MULTITASK LEARNING WITH BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS FOR SENTIMENT ANALYSIS AND SARCASM DETECTION

TAN YIK YANG

FACULTY OF ENGINEERING
UNIVERSITI MALAYA
KUALA LUMPUR

2024

MULTITASK LEARNING WITH BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS FOR SENTIMENT ANALYSIS AND SARCASM DETECTION

TAN YIK YANG

DISSERTATION SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ENGINEERING SCIENCE

FACULTY OF ENGINEERING
UNIVERSITI MALAYA
KUALA LUMPUR

2024

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Tan Yik Yang

Matric No: 17076420/02

Name of Degree: Master of Engineering Science

Title of Dissertation: Multitask Learning with Bidirectional Encoder Representations from Transformers for Sentiment Analysis and Sarcasm Detection

Field of Study: Computer/Data Network

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the Universiti Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date: 24/1/2024

Subscribed and solemnly declared before,

Witness's Signature Date: 24/1/2024

Name:

Designation

MULTITASK LEARNING WITH BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS FOR SENTIMENT ANALYSIS AND SARCASM DETECTION

ABSTRACT

In recent years, sentiment analysis has garnered significant interest in social media analytics, aiming to categorize people's thoughts, emotions, and feelings into positive, negative, or neutral categories. However, the increasing volume, complexity, and authenticity of social media data have introduced challenges such as misunderstanding, uncertainty, and inaccuracy. Particularly notable is the difficulty of identifying sarcasm in textual data, where negative intentions are expressed through positive sentences, presents a significant obstacle to sentiment analysis on social media platforms. This thesis proposes a novel multi-task learning framework that leverages Bidirectional Encoder Representations from Transformers (BERT), a state-of-the-art language model, to establish a correlation between sentiment analysis and sarcasm detection. The primary objective is to enhance the overall performance of sentiment analysis by identifying instances of sarcasm. The model's efficacy is demonstrated through comprehensive experiments, showing a notable improvement in F1-scores ranging from 2.5 to 6.5 percent upon incorporating sarcasm detection. The proposed approach not only enhances the sentiment classifier's performance but also significantly reduces training time and computational resources, offering substantial practical advantages. The findings underscore the importance of recognizing sarcasm in sentiment analysis and highlight how improved sentiment analysis aids in understanding sarcastic expressions in social media data. In the past, most sentiment analysis work treated the task as a standalone process. However, this thesis provides valuable insights into the influence of sarcasm on sentiment analysis, showing that accuracy can be improved in sentiment analysis by detecting sarcasm.

Keywords: Sentiment Analysis, Sarcasm Detection, Deep Learning, Transformers, Bidirectional Encoder Representations from Transformers (BERT), Multitask Learning

MULTITASK LEARNING WITH BIDIRECTIONAL ENCODER REPRESENTATIONS FROM TRANSFORMERS FOR SENTIMENT ANALYSIS AND SARCASM DETECTION

ABSTRAK

Sejak kebelakangan ini, analisis sentimen telah menarik minat yang ketara dalam analisis media sosial, bertujuan untuk mengategorikan pemikiran, emosi, dan perasaan individu ke dalam kategori positif, negatif, atau neutral. Namun, peningkatan jumlah, kerumitan, dan keaslian data media sosial telah memperkenalkan cabaran seperti salah tafsir, ketidakpastian, dan ketidaktepatan. Khususnya, mengenali sindiran dalam data teks, di mana niat negatif dinyatakan melalui ayat positif, menjadi halangan utama kepada analisis sentimen di platform media sosial. Tesis ini mengemukakan satu rangka kerja pembelajaran pelbagai tugas yang baru, yang menggunakan Bidirectional Encoder Representations from Transformers (BERT), satu model bahasa yang terkini, untuk menjalin korelasi antara analisis sentimen dan pengesanan sindiran. Matlamat utama adalah untuk meningkatkan prestasi keseluruhan analisis sentimen dengan mengenal pasti contoh sindiran. Keberkesanan model ini ditunjukkan melalui eksperimen komprehensif, yang menunjukkan peningkatan yang ketara dalam skor F1 yang berkisar antara 2.5 hingga 6.5 peratus dengan pengesanan sindiran diintegrasikan. Pendekatan yang dicadangkan tidak hanya meningkatkan prestasi pengklasifikasi sentimen, tetapi juga secara signifikan mengurangkan masa latihan dan sumber komputasi, menawarkan kelebihan praktikal. Penemuan ini menekankan kepentingan mengenali sindiran dalam analisis sentimen dan menyoroti bagaimana analisis sentimen yang lebih baik membantu memahami ungkapan sindiran dalam data media sosial. Sebelum ini, kebanyakan kerja analisis sentimen menganggap tugas tersebut sebagai proses berdiri sendiri. Namun, tesis ini memberikan pandangan berharga tentang impak sindiran terhadap analisis sentimen,

yang menunjukkan bahawa ketepatan boleh ditingkatkan dalam analisis sentimen dengan mengesan sindiran.

Keywords: Analisis Sentimen, Pengesanan sindiran, Deep Learning, Transformers, Bidirectional Encoder Representations from Transformers (BERT), Multitask Learning

ACKNOWLEDGEMENTS

I would like to begin by expressing my heartfelt gratitude to my supervisor, ASSOCIATE PROF. IR. DR. CHOW CHEE ONN, for his invaluable support throughout the entire research journey. His expertise and guidance have been instrumental in pushing me to achieve higher levels of excellence and ensuring that my progress remained on track. I am grateful for his attentive listening and willingness to address my concerns, providing me with valuable advice that greatly contributed to the successful completion of this project.

Additionally, I extend my sincere appreciation to ASSOCIATE PROF. IR. DR. CHUAH JOON HUANG for his continuous support as my co-supervisor. His contributions have been crucial in enriching the research process and fostering a conducive learning environment.

I am truly fortunate to have such dedicated and experienced supervisors who have played a significant role in shaping the outcomes of this research. Their encouragement and mentorship have been invaluable assets, and I am grateful for the opportunity to work with them.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

SA	:	Sentiment Analysis
NLP	:	Natural Language Processing
BERT	:	Bidirectional Encoder Representations from Transformers
MTL	:	Multitask Learning
LLMs	:	Large Language Models
DNN	:	Deep Neural Network
GPT	:	Generative Pre-trained Transformers
Bi-LSTM	:	Bidirectional Long Short-Term Memory

CHAPTER 1: INTRODUCTION

1.1 Background

The rapid growth of the internet, a phenomenon that has reshaped the modern world, has led to an unprecedented number of active users worldwide. As of April 2023, the global digital landscape boasts a staggering 5.18 billion individuals connected, encompassing an impressive 64% of the entire global population (Statista, n.d.). This transformative surge in connectivity has given rise to a unique social ecosystem, where virtual interactions have become as integral as physical interactions. Within this realm, social media platforms such as Twitter, Reddit, and Facebook have become an integral part of modern life. They offer a space for individuals to share not only personal opinions and emotions but also significant social and business events, fostering a real-time exchange of information. Consequently, this interconnectedness has triggered ripple effects across various dimensions, notably influencing social dynamics, political discourse, and economic interactions on a global scale.

Businesses, recognizing the potential of social media, have embraced it as a direct means of engaging with consumers, gaining insights into their preferences, and advertising products and services. Consumers, on the other hand, wield significant power in shaping a company's success or failure through their reviews and responses. Studies have indicated that a staggering 93% of internet users are influenced by customer reviews when making purchasing decisions (Podium, 2017). Therefore, companies that can swiftly respond to consumer feedback and adapt their strategies gain a competitive advantage.

Furthermore, social media's influence has extended to unprecedented levels during critical global events like the COVID-19 pandemic, which has caused immense fear, stress, and disruptions worldwide. Social media has become a vital platform for

individuals to express their emotions and share information, offering a unique opportunity to analyze sentiments and emotions directly from users' newsfeeds.

To extract meaningful insights from the vast amount of unstructured data generated on social media, sentiment analysis (SA) has emerged as a critical process. SA involves identifying and classifying subjective information in text using computational linguistic techniques within the realm of Natural Language Processing (NLP) (Zhao, Liu, & Xu, 2016). It allows us to determine the polarity of sentences and plays a significant role in various fields, such as product reviews, stock market forecasts, and responses to major events like terrorist attacks (Dave, Lawrence, & Pennock, 2003; Nassirtoussi, Aghabozorgi, Wah, & Ngo, 2014; Burnap, et al., 2014). Notably, SA has proven invaluable in understanding people's feelings during the COVID-19 pandemic, aiding governments in formulating appropriate measures (Arunachalam & Sarkar, 2013).

1.2 Research Problem

Automated sentiment analysis holds tremendous potential for extracting valuable insights across various domains. However, the intricate nature of human language and the subjective content found on social platforms pose significant challenges. A notable hurdle in this landscape is the accurate detection of sarcasm— the use of remarks that clearly mean the opposite of what they say, made in order to hurt someone's feelings or to criticize something in a humorous way ("Sarcasm", n.d.). The use of mockery in sarcastic remarks introduces a level of complexity that can confound automated sentiment analysis systems.

The technical challenges and problems inherent in this research involve overcoming the nuanced and context-dependent nature of sarcasm. Unlike straightforward sentiment expressions, sarcasm often employs negative language to convey positive sentiments, making it challenging for conventional sentiment analysis models. Moreover, sarcasm relies heavily on subtle cues such as tone and context which are elements that automated

systems may struggle to interpret without explicit programming to account for these complexities.

The significance of this research lies in its potential to advance the capabilities of automated sentiment analysis systems, making them more robust and adaptable in real-world applications where sarcasm is prevalent. By leveraging multitask learning with BERT, the proposed model seeks to capture the intricacies of both sentiment and sarcasm, addressing a critical gap in current automated analysis systems. The outcomes of this research could have far-reaching implications, extending from the enhancement of sentiment analysis systems to enabling Artificial Intelligence to understand nuances in languages, particularly in the domain of sarcasm.

1.3 Motivation

The rapid growth of the internet and the widespread adoption of social media platforms have transformed the way individuals communicate, share information, and express emotions. With billions of active users engaging in online interactions, the potential impact of these platforms on society, business, and global events is undeniable. To harness the extensive information generated on these platforms, SA has emerged as a vital process within the realm of Natural Language Processing (NLP) and becomes a driving force, pushing industries and researchers to delve into sentiment analysis using various text pre-processing methods and machine learning techniques.

However, challenges arise when intense emotions are expressed. Individuals often convey messages with the use of sarcasm. Since sarcasm can alter how sentiment analysis systems interpret the sentiment behind the message, promptly identifying and addressing it is vital in automated sentiment analysis systems.

Despite the challenges encountered in sentiment analysis, significant advancements have been made in Natural Language Processing (NLP), especially with the rise of the Transformer model. This model serves as the backbone for Large Language Models (LLMs) like BERT and GPT. These advancements have shown promising results in addressing the limitations previously encountered. The attention mechanism of the Transformer model enables it to capture context more effectively, while LLMs can leverage vast amounts of data to improve the accuracy of sentiment analysis models. This progress has opened new possibilities for understanding and interpreting human emotions, sentiments, and even detecting instances of sarcasm in the digital age. As a result, sentiment analysis has become more robust and capable of handling complex language patterns and nuances, paving the way for more sophisticated and insightful applications in various domains.

1.4 Research Aim and Objectives

In this research, the primary aim is to harness the capabilities of transformers-based Large Language Models (LLMs) to enhance the accuracy of sentiment analysis and finding the correlation between sarcasm detection. This is achieved by developing an intelligent information extraction system capable of precisely classifying sentiments with underlying sarcasm context within social media data. Specifically, various LLM algorithms are studied and compared against the conventional deep learning methods, including Bi-directional Long Short-Term Memory (Bi-LSTM).

Additionally, we intend to evaluate the impact of integrating sarcasm detection into sentiment analysis frameworks. Through meticulous experiments and analysis, we seek to evaluate and uncover the intricate relationship between sarcasm and sentiment, and to determine the extent to which accurate sentiment interpretation can be achieved.

Lastly, we advocate for the use of deep multi-task learning and present a novel framework that simultaneously trains classifiers for sentiment and sarcasm. This approach involves training two models simultaneously—one for sentiment analysis and the other for sarcasm detection. By sharing the underlying representations learned by the LLMs between these two tasks, the goal is to reduce model complexity and enhance overall efficiency, leading to improved performance in both sentiment analysis and sarcasm detection.

The objectives of this research are as follows:

I. To assess the significance and impact of sarcasm detection on sentiment analysis by quantifying its impact on overall model effectiveness.

II. To propose a framework using BERT-based multi-task learning to simultaneously train sentiment and sarcasm classifiers, reducing training time, computation power, and enhancing model performance.

1.5 Organization of Thesis

This thesis comprises five distinct chapters, each serving a specific purpose in effectively introducing and presenting the research details and findings.

Chapter 2 is dedicated to an extensive literature review, where existing works related to sentiment analysis and sarcasm detection ranging from traditional methods to state-of-the-arts are studied. This chapter serves as a foundation for the research, showcasing the current state of the field and identifying research gaps and shortcomings in previous approaches. From these insights, we can leverage the shortcomings and formulate an improved approach as the proposed method.

Chapter 3 delves into the methodology behind the proposed method. Here, the paper introduces a novel multi-task learning framework devised for sentiment analysis and

sarcasm detection. The section delves into the fundamental theoretical principles, pre-processing methods, and the model selection criteria for the proposed method. By explaining the architecture and approach, this section provides the reader with a comprehensive understanding of the proposed solution.

Chapter 4 provides in-depth evaluation of the proposed method based on various carefully devised experiments. By doing so, the experiments aim to evaluate the proposed method's effectiveness against other benchmarked state-of-art works and identify the shortcomings of the proposed method.

Chapter 5 recaps the research objectives and summarizes the key findings, highlighting the contributions made to the field of sentiment analysis and sarcasm detection. The chapter also reflects on the limitations of the proposed method and suggests possible future research.

CHAPTER 2: LITERATURE REVIEW

In the realm of natural language processing, algorithms primarily relied on intricate sets of hand-written rules until the 1980s. However, a transformative shift in NLP unfolded with the advent of machine learning algorithms. Chapter 2 serves the purpose of delving into the literature to identify gaps and justify their novelties. The discussion within this section navigates through various sentiment classification methods, commencing with early works employing conventional machine learning and deep learning techniques. It progresses to explore the evolution of approaches, encompassing the current state-of-the-art models such as the transformer model and multi-task learning. Additionally, the literature review will explore research that has enhanced sentiment classification through sarcasm detection, which shares similarities with our approach.

