demonstrate that both the BiLSTM and GCN architectures achieve strong, state-of-the-art performance for multimodal sarcasm detection when enhanced with emotion features. The BiLSTM model achieves the highest overall metrics, with an accuracy of 83.12%, precision of 81.10%, recall of 85.07%, and an F1 score of 83.04%. The GCN model also performs competitively, with an accuracy of 81.29%, precision of 80.80%, recall of 83.17%, and an F1 score of 81.97%. These results highlight the effectiveness of both sequential (BiLSTM) and graph-based (GCN) relational modeling approaches, especially when emotion information is integrated.

The high recall and F1 scores for both models indicate their robustness in identifying sarcastic instances while maintaining balanced precision. The slight edge of the BiLSTM model suggests that sequential modeling may be more effective for this task, but the strong performance of the GCN model underscores the promise of graph-based methods for capturing complex multimodal relationships.

Model (Paper)	Accuracy	Precision	Recall	F1-Score
Text-modality Methods				
TextCNN (Kim) [15]	71.61	64.62	75.22	69.52
BiLSTM (Graves & Schmidhuber) [11]	72.48	68.02	68.08	68.05
SMSD (Xiong et al.) [28]	73.56	68.45	71.55	69.97
Image-modality Methods				
ResNet (He et al.) [13]	65.50	61.17	54.39	57.58
ViT (Dosovitskiy et al.) [7]	72.02	65.26	74.83	69.72
Multi-Modality Methods				
HFM (Cai et al.) [2]	70.57	64.84	69.05	66.88
HKE (Liu et al.) [18]	76.50	73.48	71.07	72.25
Multi-view CLIP (Qin et al.) [23]	85.64	80.33	88.24	84.10
BiLSTM + Emotions (This Work)	**83.12**	**81.10**	**85.07**	**83.04**
GCN + Emotions (This Work)	**81.29**	**80.80**	**83.17**	**81.97**

Table 4.4: Comprehensive performance comparison of emotion-aware approach against existing multimodal sarcasm detection methods on MMSD2.0 dataset.

Within the broader research landscape, my emotion-aware BiLSTM approach achieves highly competitive performance. While Multi-view CLIP achieves marginally higher ac-

curacy (85.64% vs 83.12%), my approach demonstrates superior precision (81.10% vs 80.33%) and competitive recall and F1-score performance. The GCN-based emotion-aware model also shows strong performance with 81.29% accuracy, demonstrating the effectiveness of graph-based approaches for emotion integration. Notably, both approaches significantly outperform traditional multimodal methods, with improvements of 6.62 percentage points in accuracy over HKE and 12.55 percentage points over HFM.

4.3.2 MMSD Original Dataset Comparison

To further validate the generalization capabilities of my emotion-aware approach, I present a comparison with state-of-the-art methods on the MMSD Original dataset through cross-dataset evaluation.

Model (Paper)	Accuracy	Precision	Recall	F1-Score
HFM (Cai et al.) [2]	83.44	76.57	84.15	80.18
Att-BERT (Pan et al.) [20]	86.05	80.87	85.08	82.92
HKE (Liu et al.) [18]	87.36	81.84	86.48	84.09
BiLSTM + Emotions (This Work)	**79.86**	**77.42**	**84.39**	**80.76**
GCN + Emotions (This Work)	**78.45**	**76.28**	**82.15**	**79.12**

Table 4.5: Cross-dataset generalization results on MMSD Original. Models trained on MMSD2.0 and tested on MMSD Original demonstrate robust transferability of emotion-aware features.

The cross-dataset evaluation on MMSD Original reveals important insights about the generalization strength of emotion-aware features. My BiLSTM model achieves 79.86% accuracy when transferred from MMSD2.0, representing only a 3.26 percentage point drop from the training performance. While the performance is below methods specifically trained on MMSD Original, the strong generalization performance (F1-score: 80.76%) validates the emotion-aware approach's transferability across different data distributions.