2.1 Sentiment Analysis using Machine Learning Method

In the early stages of sentiment analysis research, the focus centered on categorizing sentiments as positive or negative, as exemplified in (Pang & Lee, 2004). In this research, three machine learning algorithms were utilized for sentiment classification: Support Vector Machine (SVM), Naïve Bayes classifier, and Maximum Entropy. The classification process used the n-gram method, encompassing unigram, bigram, and a combination of both. Additionally, the bag-of-words (BOW) paradigm was also introduced to facilitate machine learning algorithms. The outcomes of their experiments displayed promising performance, indicating substantial potential in this approach.

Another study explored document-level sentiment analysis by utilizing the syntactic relation between words (Matsumoto, Takamura, & Okumura, 2005). They derived sub-sequences of frequent terms and sub-trees of dependence from sentences, which served as features for the SVM algorithm. Additionally, they extracted unigram, bigram, word subsequence, and dependence from each sentence in the dataset. Similarly, a combination

of unsupervised and supervised techniques was employed in (Maas, et al., 2011) to learn word vectors and capture semantic term (document information) and rich sentiment contents.

In another work by Bespalov et al. (Bespalov, Bai, Qi, & Shokoufandeh, 2011), a mechanism that embeds higher-order n-gram phrases within a low-order dimensional semantic latent space were proposed to define a sentiment classification function. They employed SVM to construct a discriminative system that estimates latent space parameters with a bias towards the classification task. This method can handle both binary classifications and multi-score sentiment classifications, where prediction involves a set of sentiment scores.

Another approach involved a sentiment classification method using an entropy-weighted genetic algorithm (EWGA) and SVM (Abbasi, Chen, & Salem, 2008). Various sets of features, consisting of syntactic and stylistic characteristics, were evaluated. Stylistic aspects encompassed measures of word length distribution, vocabulary richness, and frequency of special characters. Weights were allocated for different sentiment attributes before the genetic algorithm was employed to optimize sentiment classification. The model was validated using SVM with a ten-fold cross-validation technique, yielding promising results.

Table 2.1: Table of Comparison for Machine Learning Method

Work	Advantages	Disadvantages
Pang, B., & Lee, L. (2004)	Employed n-gram method and Bag-of-Words paradigm.	Lack of exploration into more nuanced SA tasks beyong binary categories
Matsumoto, Y., Takamura, H., & Okumura, M. (2005)	Explored document-level by using syntactic relations between words.	Feature extraction complexity may hinder adaptability to diverse datasets
Bespalov, I., Bai, Y., Qi, G. J., & Shokoufandeh, A. (2011)	Proposed a mechanism embedding higher-order n-gram phrases in a low-order dimensional semantic latent space.	Limited explanation of the latent space parameters and their impact on classification.
Abbasi, A., Chen, H., & Salem, A. (2008)	Successfully optimized sentiment classification through genetic algorithm and SVM validation.	Limited exploration of the impact of stylistic features on sentiment classification.

2.2 Sentiment Analysis using Deep Learning Method

Recently, deep learning has garnered significant attention due to its ability to eliminate the need for traditional, task-specific feature engineering, making it a potent alternative for sentiment analysis.

Yanagimoto et al. proposed, a novel architecture utilizing a deep neural network to determine document similarity (Yanagimoto, Shimada, & Yoshimura, 2013). The model was trained to generate vector representations for articles by leveraging multiple market news sources obtained from T&C. The cosine similarity was then calculated among labelled papers, considering their polarity while disregarding the contents. The proposed method exhibited outstanding performance in estimating article similarities.

In another study conducted by Yafoz et al. (Yafoz & Mouhoub, 2021), sentiments in datasets containing reviews of cars and real estate in Arabic online platforms were examined. They explored the effectiveness of various deep learning algorithms, namely Bi-LSTM (Bidirectional Long Short-Term Memory), LSTM (Long Short-Term Memory), GRU (Gated Recurrent Unit), CNN (Convolutional Neural Networks), and CNN-GRU, in conjunction with the BERT word embedding model. Among the combinations tested, utilizing the BERT model with LSTM resulted in the highest F1 score of 98.71% for the vehicle dataset. Conversely, for the real estate dataset, the maximum F1 score of 98.67% was achieved using the BERT model with CNN.

Table 2.2: Table of Comparison for Deep Learning Method

Work	Advantages	Disadvantages
Yanagimoto et al. (2013)	Proposed a novel framework using consine similarity and deep learning for SA task and demonstrated outstanding performance in estimating article similarities.	Cosine similarity calculation may oversimplify document similarity and does not content.
Yafoz et al. (2021)	Explored the effectiveness of various deep learning algorithms such as DNN, CNN and RNN and Achieved high F1 scores, with BERT model and LSTM outperforming others.	Does not consider sarcasm in SA, where the proposed model may misclassify sarcasm context in dataset.

2.3 Sentiment Analysis using Transformer-based Large Language Models (LLMs)

In recent times, researchers have extensively explored transformer-based models, particularly the BERT model, for various NLP tasks, including sentiment analysis. Noteworthy publications in this domain highlight the advancements made in Sentiment Analysis through pre-trained BERT models.

One significant contribution was the introduction of BERTweet, a large-scale, open-source, pre-trained BERT model specifically designed for English Tweets by Nguyuen et al. (Nguyen, Quoc, Vu, & Nguyen, 2020). BERTweet not only exceled in text classification but also showcased impressive performance in named entity recognition and parts of speech (POS) tagging. Experimental results indicated its superiority over BERT-base architecture-based models like XLM-Rbase and RoBERTbase.

Another study by Phan et al. (H. V Phan & Do, 2020) presented a question-answering (QA) system based on LSTM, multilingual BERT, and BERT+vnKG models, utilizing a Vietnam tourism QA dataset for comparison. The proposed model exhibited superior accuracy and speed compared to the other models tested.

For SA on Spanish-language Twitter data, researchers compared RNN+LSTM and BERT models to identify cyberbullying (A Andrade-Segarra1 & A. Leon-Paredes, 2021). The findings revealed that the BERT model outperformed RNN+LSTM by a significant margin of 20% in terms of accuracy and performance. BERT-base multilingual-uncased and BERT-large-uncased models showed higher accuracy as well, but their complex architecture poses challenges in terms of computational resources and infrastructure investment.

Various other experimental research, case studies, and review papers have been published on BERT-based sentiment analysis for Italian Twitter SA (Pota, Ventura, Catelli, & Esposito, 2020), Arabic aspect-based (Abdelgwad, Soliman, & Taloba, 2022), and Bangla-English Machine Translation (Akhand, Roy, Dhar, & Kamal, 2021), further showcasing the versatility and applicability of BERT models.

In a research paper from Geetha et al. (Geetha & Renuka, 2021), a comparison between SA models, machine learning models like Naive Bayes (NB) and Support Vector Machine (SVM), LSTM, and BERT-based uncased models was conducted on customer reviews and rating datasets from e-commerce platforms. The deep learning BERT uncased model outperformed other machine learning models in terms of performance and sentiment prediction accuracy.

Additionally, Xu et al. (Xu, Shu, Yu, & Liu, 2020) investigated the hidden representations of pre-trained BERT models for aspect-based sentiment analysis (ABSA) on reviews datasets. Their findings revealed that BERT effectively utilizes aspect representation and self-attention head encoding to encode word context, thereby enhancing aspect-based sentiment analysis.

Another study by Pang et al. (Pang, et al., 2021) explored aspect-based sentiment analysis (ABSA) using the ALM-BERT model on consumer datasets, demonstrating the superior performance of the ALM-BERT model compared to other models.

For aspect-based SA, Karima et al. (Karimi, Rossi, & Prati, 2020) employed parallel and hierarchical aggregation approaches based on the hierarchical transformer model, developing two BERT models for aspect extraction and sentiment classification. Their proposed model strategy improved performance significantly, eliminating the need for additional model training.

In research by Batra et al. (Batra, Punn, Sonbhadra, & Agarwal, 2021), BERT-based SA models were explored using datasets from Software Engineering sources like GitHub, Jira web portal comments, and Stack Overflow postings. The ensemble and compressed BERT models outperformed other models by 6-12% in terms of F1 score measurements, validating their efficacy in sentiment analysis tasks.

In conclusion, recent advancements in sentiment analysis have prominently featured the widespread adoption of transformer-based models, particularly the BERT model, across diverse natural language processing (NLP) tasks. Notable contributions include BERTweet, a specialized BERT model tailored for English Tweets, showcasing superior performance in text classification, named entity recognition, and parts of speech tagging. Similarly, studies like the Vietnam tourism question-answering system and the exploration of sentiment analysis on Spanish-language Twitter data demonstrated the efficacy of deep learning models, such as LSTM and BERT, in achieving heightened accuracy and efficiency. The versatility of BERT models is further highlighted in various experiments, case studies, and review papers spanning Italian Twitter sentiment analysis, Arabic aspect-based sentiment analysis, Bangla-English machine translation, and more. These endeavours consistently reported state-of-the-art results, with BERT models outperforming other architectures in terms of accuracy and predictive capabilities. However, it is crucial to note that despite these achievements, the addressed works do not explicitly delve into the intricacies of handling sarcasm in sentiment analysis.

2.4 Sentiment Analysis using Multitask Learning

Recently, multi-task learning has gained substantial attention in deep learning research. Multi-task learning allows multiple tasks to be performed simultaneously by a shared model. Yousif et al. (Yousif, Niu, Chambua, & Younas, 2019) proposed a multi-task learning approach based on CNN and RNN. This model jointly learns the citation

sentiment classification (CSC) and citation purpose classification (CPC) to boost the overall performance of automated citation analysis and simultaneously ease the problem of inadequate training data and time-consuming for feature engineering. While the discussed models showcase remarkable capabilities, the incorporation of sarcasm detection mechanisms would further enhance their applicability and broaden the scope of sentiment analysis in real-world scenarios.

2.5 Sentiment Analysis with Sarcasm Detection

Thus far, not much research has been conducted on sarcasm detection. Yunitasari et al., proposed a method to enhance standalone sentiment classifiers using sarcasm detection (Yunitasari, Musdholifah, & Sari, 2019). This method involved training two explicit models: one for sentiment classification and the other for sarcasm detection. For sarcasm detection, the method employed top word features, unigram, and 4 Boaziz features, which encompassed punctuation-related, sentiment-related, lexical, and syntactic features. The sarcasm detection utilized the Random Forest algorithm, while the sentiment classification was performed using the Naïve Bayes algorithm. The results of the model evaluation were encouraging, with an accuracy of 80.4%, a recall of 91.3%, and a precision of 83.2%. Importantly, the evaluation demonstrated that integrating sarcasm detection into sentiment analysis led to an improvement of approximately 5.49% in performance. This finding highlights the potential impact of sarcasm detection in enhancing the accuracy and reliability of sentiment analysis models.

In the exploration of sarcasm detection and sentiment analysis, Mahdaouy et al. study focuses on employing a deep multi-task model named MARBERT (Mahdaouy, et al., 2021). Trained on the ArSarcasm dataset, which involves a shared task between sentiment analysis and sarcasm, the model utilizes a BERT encoder, multitask attention mechanism, and two task classifiers, mirroring a similar framework to previous works. The

experimental evaluation is conducted on the ArSarcasm dataset comprising 3000 instances for both test and development sets. Despite not achieving the highest score compared to other models, the proposed MARBERT model demonstrates notable performance in classifying sarcastic content as negative in sentiment analysis, implying a significant impact of sarcasm on sentiment.

2.6 Research Gap

This literature review introduces different techniques used by researchers in the field of Sentiment Analysis, including N-gram, Hybrid Multi-task Learning (MLT), Deep learning methods, and Transformer-based models. Researchers commonly rely on finding suitable datasets, performing data pre-processing, and converting data into numerical vector form to achieve better results. The accuracy obtained from these methods is notably high; for instance, the N-gram method achieved an accuracy of 94.6% using SVM (Bespalov, Bai, Qi, & Shokoufandeh, 2011), and the hybrid MLTs method achieved 91.7% using a combination of EWGA and SVM (Abbasi, Chen, & Salem, 2008). However, with the rise of transformer based LLMs models, it is evident that many of the LLMs employed in these approaches outperform conventional methods. Nevertheless,

Table 2.3: Table of Comparison for Sentiment Analysis with Sarcasm Detection

Work	Advantages	Disadvantages
Yunitasari et al. (2019)	Improved sentiment analysis accuracy as compared to standalone classfiers. Showcased the importance of saracsm in SA task.	The model explicit trained two classifers sentiment and sarcasm classifiers. This contribute to longer training time and more computation required.
Mahdaouy et al. (2021)	Implemented state-of-art models for SA and sarcasm detection such as BERT and multi-task learning. Also, successfully showcased the importance of sarcasm in SA task.	Language is in Arabic and the proposed method does not achieve the best score compared to other models.

certain shortcomings are observed in these techniques, particularly in handling sarcasm contexts, which significantly impacts sentiment classification. Failure to account for sarcasm in the system often results in misclassification, with sarcastic text being incorrectly labelled as positive sentiment.

The work proposed by Yunitasari et al. (Yunitasari, Musdholifah, & Sari, 2019) has successfully addressed this issue, but also introduced higher complexity, longer processing time, and a potential risk of overfitting. Furthermore, the work proposed by Mahdaouy et al. showcased promissing result in improving sentiment analysis using sarcasm detection in Arabic langauge.

With this in mind, it is desired to design a framework that can maintain the computational complexity while increasing the accuracy of sentiment analysis by incorporating sarcasm detection. A summary of the existing work for benchmarking has been summarized in Table 2.4.

Table 2.4: An Overview of State-of-the-Art Models for Sentiment Analysis

Work	Model Name	Details of the Model	Dataset
(Zhao, et al., 2018)	RNN-Capsule	RNN-based capsule model	Reviews
(Rezaeinia, Rahmani, Ghodsi, & Veisi, 2019)	IWV	Improved word vectors with CNN	Reviews
(Chen, Xu, He, & Wang, 2017)	BiLSTM-CRF	BiLSTM-CRF with 1-D CNN	Reviews
(Huang, Jin, & Rao, 2020)	SAT	BERT-based with two-stage training strategy	Reviews
(Lei, Yang, & Yang, 2018)	SAAN	Sentiment -aware multi-head attention with CNN model	Reviews
(Liu & Guo, 2019)	AC-BiLSTM	Bi-LSTM model with attention mechanism and convolution layer	Reviews
(Li, Qi, Tang, & Yu, 2020)	SAMF-BiLSTM	Bi-LSTM with self-attention mechanism and multi-channel features	Reviews

(Usama, et al., 2020)	ATTPolling	RNN with CNN-based attention	Reviews
(Zhang, Wang, & Zhang, 2021)	MVA	Multiview attention model to learn sentence representations from multiple angles	Reviews, Tweets, Questions, News
(Naderalvojoud & Sezer, 2020)	SAWE	Sentiment-aware word embeddings using refinement with senti-contextualized learning method	Reviews, Tweets
(Basiri, Nemati, Abdar, Cambria, & Acharya, 2021)	ABCDM	Bidirectional layer with LSTM and GRU with CNN and pooling layer	Reviews
(Onan, 2022)	RCNNGWE	Bi-Conv-RNN with group-wise enhancement mechanism	Reviews, Tweets
(Wang, Zhang, Yu, & Zhang, 2022)	CoSE	Contextual sentiment embeddings with 2-layer GRU	Reviews, Tweets
(Khan, Ahmad, Khalid, Ali, & Lee, 2023)	SCA-HDNN	Sentiment knowledge embeddings from BERT with Bi-LSTM, attention mechanism and CNN layer	Review, Tweets, Question, News

CHAPTER 3: PROPOSED METHOD

Sarcasm is a pervasive phenomenon on social media platforms, presenting a significant challenge for sentiment analysis. Conventional sentiment analysis models encounter difficulties in detecting sarcasm due to the intricate language used in sarcastic expressions. Sarcasm involves conveying negative intentions through positive sentences, posing a challenge for models to grasp without adequate context.

To tackle this issue, researchers have turned to transformer based LLMs such as BERT and GPT, which have exhibited exceptional performance in identifying sarcasm and understanding nuanced language. Their ability to contextualize words within the broader context of a sentence or document enables them to effectively handle complex language structures.

This research proposed a method that leverages BERT in a multi-task learning framework to enhance sentiment analysis by proficiently identifying sarcasm. By simultaneously training the model to predict both sentiment and sarcasm, the approach aims to capture the underlying correlation between these two tasks, thereby improving the overall performance of sentiment analysis.

The utilization of BERT in a multi-task learning framework is expected to yield superior results compared to traditional models like Bi-LSTM, as it equips the model to handle the intricacies of language and nuances more effectively. By identifying sarcasm more accurately, the proposed method seeks to enhance the accuracy and reliability of sentiment analysis on social media platforms.

The schematic representation of the proposed framework's overall architecture is depicted in Figure 3.1.

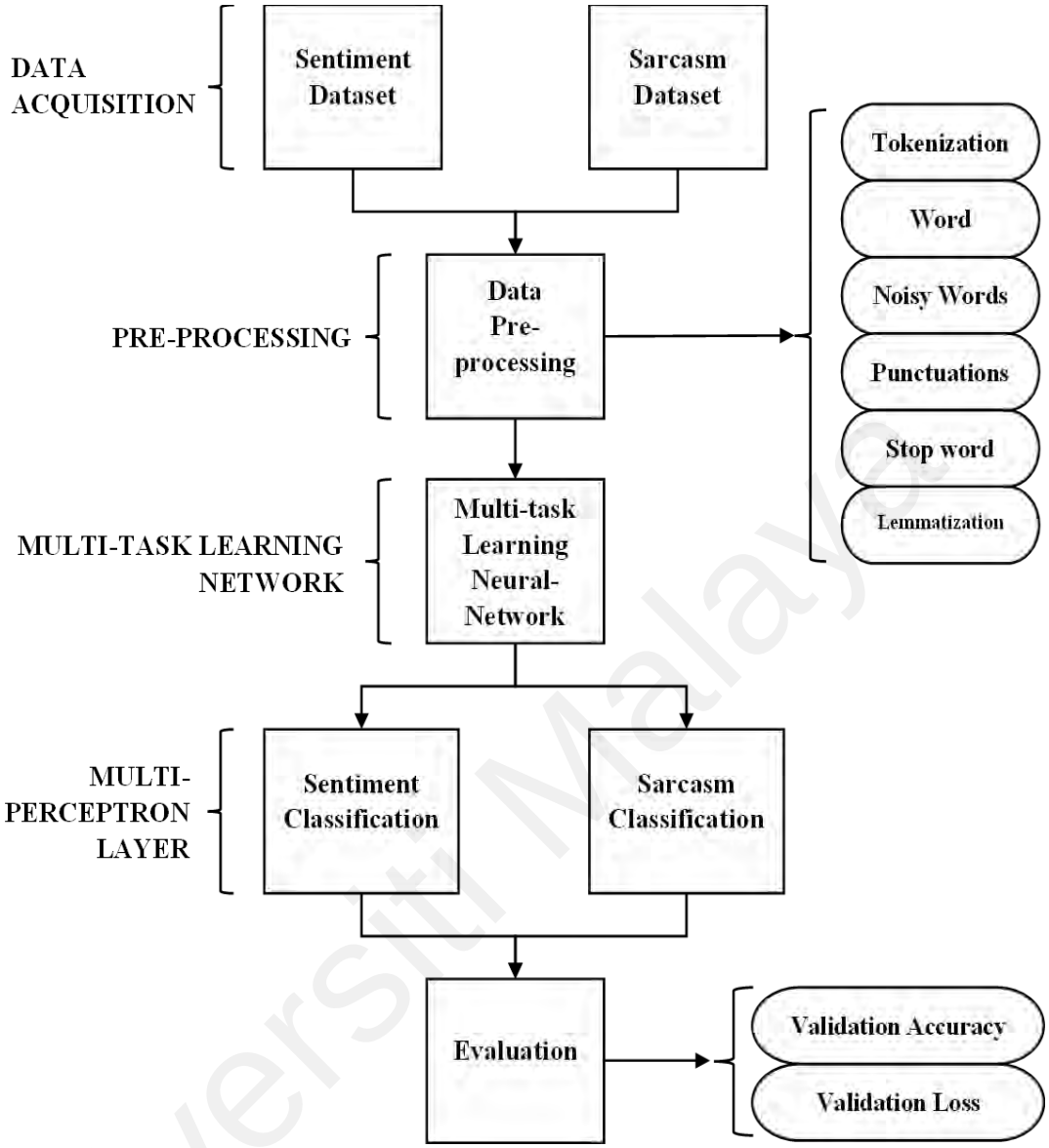


Figure 3.1: Sentiment Analysis and Sarcasm Detection using Deep Multitask Learning: Overall Framework

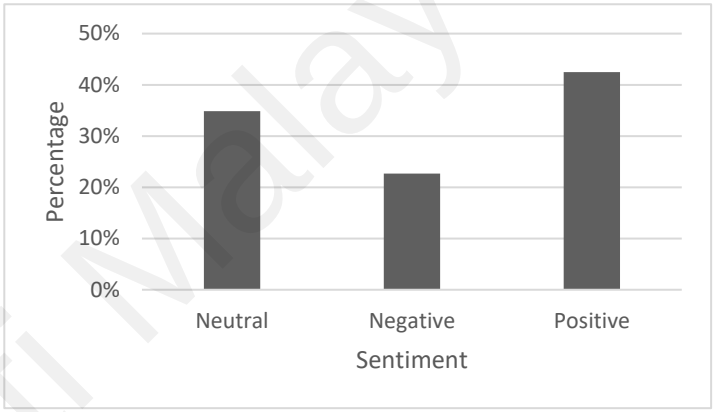
3.1 Data Acquisition

Data acquisition represents the initial stage in machine and deep learning, often demanding considerable time and resources to gather potentially relevant data. In the proposed framework, two datasets are required: a sentiment dataset and a sarcasm dataset. The sentiment dataset consists of 227,600 tweets, categorized as 0 for neutral, 1 for negative sentiment, and 2 for positive sentiment (Chaithanya, 2020). Figure 3.2(a) displays sample headers from the dataset, while the distribution is depicted in Figure

3.2(b). Notably, the dataset exhibits a slight imbalance, with the majority being labeled as neutral and the minority as negative sentiment.

As for the sarcasm dataset, it comprises 50,722 tweets labeled with 0 for non-sarcastic and 1 for sarcastic (Misra, 2019). Figure 3.3(a) illustrates sample headers from tweets categorized as sarcastic and non-sarcastic, and Figure 3.3(b) depicts the distribution, indicating a similar slight imbalance, with more tweets classified as non-sarcastic (type 0).

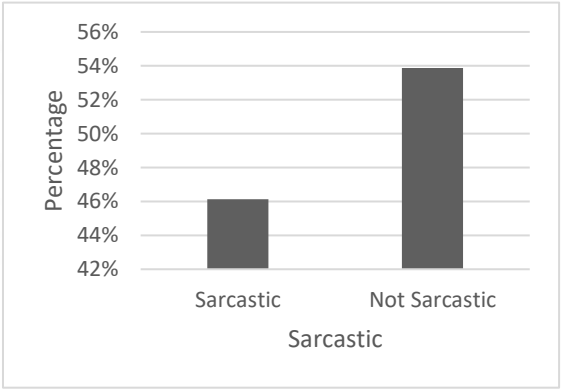
	clean_comment	label
0	I`d have responded, if I were going	neutral
1	Wow... u just became cooler.	positive
2	Sooo SAD	negative
3	bullying me	negative
4	leave me alone	negative



(a) (b)

Figure 3.2: (a) Sentiment Dataset Header; (b) Data Distribution in Sentiment Dataset According to Labels

	is_sarcastic	headline
0	1	thirtysomething scientists unveil doomsday clo...
1	0	dem rep. totally nails why congress is falling...
2	0	eat your veggies: 9 deliciously different recipes
3	1	inclement weather prevents liar from getting t...
4	1	mother comes pretty close to using word 'strea...



(a) (b)

Figure 3.3: (a) Sarcasm Dataset Header; (b) Data Distribution in Sarcasm Dataset According to Labels

3.2 Data Pre-Processing

Most of the datasets available are noisy and unstructured in nature, so pre-processing is needed to transform the noisy datasets into an understandable format for the training to ensure high accuracy. Noisy words in the context of sentiment analysis refer to terms that do not contribute significant meaning to the sentiment expressed in a sentence. This includes words like hyperlinks, retweets, stock market tickers, and common stop words. Removing noisy words not only reduces the dimensionality of the dataset but also helps in focusing on the essential content. Moreover, it is crucial to retain words that are semantically relevant to sentiment. Sentiment analysis relies heavily on the emotional context conveyed by words. For instance, the word "happy" conveys positive sentiment, while "unhappy" conveys negative sentiment. Removing irrelevant words ensures that the sentiment analysis model focuses on these crucial terms, improving its ability to accurately predict sentiment.

With the consideration in mind, the proposed method employs the pre-processing steps as listed below:

i. Removal of irrelevant words such as hyperlinks and noisy words (such as retweet and stock markets tickers and stop words).
ii. Removal of punctuation.
iii. Word lemmatization to reduce inflected to tier word stem or base.
iv. Tokenization to convert text to vector representation.

3.3 Model Selection

In the realm of NLP, particularly in sentiment analysis and sarcasm detection, the choice of an appropriate model plays a pivotal role in attaining optimal performance for a given task. Notably, the Bi-LSTM, along with Transformer-based models like BERT and GPT, have gained significant popularity in recent years.

In this section, a careful examination of these models is conducted to identify the most suitable candidate for incorporation into our framework.

3.3.1 Bidirectional Long Short-Term Memory Network (Bi-LSTM)

Deep recurrent neural networks (RNNs) have demonstrated remarkable effectiveness in sentiment analysis due to their ability to maintain information as memory over time through feedback loops in the recurrent layer. However, traditional RNNs encounter challenges in learning long-term temporal relationships due to the vanishing gradient problem, where the gradient of the loss function diminishes exponentially with time.

To address this issue, the LSTM network has been introduced. The LSTM network, which is a type of RNN, employs memory cells and three distinct gates—the forget gate, input gate, and output gate—to mitigate the vanishing gradient problem. These explicit gating mechanisms allow the cell to decide whether to read from, write to, or erase the state vector at each step. Figure 3.4 illustrates the fundamental topology of the LSTM network. The ability of LSTM to "memorize" and "forget" information at different stages makes it a practical solution for sentiment and sarcasm detection, as each word's significance in a sentence can be effectively preserved. Details of the equation of forget, input and gate are as follows:

i. Forget Gate: The forget gate determines what information from the cell state should be discarded or kept. It takes the previous cell state (C_{t-1}) and the current input (X_t) and produces a forget gate activation vector (f_t) between 0 and 1. Where: W_f is the weight matrix of forget gate, $[h_{t-1}, x_t]$ is the concatenation of the previous hidden state and the current input., b_f is the bias of forget gate and σ is the sigmoid activation function.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

ii. Input Gate: The input gate updates the cell state with new information. It also uses the previous cell state (C_{t-1}) and the current input (X_t) to compute an input gate activation vector (i_t) and a candidate cell state update (C_t). Where: W is the weight matrix of forget gate, $[h_{t-1}, x_t]$ is the concatenation of the previous hidden state and the current input., b is the bias of forget gate and σ is the sigmoid activation function.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$C_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

iii. Output Gate: The output gate determines the next hidden state (h_t) based on the updated cell state. It takes the previous hidden state (h_{t-1}), the current input (X_t), and the updated cell state (C_t) to produce an output gate activation vector (o_t). Where: W is the weight matrix of forget gate, $[h_{t-1}, x_t]$ is the concatenation of the previous hidden state and the current input., b is the bias of forget gate and σ is the sigmoid activation function tanh is the hyperbolic tangent activation function.

$$o_t = \sigma(W_t \cdot [h_{t-1}, x_t] + b_0)$$

$$h_t = o_t \cdot \tanh(C_t)$$

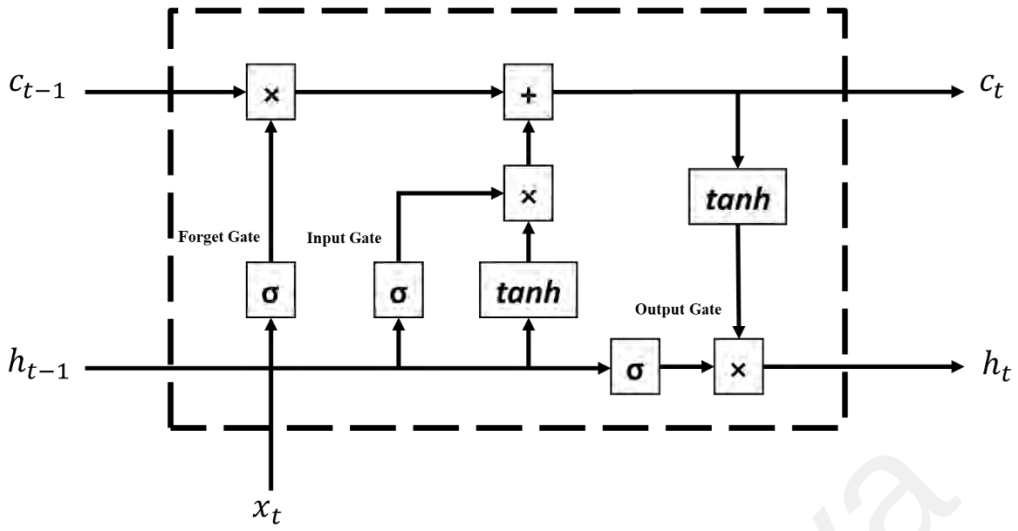


Figure 3.4: Architecture of Long Short-Term Memory Cell

Moreover, the concept of Bidirectional has been introduced. Bi-LSTM is a sequential processing model that involves two LSTMs: one processes input in the forward direction, while the other processes input in the reverse direction. By leveraging bidirectional processing, Bi-LSTM offers a broader contextual understanding, enabling quicker and more comprehensive learning of complex patterns. Its ability to store information from both the past and future makes it well-suited for sentiment and sarcasm analysis.