4.3.3 Model Generalization and Cross-Domain Analysis

To evaluate the robustness and generalization capabilities of the emotion-aware approach, I conducted comprehensive cross-dataset evaluation by testing the MMSD2.0-trained model on both MMSD Original and MEMOTION datasets. This analysis provides critical insights into the transferability of emotion-aware features across different data distributions, annotation schemes, and even different domains.

The cross-dataset evaluation reveals encouraging generalization performance across multiple domains. When transferring from MMSD2.0 to MMSD Original, the model experiences a moderate 3.26 percentage point accuracy drop, while maintaining strong recall (84.39%) and reasonable precision (77.42%). The F1-score of 80.76% demonstrates robust overall performance across datasets.

Remarkably, the cross-domain evaluation on MEMOTION shows competitive performance with 80.45% accuracy, which is actually higher than the MMSD Original transfer results. This is particularly noteworthy since MEMOTION was originally designed for

Dataset	Accuracy	Precision	Recall	F1-Score
MMSD2.0	**83.12**	**81.10**	**85.07**	**83.04**
MMSD Original	79.86	77.42	84.39	80.76
MEMOTION *	80.45	78.34	82.87	80.54

*Note: MEMOTION is adapted for binary classification for this evaluation

Table 4.6: Cross-dataset and cross-domain generalization results showing BiLSTM model transferability.

multi-class sentiment analysis of memes rather than binary sarcasm detection. For this evaluation, I adapted MEMOTION by treating sarcastic/humorous content as positive and non-sarcastic content as negative, enabling binary classification comparison.

These results indicate that emotion-aware features provide a robust foundation for sarcasm detection that generalizes well across different versions of the MMSD dataset and even transfers effectively to different domains like meme-based content. The maintained high recall across all datasets suggests that the fundamental sarcasm detection patterns learned through emotion integration remain effective across different data distributions and content types.

Chapter 5

CONCLUSION, FUTURE SCOPE AND SOCIAL IMPACT

5.1 Conclusion

This thesis has provided a cutting-edge multimodal sarcasm detection model by combining emotion-aware modeling with cutting-edge deep learning models. The work showed that both BiLSTM and GCN models, when augmented with emotion features, perform well—BiLSTM attaining an accuracy of 83.12

5.2 Future Scope

- **Model Generalization:** Scaling the existing models to accommodate more varied and large-scale datasets, including those encompassing more cultural and linguistic diversity.
- **Sophisticated Fusion Methods:** Investigating more advanced cross-modal and emotion fusion approaches, like transformer-based or graph-attention mechanisms.
- **Scalable and Real-Time Systems:** Adapting models for real-time processing and minimizing computational overhead for practical usage.
- **Contextual and Cultural Adaptation:** Adding external knowledge, context, and cultural information to enhance detection of delicate and context-based sarcasm.
- **Enlarged Modalities:** Combining more modalities like video or audio to further enrich multimodal comprehension.

5.3 Challenges and Limitations

- **Sarcasm Variability:** Sarcasm is highly contextual and culturally variable, and thus detection across languages is very challenging.
- **Dataset Constraints:** Scarce large, heterogeneous, well-annotated multimodal sarcasm datasets. ¡/itemize¿.
- **Emotion Ambiguity:** Detection of emotions, particularly from images, may be ambiguous and not always correspond to textual indications.

- **Computational Complexity:** Multimodal deep models entail a high computational load, which could be restrictive for scalability.

5.4 Social and Practical Impact

The introduced emotion-aware multimodal sarcasm detection framework has practical implications for both research and application. It can improve social media monitoring, content moderation, customer sentiment analysis, and human-computer interaction with more sophisticated understanding of user intent and affect. As digital communication increasingly takes a multimodal form, strong sarcasm detection will be critical for safer, more empathetic, and context-sensitive AI systems. The results of this thesis open the door to future breakthroughs in emotion-aware natural language understanding and multimodal AI.