3.3.2 Bidirectional Encoder Representations from Transformers (BERT)

BERT (Bidirectional Encoder Representations from Transformers) is a powerful pre-trained transformer-based model that has achieved a state-of-the-art performance on a wide range of NLP tasks, including sentiment analysis and sarcasm detection (Vaswani, et al., 2017). At its core, BERT uses a transformer-based architecture composed of a multi-layer bidirectional encoder. The encoder consists of several transformer blocks, each of which includes a self-attention mechanism and a feedforward neural network. The self-attention mechanism enables BERT to model long-range dependencies between words in a sentence by attending to all other words in the sentence, not just those that are immediately adjacent. This mechanism enables the model to capture complex

relationships between words helping it in achieving high accuracy on a wide range of NLP tasks.

BERT is pre-trained on massive amounts of unlabeled text using a masked language modeling (MLM) task. The pre-training dataset typically consists of billions of words sourced from a diverse range of texts, including books, articles, and websites. This extensive dataset allows BERT to capture a broad understanding of language syntax and semantics. The pre-training process involves randomly masking words in a sentence and training the model to predict the masked words based on their surrounding context.

The pre-training phase involves training BERT on large-scale datasets, often using millions or even billions of sentences, with parameters such as the number of attention heads, layers, and hidden units defined in the model architecture. For example, a common version of BERT-base has 12 attention heads, 12 layers, and 110 million parameters.

To fine-tune BERT for sentiment analysis and sarcasm detection, the pre-trained model is further trained on labeled datasets specific to these tasks. The fine-tuning dataset includes texts annotated with sentiment labels for sentiment analysis and labeled data for sarcasm detection. The fine-tuning process refines the model's parameters, adapting it to the nuances of sentiment expression or sarcasm based on the contextual cues present in the labeled datasets. The final fine-tuned BERT model becomes highly effective for these downstream NLP tasks, leveraging its pre-trained knowledge and task-specific adaptations.

In sentiment analysis, BERT is particularly effective because of its ability to capture the context in which words are used. By considering the context in which a word appears, BERT can accurately determine the sentiment of a sentence, even when the sentiment is expressed in a subtle or complex manner. Additionally, the bidirectional nature of BERT

allows it to consider both the preceding and succeeding context of a given word, further enhancing its ability to accurately determine the sentiment of a sentence.

In sarcasm detection, BERT's ability to model long-range dependencies and capture context is also highly effective. Sarcasm is often conveyed through the use of words or phrases that are contradictory to their literal meaning, making it difficult for traditional NLP models to detect. However, by considering the context in which these words or phrases are used, BERT can accurately detect sarcasm with a high degree of accuracy.

In short, BERT is a highly effective model for sentiment analysis and sarcasm detection, thanks to its transformer-based architecture, self-attention mechanism, and pre-training on large amounts of unlabeled text. Its ability to capture context and model long-range dependencies makes it highly effective at these tasks and has led to state-of-the-art performance on various benchmarks.

3.3.3 Generative Pre-trained Transformer (GPT)

GPT is another highly popular transformer-based model for natural language processing, which has also achieved a state-of-the-art performance on various NLP tasks. The architecture of GPT is similar to BERT, in that it is also composed of a multi-layer transformer encoder. However, while BERT is pre-trained using a masked language modelling task, GPT is pre-trained using a language modelling task, where the model is trained to predict the next word in a sequence given the preceding words.

GPT utilizes a transformer decoder architecture, which enables it to generate new text by predicting the next word in a sentence given the preceding words. This architecture is particularly useful for language generation tasks and has been used to generate text in a variety of contexts, including story writing, machine translation, and chatbots. GPT also makes use of a self-attention mechanism, which enables it to capture long-range

dependencies between words in a sentence. The self-attention mechanism allows the model to focus on the most relevant parts of the input sequence, enabling it to generate accurate predictions.

In sentiment analysis and sarcasm detection, GPT has also shown promising performance. Because GPT has been trained to model the relationships between words in a sentence, it is able to accurately capture the sentiment of a given text. The model can also be fine-tuned on specific sentiment analysis or sarcasm detection datasets to further improve its performance on these tasks.

In short, GPT is a highly effective transformer-based model for natural language processing that has shown promising results in sentiment analysis and sarcasm detection. Its transformer decoder architecture and self-attention mechanism enable it to accurately capture the relationships between words in a sentence, making it highly effective for text generation and other NLP tasks.

3.3.4 Bi-LSTM vs Transformer-based Models

Traditionally, deep recurrent neural networks (RNNs) such as bi-directional LSTM (Bi-LSTM) have been widely used in sentiment analysis and sarcasm detection. However, in recent years, transformer-based models, such as BERT and GPT, have emerged as powerful alternatives to Bi-LSTM. These models utilize self-attention mechanisms that enable them to capture long-range dependencies between words in a sentence, making them highly effective for tasks that require understanding of the relationships between words.

Several studies have compared the performance of Bi-LSTM and transformer-based models in sentiment analysis and sarcasm detection. For example, a study by Devlin et al. shows that BERT outperforms Bi-LSTM on various benchmarks in sentiment analysis

(Devlin, et al., 2019). Similarly, a study by Hongchan et al. shows that BERT outperforms Bi-LSTM in sentiment analysis (Hongchan, Yu, Zishuai, & Haodong, 2021).

The main advantage of transformer-based models over Bi-LSTM is their ability to capture long-range dependencies between words in a sentence. This enables them to more accurately capture the relationships between words, even when they are far apart in the sentence. Additionally, transformer-based models have been pre-trained on massive amounts of unlabelled data, which allows them to better understand the syntax and semantics of language. Despite the advantages of transformer-based models, they also have some drawbacks. For example, they require significant computational resources to train and use effectively, and they may be prone to generating biased or inappropriate responses in certain contexts.

Overall, the studies reviewed indicate that transformer-based models, such as BERT and GPT, tend to outperform Bi-LSTM networks in a variety of natural language processing tasks, including sentiment analysis and sarcasm detection. While transformer models can be computationally expensive, they offer a significant performance boost over Bi-LSTM networks. With this in mind, the proposed framework in this thesis leverages a transformer-based model.

3.3.5 BERT vs GPT

The field of NLP has witnessed remarkable advancements in recent years. Powerful language models have been extensively utilized for various language tasks, including sentiment analysis and sarcasm detection. However, the selection of the most appropriate model can significantly impact the performance and accuracy of the results.

While BERT and GPT are built using transformer architecture, they differ in their pre-training approaches. BERT is a bidirectional model pre-trained on a large corpus of text

using a masked language modelling objective. It learns contextual relations between words in a sentence and can be fine-tuned on specific tasks such as sentiment analysis or sarcasm detection with limited labelled data. On the other hand, GPT is a generative model pre-trained using unsupervised language modelling, focusing on generating text by predicting the next word in a sequence. While GPT excels at language generation, it may not be the ideal choice for, that demand a nuanced understanding of language and context, including sentiment analysis and sarcasm detection.

For sentiment analysis and sarcasm detection, BERT has demonstrated superior effectiveness compared to GPT. BERT's capacity to capture the meaning of words in context is critical for accurately detecting sentiment and sarcasm. Moreover, BERT can perform multi-task learning, enabling it to simultaneously classify sentiment and detect sarcasm within a single model. Apart from its exceptional performance, BERT has gained widespread acceptance in the research community, with numerous studies showcasing its effectiveness for various language tasks. Its popularity is further bolstered by its high customizability, allowing researchers and practitioners to fine-tune the pre-trained model for specific tasks using minimal task-specific data.

In conclusion, although both BERT and GPT are formidable language models, BERT stands out as the preferred choice for sentiment analysis and sarcasm detection tasks due to its ability to capture contextual meaning and excel in multi-task learning. Its popularity and versatility make it an appealing option for researchers and practitioners in the dynamic field of natural language processing.

3.4 BERT-based Multitask Learning Deep Neural Network

In this thesis, we aim to propose a framework is to enhance the performance of sentiment classification by incorporating sarcasm detection as an auxiliary task. Traditionally, this has been done by training two separate models for sentiment and

sarcasm detection, which can be inefficient, redundant, and computationally expensive (Yunitasari, Musdholifah, & Sari, 2019). To overcome these limitations, multi-task learning is proposed in our framework, where both sentiment analysis and sarcasm detection are trained simultaneously using a shared BERT layer. This approach offers several advantages, including reduced overfitting, improved data efficiency, and faster learning by leveraging shared representations. Additionally, multi-task learning helps address known weaknesses of deep learning methods, particularly the BERT model, such as high computational demand and large-scale data requirements (Crawshaw, 2020).

The overall architecture of the proposed BERT-based multi-task learning deep neural network is illustrated in Figure 3.5. The proposed method features two classification tasks sharing a single-layer BERT as the encoder, with a separate multilayer perceptron (MLP) layer for prediction. The detailed explanation of the architecture is given in the following subsections.

3.4.1 Embedding Layer

The embedding layer of BERT is a crucial component that transforms input text into numerical representations for processing by the neural network. It utilizes WordPiece embeddings, position embeddings, and token embeddings to create a dense representation of the input text.

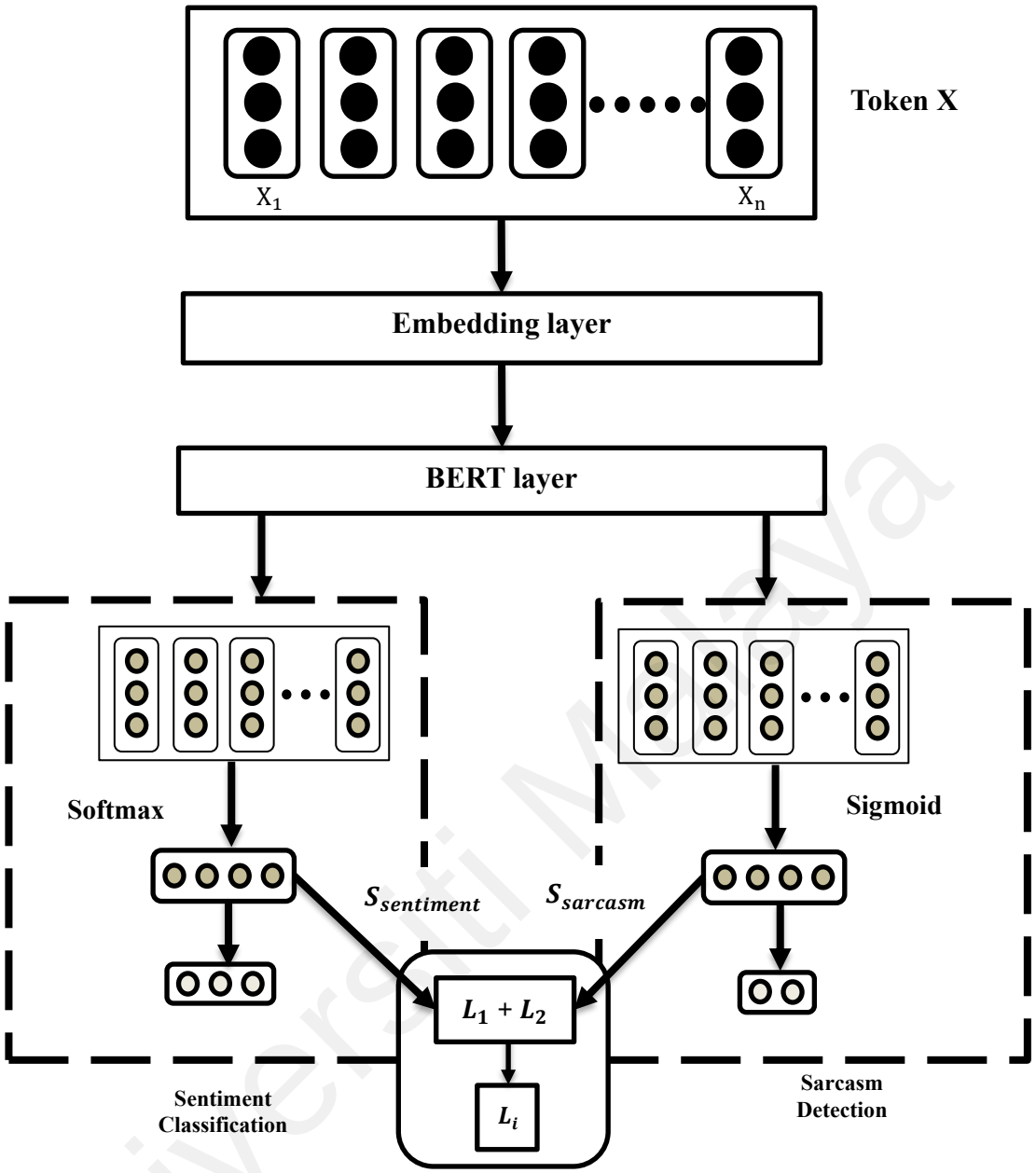


Figure 3.5: Architecture of the Proposed Method (BERT-based Multitask Learning)

WordPiece embeddings are similar to traditional word embeddings but break down words into smaller subwords. For instance, the word "unhappy" might be represented as "un," "##happy," and "##ly." Each subword is assigned a fixed-length vector representation, learned during pre-training. This enables BERT to better capture word morphology and handle out-of-vocabulary words.

Position embeddings play a crucial role in BERT by encoding the positional information of each token within the input sequence. Unlike traditional recurrent neural

networks, which inherently capture sequential information, transformers like BERT lack a built-in sense of token order. To address this limitation, position embeddings are introduced to provide the model with a sense of token position and sequence structure.

In the case of BERT, position embeddings are added to the WordPiece embeddings for each token. WordPiece embeddings represent the semantic content of individual tokens, while position embeddings help the model understand the sequential arrangement of tokens. The combination of these embeddings results in what is known as the token embedding.

The token embedding, formed by the summation of WordPiece and position embeddings, enables BERT to distinguish between different types of tokens in the input sequence. For instance, BERT uses special tokens like [CLS] (classification) and [SEP] (separator). The [CLS] token is positioned at the beginning of the sequence and serves as a representation for the entire input sequence. It is particularly important for tasks like sentence classification or sentiment analysis. The [SEP] token, on the other hand, is used to separate multiple input sequences when working with paired inputs, such as question-and-answer pairs.

Mathematically, the WordPiece embeddings are learned using an embedding matrix with a size of:

$$matrix\ size = vocabulary_{size} \times embedding_{dimension} \quad \text{Eq. (1)}$$

where vocabulary_size is the number of unique subwords in the vocabulary, and embedding_dimension is the desired dimensionality of the embeddings.

In mathematical terms, the token embedding is the element-wise sum of the corresponding elements in the WordPiece embedding vector and the position embedding vector for a given token i:

$$Token_i = WordPiece_i \times Position_i$$

This process is applied to all tokens in the input sequence, creating a sequence of token embeddings that captures both the semantic content of each word (WordPiece embeddings) and its positional information in the sequence (Position embeddings). This combined representation enables BERT to understand the context and relationships between words based on both their meanings and their positions in the input sequence.

3.4.2 BERT Layer

After the embedding layer, the BERT model uses a series of transformer layers to learn contextual representations of the input sequence for both sentiment analysis and sarcasm detection. Mathematically, the output of the embedding layer can be represented as:

$$X_{token} = [x_1, x_2, \dots, x_n] \quad \text{Eq. (2)}$$

where X_{token} is the input sequence consisting of n tokens of sentiment analysis and sarcasm detection input data, and each token is represented by a d-dimensional embedding vector.

Afterwards, X_{token} is passed to the transformer layer consists of multiple transformer blocks. The transformer layers use a self-attention mechanism to attend to different parts of the input sequence and capture long-range dependencies between tokens. The output of the i-th transformer block can be represented as:

$$H_i = TransformerBlock(H_{i-1}) \quad \text{Eq. (3)}$$

Where H_{i-1} is the output of the (i-1)-th transformer block, and TransformerBlock is the operation performed by the i-th transformer block.

The self-attention mechanism in each transformer block can be expressed mathematically as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad \text{Eq. (4)}$$

where Q, K, and V are the query, key, and value matrices, respectively, d_k is the dimensionality of the key vectors, and the softmax function takes the vector scores and transform them to a probability distribution. The output of the self-attention mechanism is then passed through a feedforward neural network and added to the residual connection of the input to obtain the output of the transformer block.

After passing through multiple transformer blocks, the output of the final transformer block is a sequence of hidden states, where each hidden state corresponds to a token in the input sequence. The transformer layers in BERT-based multi-task learning allow the model to learn a shared representation for both tasks, which is beneficial for reducing overfitting and improving accuracy. Mathematically, the output of the transformer layer for the shared representation can be represented as:

$$H_* = [h_1, h_2, \ldots, h_n] \quad \text{Eq. (5)}$$

where each h_i is the hidden state corresponding to the i-th token in the input sequence and H_* represent the shared representation of output for sentiment analysis and sarcasm detection.

3.4.3 Multi-layer Perceptron

The shared representation H_* obtained from the transformer layers is passed through a fully connected layer to obtain the output for each task. The multilayer perceptron (MLP) can be seen as a classifier that maps the shared representation to the output of each task. The dotted box in Figure 3.5 gives a deep insight of the MLP used in the proposed framework. It can be observed that the primary task and the secondary task have similar architecture except for the activation functions of the dense layer.

For sentiment analysis, the shared representation is fed into a fully connected layer followed by a softmax activation function to obtain the sentiment classification output. The softmax activation function is chosen because the sentiment analysis task involves multi-class classification. Mathematically, this can be represented as:

$$Z_{sentiment} = FC_{sentiment}(H_*) \quad \text{Eq. (6a)}$$

$$S_{sentiment} = softmax(Z_{sentiment}) \ldots. \quad \text{Eq. (6b)}$$

where $FC_{sentiment}$ is the fully connected layer for sentiment analysis.

For sarcasm detection, the shared representation is fed into a separate fully connected layer followed by a sigmoid activation function to obtain the sarcasm classification output. The sigmoid activation function is chosen because the sarcasm detection task involves binary-class classification. Mathematically, this can be represented as:

$$Z_{sarcasm} = FC_{sarcasm}(H_*) \quad \text{Eq. (7a)}$$

$$S_{sarcasm} = sigmoid\big(FC_{sarcasm}(H_*)\big) \quad \text{Eq. (7b)}$$

where $FC_{sarcasm}$ is the fully connected layer for sarcasm detection.

3.4.4 Loss Function

The configuration of the frameworks allows the secondary task to inform the training on the primary task by computing the loss of the shared model using equation 8:

$$L_{shared} = \alpha L_{sentiment} + (1 - \alpha) L_{sarcasm} \quad \text{Eq. (8)}$$

where $L_{sentiment}$ denoted as the loss for the primary task, $L_{sarcasm}$ denoted as the loss for the secondary task, L_i is the total loss, and α denoted as the learning rate of the proposed model.

The cross-entropy losses are given in equations 9 and 10, respectively, for sentiment classification $L_{(sentiment)}$ and sarcasm classification $L_{(sarcasm)}$.

$$Categorical\ Cross\ Entropy = -\sum_i y_i \log(\hat{y}_i) \quad \text{Eq. (9)}$$

$$Binary\ Cross\ Entropy = -y_i \log(\hat{y}) - (1 - y) \log(1 - \hat{y}) \quad \text{Eq. (10)}$$

where y is the true label (either 0 or 1), $\hat{y}$ is the predicted probability of the positive class (between 0 and 1).

In order to optimize the performance of the model, the Adam optimizer is used during training. At each epoch, the algorithm calculates the gradient of the loss function with respect to each parameter for every batch of data, indicating the direction for parameter updates. Adam introduces adaptivity in the learning rates by maintaining moving averages of the first-order moment (mean of gradients) and the second-order moment (uncentered variance of gradients). These moving averages are utilized to dynamically adjust the learning rates for each parameter, allowing the algorithm to handle varying scales of gradients and parameter contributions. Additionally, Adam includes bias correction to address biases introduced by the moving averages during early training epochs, ensuring unbiased estimates. The combination of adaptive learning rates and bias

correction makes Adam a powerful optimization algorithm, widely utilized in training neural networks, as it mitigates challenges associated with gradient scaling and contributes to the overall efficiency of the training process.

CHAPTER 4: RESULTS AND DISCUSSION

4.1 Simulation Settings

4.1.1 Evaluation Metrics

The following standard metrics are used in this thesis to benchmark the performance of the proposed framework.

i. **Recall**, also known as sensitivity or true positive rate, is a measure of the ability of a model to correctly detect positive instances in a dataset. It is calculated as the ratio of true positives (TP) to the total of true positives and false negatives (FN), as shown in equation 10:

$$Recall = \frac{TP}{TP+FN} \ldots \text{Eq. (10)}$$

ii. **Precision** is the accuracy of positive predictions, and it is defined as the ratio of true positives to the total predicted positives, including both true positive (TP) and false positive (FP) as given in equation 11.

$$Precision = \frac{TP}{TP+FP} \ldots \text{Eq. (11)}$$

iii. **F1-score** is a helpful metric to compare two classifiers. F1 score considers both recall and precision, which is defined as equation 12.

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \ldots \text{Eq. (12)}$$

Another commonly used parameter is accuracy, which represents the ratio of total correct predictions to the total number of predictions. However, the accuracy metric does not take into consideration false negatives and false positives in its calculation, making this metric highly unreliable. To illustrate this, let's imagine a dataset of 1000 tweets, with 900 being sarcastic and 100 not sarcastic. Suppose the classifier predicts all tweets as sarcastic, the model's performance will have a high accuracy of 90%. However, in reality, the F1-score metric, which considers both precision and recall, will be 0 due to the high

number of false negatives. Hence, the experiments conducted will utilize the F1 score as the primary metric to evaluate the performance of the proposed model. The F1 score provides a better balance between precision and recall and is more suitable for tasks where class imbalances or misclassifications need to be taken into account.

4.1.2 Experiment Overview

The initial phase of the experiment involves experiments 1 and 2, which are dedicated to establishing the proof of concept. The experiment will commence by evaluating computationally cheaper models, including the Bi-LSTM model, alongside other standalone deep learning models. The primary objective in this preliminary analysis is to assess the performance of the Bi-LSTM model in comparison to conventional deep learning methods.

This preliminary analysis aims to offer valuable insights into the strengths and weaknesses of the Bi-LSTM model, enabling the identification of areas for enhancement. Furthermore, it serves to validate the effectiveness of the proposed framework for improving sentiment analysis through sarcasm detection. By comparing the performance of the Bi-LSTM model with other computationally efficient deep learning methods, the goal is to ascertain the most suitable model for the given dataset.

Subsequent experiments, namely experiments 3, 4, and 5, involve a comprehensive evaluation of the proposed BERT model and GPT, assessing their effectiveness in comparison to the Bi-LSTM model. Through the assessment of multiple models, we intend to determine the most effective approach for sentiment analysis and sarcasm detection. The insights gained from these evaluations will significantly contribute to enhancing the accuracy and reliability of sentiment analysis models for real-world applications.

Below are the details of the conducted experiments:

i. Experiment 1: Deep Learning Model Baselines and Variants

Experiment 1 focuses on establishing the proof of concept of the proposed method by testing computationally cheaper models, such as various deep learning models (DNN, CNN, RNN, and LSTM).

ii. Experiment 2: Evaluation of Deep Learning Model on Unbiased Datasets

Experiment 2 incorporates a multi-task learning framework into the best deep learning model identified in Experiment 1. Sentiment analysis and sarcasm detection are jointly trained on this model and compared with the deep learning model that lacks sarcasm detection capabilities. This approach provides valuable initial insights into the proposed method's concept and identifies its weaknesses.

iii. Experiment 3: Comparison of BERT, GPT and Bi-LSTM as Proposed Methods for Sentiment Analysis

The primary focus of this experiment was to determine and assess whether BERT is the most effective machine learning method to be used as the proposed method. While the MTL-Bi-LSTM method demonstrated superior performance in previous experiments, its limitations in neutral sentiment classification necessitated an evaluation of transformer-based methods such as GPT and BERT. This investigation aims to further enhance the proposed method by comparing its performance with these powerful language models.

iv. Experiment 4: Comparison of the Proposed Methods with Other Baseline Models

This experiment aimed to compare the proposed method's performance against other existing state-of-the-art methods, using different widely benchmarked datasets. By evaluating the proposed method alongside established methods on diverse datasets, experiment 4 able to assess its effectiveness and competitiveness in various scenarios.

v. Experiment 5: Evaluation of the Impact of Sarcasm Detection on the Sentiment Analysis Performance

This experiment focused on studying the impact of sarcasm detection on sentiment analysis. Specifically, a dataset containing instances of sarcasm was carefully chosen for this study. The primary objective was to evaluate how the presence of sarcasm influences the accuracy and effectiveness of the proposed method. By comparing the proposed method's performance with and without sarcasm detection on sentiment analysis datasets with sarcastic comments, the experiment gains valuable insights into the challenges and opportunities of incorporating sarcasm detection into sentiment analysis algorithms.

4.1.3 Environment and Hyperparameter Settings

The experimental configurations for both Experiment 1 and Experiment 2 are outlined in Table 4.1. Word2Vec was employed with a dimension of 200 for embedding words. A dropout rate of 0.4 countered overfitting, while training extended across 50 epochs. A learning rate of 0.001 facilitated gradual weight adjustments, with batches of 128 examples processed per iteration. Cross-entropy served as the loss function for classification, and the RMS Prop optimizer fine-tuned parameters effectively. These hyperparameters collectively laid the foundation for successful model training and subsequent analysis.

Table 4.1: Hyperparameters Settings for Experiment 1 and 2

Parameters	Values
Embedding Dimension	Word2Vec = 200
Dropout	0.4
Epoch	50
Learning Rate	0.001
Batch Size	128
Loss Function	Cross-entropy
Optimizer	RMS Prop

In the subsequent experiment, particularly Experiment 4 which involves benchmarking against other methods, the setups were carefully made to ensure fairness by adhering to the experimental settings and hyperparameters of the benchmarked works. The implementation utilized the Keras library with TensorFlow, and the model was trained and tested on Google Colab with T5 GPU acceleration. For word embeddings, WordPiece embeddings were employed and the uncased version of the BERT-BASE model, which consists of 12 network layers, a hidden layer dimension of 768, and 12 attention heads was used. The model was pretrained with over 110 million parameters. For optimization, the Adam optimizer with a learning rate of 2e-5 was utilized, and the dropout rate was set to 0.5 to prevent overfitting. The sigmoid function was employed in the dense layer to calculate the probabilities of the class labels. The architecture was trained for a total of 10 epochs. The details of the hyperparameters are summarized in Table 4.2.

4.1.4 Sentiment and Sarcasm Truth Table

Table 4.3 illustrates the relationship between sentiment and sarcasm classifications. In scenarios where a tweet exhibits positive sentiment while concurrently conveying sarcasm, it is assigned a negative classification. A concrete illustration of this

Table 4.2: Hyperparameter Settings for Experiment 3, 4 and 5

Parameters	Values
GPT Model	GPT-2
BERT Model	BERT-BASED uncased
Embedding Dimension	300
Dropout	0.5
Epoch	10
Learning Rate	0.00002
Batch Size	16
Loss Function	Binary cross-entropy
Optimizer	Adam Optimizer

phenomenon can be found in the following tweet: "So many assignments, I love school life so much."

The sentiment classifier identifies the sentiment within this tweet as positive due to the presence of the phrase "Love school life so much," which conveys a positive sentiment. However, the sarcasm classifier recognizes the underlying sarcasm stemming from the phrase "So many assignments." As a result of this duality, the tweet is ultimately classified as negative.

For forthcoming implementations centered around sarcasm detection, the classification protocol for sentences imbued with sarcastic sentiment will be modeled after the aforementioned example. This approach allows for a more nuanced and accurate interpretation of tweets that exhibit both positive sentiment and subtle sarcasm.