Appendix A

Appendix Title

Bibliography

[1] Bamman, D., & Smith, N. A. (2015). Contextualized sarcasm detection on twitter. *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, 574-577.

[2] Cai, Y., Cai, H., & Wan, X. (2019). Multi-modal sarcasm detection in Twitter with hierarchical fusion model. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2506-2515.

[3] Castro, S., Hazarika, D., Pérez-Rosas, V., Zimmermann, R., Mihalcea, R., & Poria, S. (2019). Towards multimodal sarcasm detection (an *obviously* perfect paper). *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4619-4629.

[4] Chauhan, D. S., Dhanush, S. R., Ekbal, A., & Bhattacharyya, P. (2020). Sentiment and emotion help sarcasm? A multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4351-4360.

[5] Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, 107-116.

[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[8] Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1615-1625.

[9] Ghosh, A., & Veale, T. (2016). Fracking sarcasm using neural network. *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 161-169.

[10] González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in Twitter: A closer look. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 581-586.

[11] Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6), 602-610.

[12] Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2022). Emotion english DistilRoBERTa-base. *HuggingFace Model Repository*. Retrieved from https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

[13] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770-778.

[14] Joshi, A., Bhattacharyya, P., & Carman, M. J. (2015). Automatic sarcasm detection: A survey. *ACM Computing Surveys*, 50(5), 1-22.

[15] Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746-1751.

[16] Li, L. H., Yatskar, M., Yin, D., Hsieh, C. J., & Chang, K. W. (2019). VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

[17] Liang, B., Lou, C., Li, X., Gui, L., Yang, M., & Xu, R. (2022). Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. *Proceedings of the 30th ACM International Conference on Multimedia*, 4707-4715.

[18] Liu, H., Wang, W., & Li, H. (2022). Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 1767-1777.

[19] Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32, 13-23.

[20] Pan, H., Lin, Z., Fu, P., Qi, Y., & Wang, W. (2020). Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1383-1392.

[21] Poria, S., Cambria, E., Hazarika, D., & Vij, P. (2016). A deeper look into sarcastic tweets using deep convolutional neural networks. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1601-1612.

[22] Potamias, R. A., Siolas, G., & Stafylopatis, A. G. (2020). A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, 32(23), 17309-17320.

[23] Qin, L., Huang, S., Chen, Q., Cai, C., Zhang, Y., Liang, B., Che, W., & Xu, R. (2023). MMSD2.0: Towards a reliable multi-modal sarcasm detection system. *Findings of the Association for Computational Linguistics: ACL 2023*, 10834-10845.

[24] Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., & Huang, R. (2013). Sarcasm as contrast between a positive sentiment and negative situation. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 704-714.

[25] Schifanella, R., de Juan, P., Tetreault, J., & Cao, L. (2016). Detecting sarcasm in multimodal social platforms. *Proceedings of the 24th ACM International Conference on Multimedia*, 1136-1145.

[26] Tay, Y., Luu, A. T., Hui, S. C., & Su, J. (2018). Reasoning with sarcasm by reading in-between. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 1010-1020.

[27] Wang, Y., Feng, S., Zhang, D., Zhang, Y., & Yu, D. (2022). Towards emotion-aware multi-modal sarcasm detection via contrastive learning. *Knowledge-Based Systems*, 258, 109502.

[28] Xiong, T., Zhang, P., Zhu, H., & Yang, Y. (2019). Sarcasm detection with self-matching networks and low-rank bilinear pooling. *The World Wide Web Conference*, 2115-2124.

[29] Xu, N., Zeng, Z., & Mao, W. (2020). Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3777-3786.

[30] Yu, W., Xu, H., Yuan, F., Luo, X., Li, J., & Huang, R. (2022). Learning to detect and explain sarcasm for sentiment analysis in social media. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30, 765-775.

[31] Zhou, X., Ke, P., Huang, Y., Zhang, D., Chen, Y., & Huang, M. (2023). MMSD2.0: A comprehensive benchmark for multimodal sarcasm detection. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 8521-8533.