Table 4.3: Sentiment and Sarcasm Relationship Table

Sentiment score	Sarcasm score	Output
Negative	Not sarcastic	Negative
Negative	Sarcastic	Negative
Positive	Not sarcastic	Positive
Positive	Sarcastic	Negative

4.2 Experiment 1: Deep Learning Model Baselines and Variants

4.2.1 Deep Learning Model Variants

The dataset shown in Figure 3.2(a) and (b) is split into 80% for training and 20% for testing. The following baselines and variations of the models are compared.

i. Standalone classifier with deep learning algorithms such as (RNN, CNN, Bi-GRU and Bi-LSTM) with the following characteristics.

$$Z_{sentiment} = FCLayer_{sentiment}(RNN(X_{sentiment}))$$

$$Z_{sarcasm} = FCLayer_{sarcasm}(RNN(X_{sarcasm}))$$

$$S_{sentiment} = Softmax(Z_{sentiment})$$

$$S_{sarcasm} = Sigmoid(Z_{sarcasm})$$

In this case, there are two standalone models respectively, for sentiment and sarcasm models. X is the list of input sentences obtained from the word embeddings. Then, X was fed to the deep learning algorithm and then pass the output through a fully connected layer (FCLayer) to get the sentence representation Z. Finally, Z is passed through the dense layer, which is softmax and sigmoid for sentiment and sarcasm classification, respectively ($S_{sentiment}$ & $S_{sarcasm}$).

ii. Multi-task learning using Bi-LSTM with the following characteristics:

$$Z_* = FCLayer_*(bidirection(LSTM(X)))$$

$$S_{sentiment} = Softmax(Z_*)$$

$$S_{sarcasm} = Sigmoid(Z_*)$$

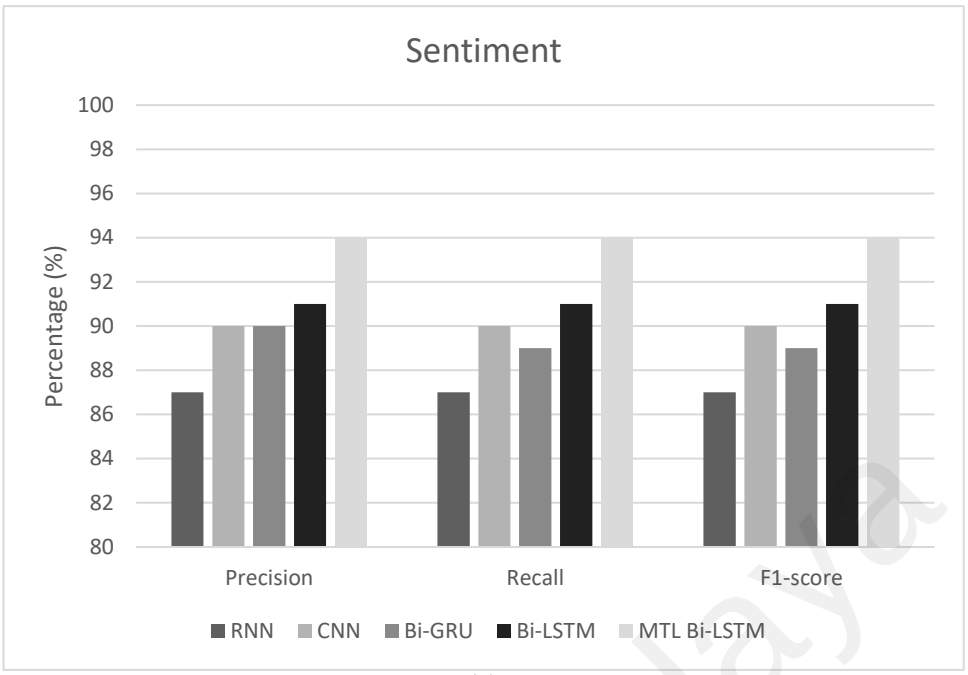
where * represents the shared layer between two tasks sentiment.

4.2.2 Results and Discussion

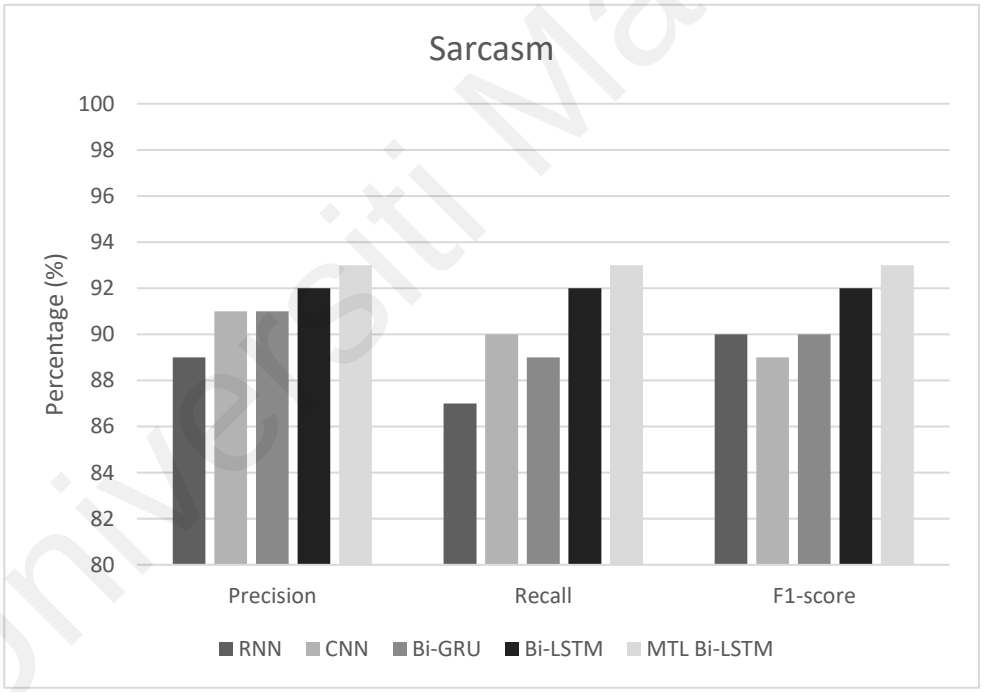
The experiment results presented in Figure 4.1 reveal valuable insights into the performance of different models for sentiment and sarcasm classification.

Comparing the four standalone models, it is evident that conventional neural networks, such as RNN, CNN, and Bi-GRU, yield inferior outcomes compared to the more recent Bi-LSTM network. The Bi-LSTM model outperforms these methods by a margin of 0.4%, 0.1%, and 0.1%, respectively. Notably, the Bi-LSTM model achieves remarkable F1-scores of 91% for sentiment classification and 92% for sarcasm classification. This improvement is attributed to the utilization of memory gates, allowing the network to preserve important information while discarding irrelevant data. Additionally, the Bi-LSTM network enables bidirectional learning, facilitating the retention of information from both the past and the future for more effective classification.

Moving on to the comparison with the multi-task learning Bi-LSTM model, it becomes evident that the multi-task learning approach outperforms all other models in both sentiment and sarcasm classification, achieving F1-scores of 94% and 93%, respectively. Compared to the standalone Bi-LSTM model, the multi-task learning Bi-LSTM model demonstrates a significant improvement of 3% and 1% in sentiment and sarcasm classification, respectively. This enhancement is due to multi-task learning utilizing a



(a)



(b)

Figure 4.1: Experiment Results using Different Variety of Models: (a) Sentiment Classification; (b) Sarcasm Classification

shared layer, which reduces the risk of overfitting during gradient descent and ensures efficient learning. It is noteworthy that the sarcasm classifier contributes to the enhanced performance of the sentiment classifier, showcasing the benefits of jointly training on both tasks. The improvement in sentiment classification is more substantial than in sarcasm classification, as the latter is a subtask of sentiment analysis. Importantly, the

performance of the multi-task learning Bi-LSTM model remains consistent when analyzing precision and recall values, showing a similar pattern as the F1-Score. The margin of improvement is approximately 3% and 1% for sentiment and sarcasm classifications, respectively. This indicates the robustness of the multi-task learning approach across different evaluation metrics.

Overall, the results demonstrate the effectiveness of the multi-task learning Bi-LSTM model in sentiment and sarcasm classification, outperforming both conventional neural networks and standalone Bi-LSTM models. The utilization of memory gates, bidirectional learning, and joint training of both tasks contribute to the significant improvements observed in the model's performance.

4.3 Experiment 2: Evaluation of Deep Learning Models on Unbiased Datasets

In this experiment, we aim to evaluate the performance of the multi-task learning Bi-LSTM model using different datasets and compare it with the Bi-LSTM standalone model, which was identified as the best standalone model in the previous experiment. The primary objective is to analyze how well the model performs on unseen data and assess its effectiveness in real-life scenarios.

4.3.1 Datasets

Three datasets obtained from Kaggle are utilized for sentiment analysis task, and they are summarized in Table 4.4, as explained below:

i. **Reddit Dataset**:

The Reddit dataset is collected from the comment section of Reddit posts. This dataset is characterized by being unbalanced, with a majority of positive sentiments. It serves the purpose of the experiment well as it provides data from

a different social media platform, potentially having a different writing style and structure compared to the datasets used in previous experiments.

ii. **Twitter US Airline Dataset**:

This dataset consists of reviews written by customers of US airlines and collected from Twitter. Contributors were asked to classify the nature of their comments as positive, negative, or neutral tweets, followed by providing reasons for their comments. Similar to the Reddit dataset, this dataset is also unbalanced, with the majority of data belonging to negative sentiments.

iii. **Twitter Dataset**:

The Twitter dataset was scraped from Twitter using the Tweepy module. It is an unbalanced dataset as well, with negative sentiments being the majority of the data. This dataset is the closest in nature to the dataset used to train models in previous experiments, providing a more direct comparison with the performance achieved earlier.

4.3.2 Results and Discussion

In Figure 4.2, the F1-scores achieved by the sentiment classifiers on various datasets are showcased. Among the three datasets, the Twitter dataset exhibits the highest F1-score, standing at 58% for the standalone sentiment classifier and 63% for the proposed multi-task learning Bi-LSTM model. The Reddit dataset follows closely with F1-scores

Table 4.4: Distribution of Sentiment Dataset

Datasets	Positive	Negative	Neutral
Reddit	15830 (42%)	8277 (22%)	13142 (35%)
Twitter US Airline	2363(16%)	9178 (63%)	3090 (21%)
Twitter	1103 (17%)	4001 (61%)	1430 (22%)

of 55% for the standalone sentiment classifier and 61% for the proposed approach. Meanwhile, the Twitter US airline dataset records an F1-score of 52% for the standalone sentiment classifier and 54% for the proposed method.

The superior performance on the Twitter dataset could be attributed to its similarity to the data used for training the model but with a broader scope. This similarity might have facilitated better generalization of the model to handle sentiments present in this dataset effectively.

However, when assessing the performance of the proposed multi-task learning Bi-LSTM model on unbiased data in general, the average performance is relatively low, at around 59%. To better understand this issue, the neutral label's F1-scores for the proposed method on these datasets are examined, as shown in Figure 4.3. The poor overall performance is largely influenced by the deficient neutral label F1-scores, which are 48%, 51%, and 44% for the Reddit, Twitter, and Twitter US airline datasets, respectively. The low F1-scores for the neutral class can be primarily attributed to the low precision scores

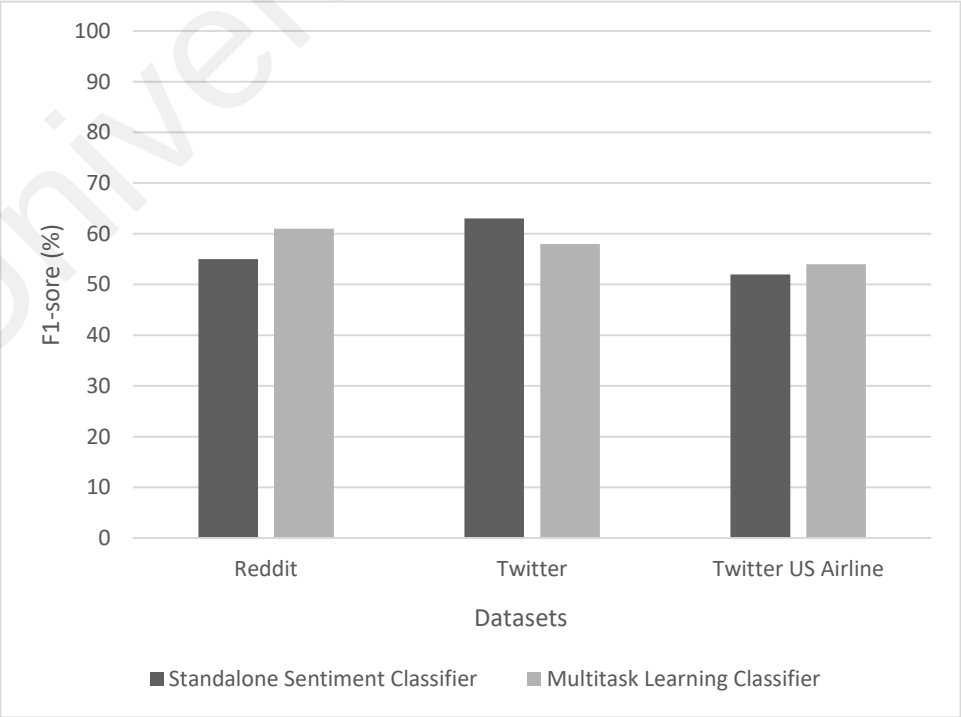


Figure 4.2: Experiment Results of Standalone and Multitask Classifiers on Unbiased Datasets

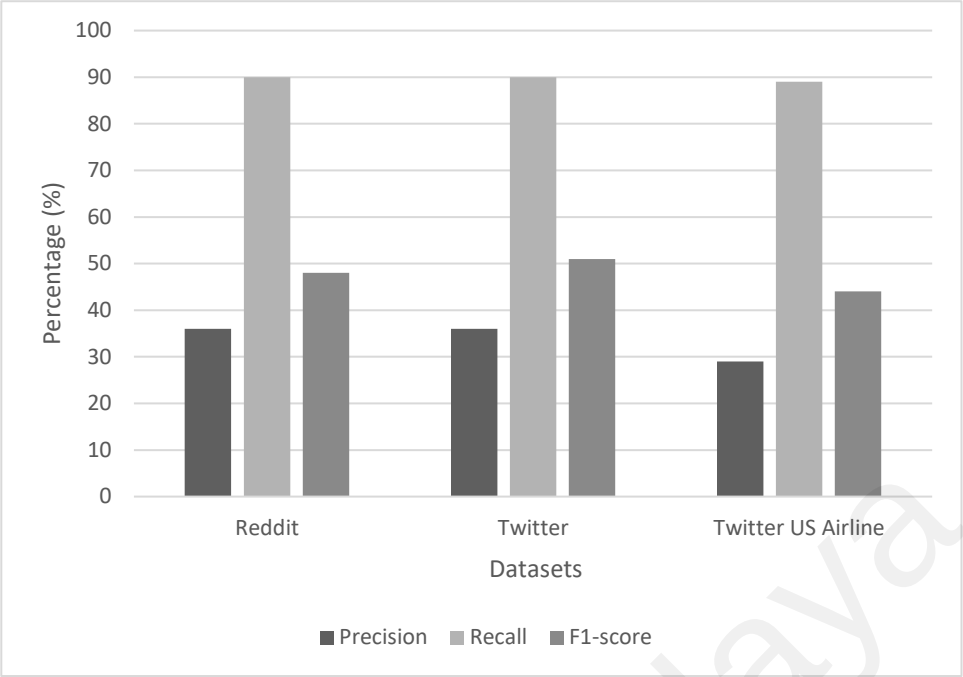


Figure 4.3: Neutral Score of the MTL-Bi-LSTM Model on Unbiased Datasets

when classifying neutral sentiments, which are 33%, 36%, and 29% for the same datasets, respectively. This indicates that a significant number of neutral sentiment sentences are being misclassified as false negatives.

Upon further examination, it is observed that the low recall for neutral sentiments is due to the presence of many neutral words, mostly stop words, which are removed during the pre-processing stage. Consequently, the model encounters difficulty in accurately classifying neutral tweets. Figure 4.4 shows the word cloud for the neutral label in these datasets, highlighting the neutral words that may contribute to the misclassification.

Despite these challenges, the proposed multi-task learning Bi-LSTM model still outperforms existing methods when analysing unbiased datasets. This indicates the model's potential and effectiveness in sentiment analysis tasks even though it faces difficulties with neutral sentiment classification on some datasets.

To address these limitations, transformer-based large language models (LLMs) like BERT and GPT have emerged as popular solutions. LLMs are pre-trained on massive

amounts of data and have the ability to capture complex language structures and nuances, including those present in neutral sentiment. Their attention mechanisms enable them to focus on the most relevant parts of the text, resulting in improved accuracy.

In light of these findings, we propose a BERT-based multi-task learning framework to handle sentiment analysis and sarcasm detection. By leveraging the benefits of multi-task learning with the powerful capabilities of BERT, this framework aims to effectively use resources and improve performance. It is expected that using a BERT-based model will lead to significant improvements in neutral sentiment detection, resulting in more accurate sentiment analysis and sarcasm detection.

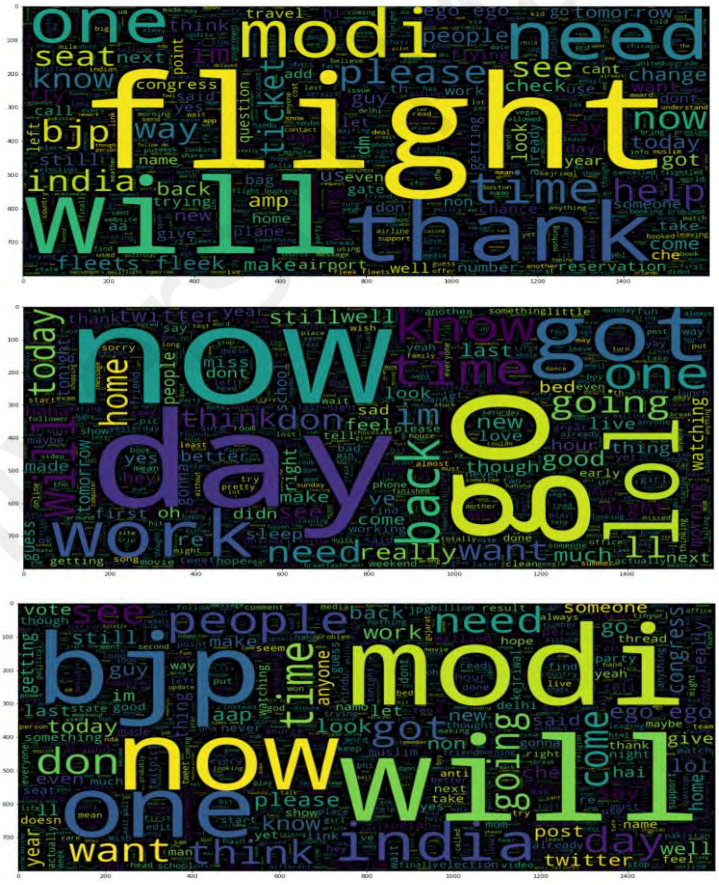


Figure 4.4: WordCloud for Neutral Label in Unbiased Datasets

4.4 Experiment 3: Comparison of BERT, GPT and Bi-LSTM as Proposed Methods for Sentiment Analysis

4.4.1 Datasets

This experiment evaluated the MTL-Bi-LSTM, MTL-GPT model and the proposed MTL- BERT model using the dataset shown in Table 4.4, divided into 80% for training and 20% for testing. The primary objective was to assess the performance of the popular transformer-based methods and compare them with the performance of the MTL-Bi-LSTM model to determine the most effective choice for the proposed method.

4.4.2 Results and Discussion

Figure 4.5 presents the experimental results of the proposed method compared to the MTL-Bi-LSTM and MTL-GPT models in Experiment 3. The proposed model achieved the best performance with an impressive 97.5% and 99.6% F1-score for sentiment analysis and sarcasm detection, respectively. It outperforms the MTL-Bi-LSTM model by 3.5% and 6.6% in sentiment analysis and sarcasm detection, respectively. However, the MTL-GPT model performs poorly in sentiment classification, achieving an 89% F1-score. On the other hand, the MTL-Bi-LSTM model struggles in sarcasm classification, achieving only a 93% F1-score.

The superiority of the proposed model in sentiment analysis and sarcasm detection can be attributed to its use of the BERT algorithm, which is highly effective in capturing complex language structures and nuances present in sentiment and sarcasm. Unlike the MTL-GPT and MTL-Bi-LSTM models, the proposed model's bidirectional mechanism allows it to capture contextual information from both left to right directions, providing a more comprehensive understanding of the text. This enables the model to capture dependencies and relationships between words more effectively, enhancing its accuracy in sentiment and sarcasm classification.

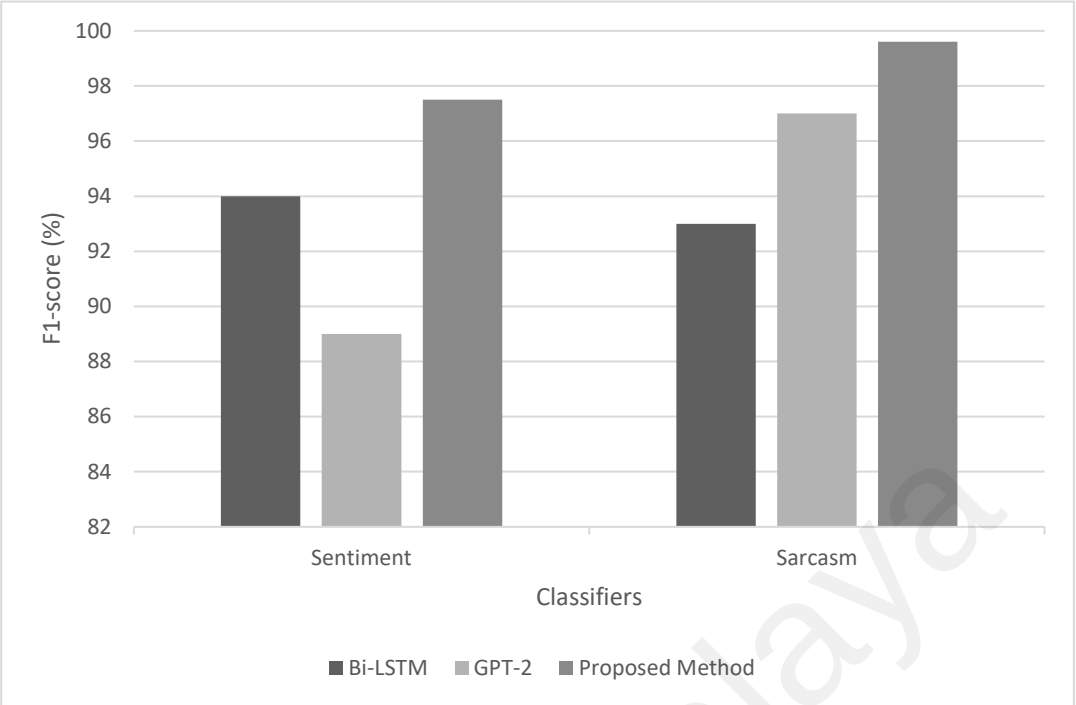


Figure 4.5: Experiment Results of Bi-LSTM vs GPT vs Proposed Method on Experiment 3

Furthermore, the proposed method utilizes the transformer's attention mechanism, which helps it capture long-range dependencies within the text and assign varying levels of importance to different words. This capability is crucial in identifying "sentiment clues" and "sarcasm clues" within words and phrases, thereby significantly improving the model's accuracy in sentiment and sarcasm classification tasks. In contrast, models lacking attention mechanisms, such as Bi-LSTM, struggle to effectively highlight and focus on these critical "sentiment clues" and "sarcasm clues," limiting their ability to capture global dependencies across the entire sequence and resulting in reduced performance in sentiment and sarcasm classification tasks.

Overall, the results demonstrate the effectiveness of the proposed Multitask learning BERT model, outperforming both MTL-GPT and MTL-Bi-LSTM models in sentiment analysis and sarcasm detection. The use of the BERT algorithm and transformer's attention mechanism has proven to be a powerful combination for capturing the complexities of sentiment and sarcasm in text, leading to significantly improved performance in these tasks.

4.5 Experiment 4: Comparison of the Proposed Method with Other Baseline Models

4.5.1 Datasets

A range of widely recognized benchmark datasets was carefully selected. These datasets represented different domains and problem types, ensuring a comprehensive evaluation of the proposed method's performance in various scenarios. The experiment evaluated the performance of the proposed method on the following three text sentiment classification datasets:

- **SST-2**: Also known as Stanford Sentiment Treebank's binary version, is a is a widely recognized benchmark dataset for sentiment analysis. It consists of a collection of movie reviews that have been labeled with positive and negative binary sentiment labels. The SST-2 dataset contains 9613 samples, evenly distributed between positive and negative. (Socher, et al., 2013)
- **MR**: The movie reviews dataset is a widely referenced dataset for sentiment classification. It consists of 5331 positive and 5331 negative samples. (Pang & Lee, 2005)
- **SemEval 2013 and 2014**: The SemEval 2013 task 2 (Nakov, et al., 2013) and 2014 task 9 (Rosenthal S. , Ritter, Nakov, & Stoyanov, 2014) were sentiment analysis competitions that focused specifically on Twitter data on a variety of subjects and domains. The datasets were annotated with sentiment labels such as positive, negative, or neutral and datasets are divided into train set, development set and test set.
- **Twitter Airline**: Twitter US Airline, consisting of reviews written by their customers. The contributors were asked to classify the nature of its comment as positive, negative, or neutral tweets, followed by their reasons for giving such

comments. It is also an unbalanced dataset, with negative sentiment being the majority of the data. (Wan & Gao, 2015)

In addition to the sentiment classification datasets, the proposed method will also train on a secondary task of sarcasm detection. Below are the details of the dataset:

- **SemEval 2018**: The SemEval 2018 task 3 was a competition on sarcasm or irony detection in English tweets. It consists of 3834 tweets for the training set and 784 tweets for the test set and were annotated sarcastic and not sarcastic.
- **News Headlines Sarcasm Dataset**: For the news headlines sarcasm dataset (Misra, 2019), there are 28619 tweets labeled with 0 for not sarcastic and 1 for sarcastic. The dataset consists of high quality news headlines written by professionals and low noise.

The evaluation done in experiment 4 using SST-2 and SemEval datasets will follow the study done by Wang et al. and Khan et al. (Wang, Zhang, Yu, & Zhang, 2022; Khan, Ahmad, Khalid, Ali, & Lee, 2023). As both datasets provide distinct train, test and development datasets, the datasets will be used as training and development sets to train and tune the classifier and test the performance of the classifier on the test set. In addition, the classifier trained by the SemEval-2013 data will be used to evaluate on the SemEval-2014 dataset because only test set is provided by the SemEval-2014. Besides, following their method, neutral comments will be filtered out and only take into consideration positive and negative comments. As for the MR dataset, train test split was applied to randomly split the dataset to train, development, and test sets with ratios of 0.8, 0.1, 0.1, following prior work by Zhao et al. and Zhang et al. (Zhao, et al., 2018; Zhang, Wang, & Zhang, 2021). Besides, the SemEval 2018 and News Headlines Sarcasm dataset will used as input for the secondary task of the proposed method which is sarcasm detection. In experiment 4, the main focus is on the performance of the sentiment analysis task, so the

output of the secondary task will be disregarded, and no changes will be made to the predictions made by the primary task. However, this setup allows us to investigate the impact of shared data between the two tasks in the proposed method. The statistical distributions of the sentiment analysis datasets are shown in Table 4.5:

4.5.2 Benchmarked Methods

- **RNN-Capsule:** Wang et al. (Zhao, et al., 2018) proposed a novel model using RNN-based capsule model to improve sentiment analysis accuracy.
- **IWV:** Rezaeinia et al. (Rezaeinia, Rahmani, Ghodsi, & Veisi, 2019) proposed enhanced word vectors along with a Convolutional Neural Network (CNN) model for sentiment classification.
- **BiLSTM-CRF:** Chen et al. (Chen, Xu, He, & Wang, 2017) proposed a neural network sequence model using BiLSTM-CRF to obtain target expression in opinionated sentences and divides the sentences into three categories based on the amount of target. Then, further trained the sentiment classifier using 1d-CNNs and the three categories of sentences.
- **SAT:** Huang et al. (Huang, Jin, & Rao, 2020) proposed a two-stage training strategy to train the BERT model for sentiment analysis.
- **SAAN:** Lei et al. (Lei, Yang, & Yang, 2018) proposed a sentiment-aware multi-head attention CNN-based model for sentiment analysis.

Table 4.5: Distribution of Datasets Used in Experiment 4

Datasets	Train	Dev	Test
SST-2	6920 (72%)	872 (9%)	1821 (19%)
MR	8600 (80%)	1000 (10%)	1000 (10%)
SemEval-2013	9684 (64%)	1654 (11%)	3813 (25%)
SemEval-2014	-	-	1853

- **AC-BiLSTM:** Liu et al. (Liu & Guo, 2019) proposed a BiLSTM model with attention mechanism and convolution layer to perform text classification.
- **SAMF-BiLSTM:** Li et al (Li, Qi, Tang, & Yu, 2020) proposed a BiLSTM model with multi-channel features and self-attention for sentiment analysis.
- **ATTPolling:** Usama et al. (Usama, et al., 2020) proposed a hybrid deep learning model based on RNN and CNN-based attention for sentiment analysis.
- **MVA:** Zhang et al. (Zhang, Wang, & Zhang, 2021) proposed a model that learn a sentence context representation from multiple perspectives via Multiview attention model for sentiment analysis.
- **SAWE:** Naderalvojoud et al. (Naderalvojoud & Sezer, 2020) proposed sentiment-aware word embeddings using refinement and senti-contextualized learning approach for sentiment analysis.
- **ABCDM:** Basiri et al. (Basiri, Nemati, Abdar, Cambria, & Acharya, 2021) suggested bidirectional layer that consist of BiLSTMs and BiGRU to extract global features from text documents. In addition, the attention mechanisms followed by CNN and pooling layer are utilized to identify sentiment orientation of review documents.
- **RCNNGWE:** Onan et al. (Onan, 2022) proposed a bidirectional and convolutional recurrent neural network architecture with a group-wise enhancement mechanism for sentiment analysis.
- **CoSE:** Wang et al. (Wang, Zhang, Yu, & Zhang, 2022) proposed a two layers GRU language model to construct a contextual sentiment embedding for sentiment analysis.
- **SCA-HDNN:** Khan et al. (Khan, Ahmad, Khalid, Ali, & Lee, 2023) proposed a model that leverages sentiment knowledge and employs pre-trained BERT model

incorporating BiLSTM, attention mechanism and CNN layer for sentiment classification.

4.5.3 Results and Discussion

In Experiment 4 tested the proposed method against various state-of-the-art baseline models on multiple datasets, including SST-2, MR, SemEval-2013, and SemEval-2014. The results, presented in Table 4.6, demonstrate the performance of each model on these datasets.

Upon examining the results, it becomes evident that the proposed method outperformed all other models on several datasets. Notably, it achieved the highest F1-score on SST-2 with an impressive 92.5%, surpassing all other methods. Similarly, it attained the highest accuracy on the Airline Twitter dataset with an impressive 95.2%. Additionally, the proposed method excelled on the MR and SemEval-2013 datasets, achieving F1-scores of 90% and 91.2%, respectively. Finally, it obtained the highest F1-score of 85.2% on the SemEval-2014 dataset.

Another remarkable method in Table 4.6 is SCA-HDNN, which also achieved impressive results on multiple datasets. SCA-HDNN obtained the highest F1-score on both the MR and SemEval-2013 datasets, achieving 93.5%. SCA-HDNN leverages a combination of attention-enabled deep neural networks, such as Bi-LSTM and CNN, to capture sequential dependencies in text data. Additionally, it benefits from sentiment-enhanced word embeddings from BERT, enabling the model to better understand sentiment information in text data. However, it is important to note that the proposed method still outperformed SCA-HDNN and other baseline models on most of the datasets, namely SST-2, Twitter Airline, and SemEval-2014.

Table 4.6: Experiment Results for Experiment 4. "-" denotes the method does not use the dataset to test and the best results are bolded.

Baselines Model	Dataset				
	SST-2	MR	Airline Twitter	SemEval-2013	SemEval-2014
RNN-Capsule	-	83.8	-	-	-
IWV	-	82.0	91.4	-	-
BiLSTM-CRF	88.3	82.3	-	-	-
SAT	79.4	79.4	-	-	-
SAAN	-	84.3	-	-	-
AC-BiLSTM	88.3	83.2	90.6	-	-
SAMF-BiLSTM	83.3	83.3	-	-	-
ATTPooling	83.6	83.6	91.2	-	-
MVA	88.0	88.0	-	-	-
SAWE	88.5	-	93.6	-	-
ABCDM	89.6	84.7	94.0	93.6	-
RCNNGWE	92.1	85.0	-	-	-
CoSE	89.8	87.7	-	90.8	82.8
SCA-HDNN	91.9	**93.5**	-	**93.5**	84.5
Proposed Method	**92.5**	90	**95.2**	91.2	**85.2**

The remarkable performances of the proposed method can be attributed to several key factors. Firstly, the method leverages the power of BERT, a pre-trained language model known for its ability to comprehend contextual information in text. This contextual understanding allows the model to capture intricate relationships between words and sentences, contributing to its high accuracy.

Furthermore, the multi-task learning approach employed by the proposed method plays a crucial role. By simultaneously training sentiment analysis and sarcasm detection, the model benefits from shared representations, promoting generalization and regularization, ultimately reducing overfitting. This sharing of representations enables the model to learn and capture common linguistic features and contextual cues that are relevant to both sentiment analysis and sarcasm detection tasks. By jointly optimizing the model for multiple tasks, the shared representations allow for the extraction of robust and

transferable features that generalize well across different datasets and tasks. Moreover, the sharing of representations acts as a form of regularization, preventing the model from overfitting to specific task characteristics and enhancing its ability to generalize to unseen examples.

However, it is important to acknowledge that the proposed method's excellent performance on the multi-domain datasets may not fully illustrate its value. In most real-life settings, sarcasm is often used to convey messages. Therefore, in subsequent experiments, the aim is to analyze the proposed method by including sarcasm detection on sentiment datasets that contain instances of sarcasm.

4.6 Experiment 5: Evaluation of the Impact of Sarcasm Detection on Sentiment Analysis

4.6.1 Experiment Settings

The Experiment 5 will compare the performance of the proposed method with and without sarcasm detection on a carefully chosen sentiment classification dataset that includes sarcastic contexts. This dataset containing sarcasm will be added to the datasets experimented on in Experiment 4, namely SST-2, MR, SemEval-2013, and SemEval-2014.

Besides, the Experiment 5 will be using the test dataset from SemEval-2014 Task 9 (Rosenthal S. , Ritter, Nakov, & Stoyanov, 2014). Among the test sets provided by SemEval-2014 Task 9, there is a specific test set that contains sentences with sarcasm. The author annotated the sarcasm data following a similar approach as presented in Table 4.3, where sarcastic sentences are labelled as negative.

The annotation method employed for the selected dataset aligns with the relationships defined in Table 4.3, making it an ideal dataset for examining the impact of sarcasm on

sentiment analysis. By utilizing this dataset, valuable insights can be gained into how the proposed method effectively analyses and classifies sentiments in the presence of sarcasm.

Similar to Experiment 4, the training and development sets will be used for training and tuning the model. The model will then be evaluated on the test set, which includes the SemEval 2014 test dataset with sarcasm. Additionally, the secondary task, which focuses specifically on sarcasm detection, will be trained and tuned using the SemEval-2018 sarcasm and News Headlines Sarcasm datasets.

As in Experiment 4 the neutral-labelled data will be excluded from this evaluation. By carefully considering the dataset selection, annotation method, and test scenario, experiment 5 aims to provide valuable insights into the proposed method's ability to effectively analyse and classify sentiments, particularly in the presence of sarcasm. Additionally, by comparing and contrasting the performance of the proposed method without and with sarcasm detection, the experiment seeks to analyse the impact of sarcasm detection on sentiment analysis.

4.6.2 Results and Discussion

This section provides an assessment of the impact of sarcasm detection on the performance of the proposed sentiment analysis method, using F1-score as the evaluation metric. Figure 4.6 presents a comparison of the proposed method's performance with and without the inclusion of the secondary task, namely sarcasm detection.

The results of the study were compelling. The sentiment analysis method equipped with sarcasm detection consistently demonstrated superiority over the version without this capability. This superiority was evident across all datasets, highlighting the

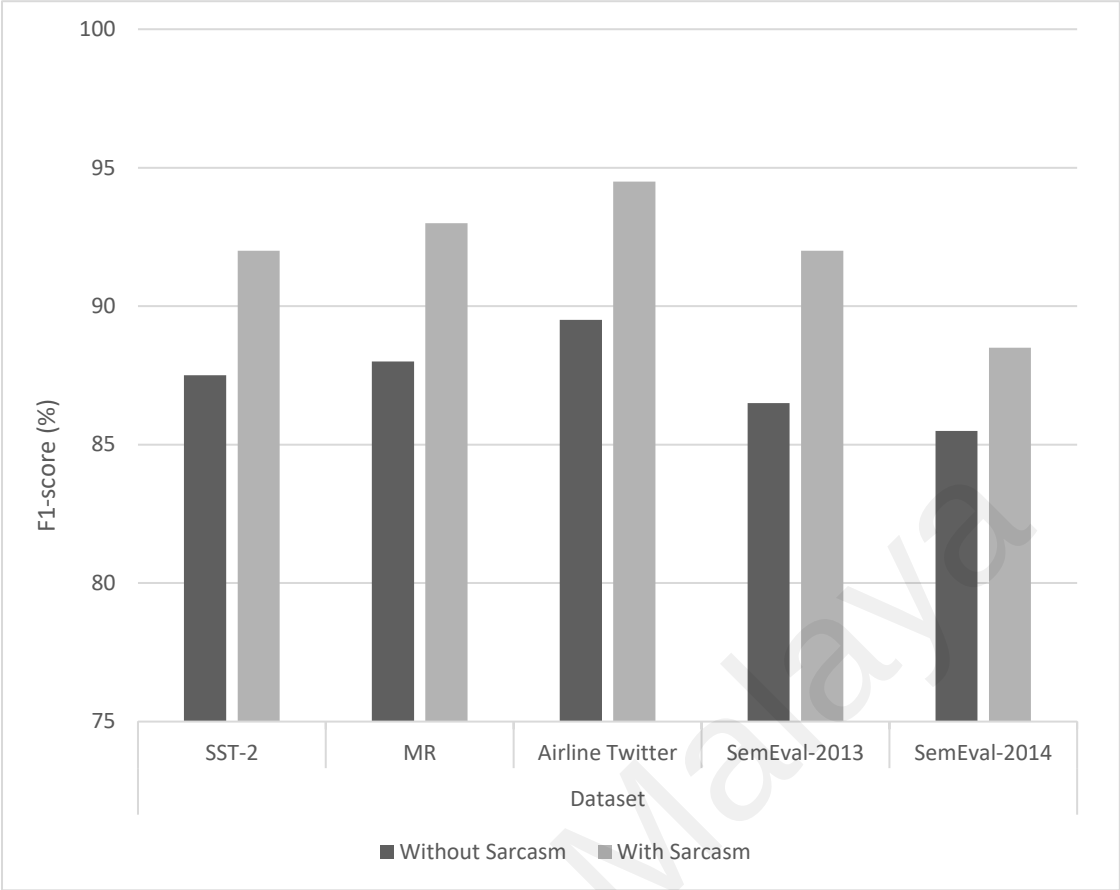


Figure 4.6: Overall Performance Achieved by the Proposed Method with and without Sarcasm Detection on Experiment 5

significance of incorporating sarcasm detection in sentiment analysis systems, particularly in real-world applications.

The influence of sarcasm detection on F1-score was indeed noteworthy. The model with sarcasm detection achieved performance scores ranging from 2.5 to 6.5 percentage points higher compared to the model without sarcasm detection. This improvement signifies the model's enhanced ability to understand nuanced sentiments and effectively handle the complexities introduced by sarcastic expressions.

The annotation method follows the relationship defined in Table 4.3, wherein positive and sarcastic sentiments are treated as negative. Thus, the experiment proceeded to assess the performance of both variants of the proposed method in terms of label-specific metrics using Table 4.7 and Table 4.8. Specifically, the model’s performance was analyzed for positive and negative sentiment labels.

For positive sentiment labels, the model with sarcasm detection exhibited higher precision, which was achieved through a reduction in false positives. This improvement directly resulted from the model's enhanced capability to recognize and distinguish sarcastic expressions, leading to fewer instances of falsely classifying sarcastic statements as genuine positive sentiments. Consequently, the model achieved more accurate classification of positive sentiments.

Conversely, for negative sentiment labels, the model with sarcasm detection demonstrated higher recall, stemming from a decrease in false negatives. The model's improved sarcasm recognition contributed to a reduced misclassification of sarcastic expressions as negative sentiments, resulting in an increased number of correctly identified negative sentiments. This heightened recall further boosted the model's performance in accurately identifying negative sentiments in the presence of sarcasm.

To further investigate, from Table 4.8, it can be observed that that the F1-score in negative is still lower than the F1-score of positive class, which contribute to a lower in total F1-score. Although the sarcasm recognition successfully improves the performance of the proposed method, there is still some misclassification in detecting sarcasm which result in this lower F1 score in negative as compared to the positive class. As the outcomes suggest, an effective sarcasm detection model directly contributes to enhancing sentiment analysis performance, particularly in real-world applications. While a performance increase is evident when comparing the two models, there remains potential for improving

Table 4.7: Performance Scores of Positive and Negative Labels for the Proposed Method without Sarcasm Detection. P = Precision, R = Recall, F = F1-score

Labels	SST-2			MR			SemEval-2013			SemEval-2014			Twitter US Airline		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Positive	85	90	88	94	93	93	93	92	93	91	88	89	94	95	94
Negative	95	81	87	86	80	83	85	76	80	83	81	82	90	80	85

the sarcasm detection model in future work. Possibly, we can incorporate another task to the multi-task learning framework which is speech or tone recognition to properly understand sarcastic context and improve the sentiment analysis performance.

These findings underscore the significance of sarcasm detection in sentiment analysis tasks, particularly in scenarios where sarcasm is prevalent in the text data. By integrating sarcasm detection, the sentiment analysis model becomes more proficient at understanding nuanced sentiments and effectively distinguishing between genuine expressions and sarcastic statements. Overall, the study highlights the critical role of sarcasm detection in sentiment analysis models, emphasizing its potential to enhance the robustness and effectiveness of sentiment analysis in real-world language data. The implications of our research are particularly relevant in real-world applications where sentiment analysis plays a pivotal role, such as social media monitoring, customer feedback analysis, and brand reputation management. In such contexts, the ability to correctly interpret sarcastic sentiments can significantly impact the accuracy and reliability of sentiment analysis results.

4.7 Model Analysis

The subsequent section will delve into a comprehensive analysis of the proposed model, focusing on its complexity, training time, and its application using multitask learning.

Table 4.8: Performance Scores of Positive and Negative Labels for the Proposed Method with Sarcasm Detection. P = Precision, R = Recall, F = F1-score

Labels	SST-2			MR			SemEval-2013			SemEval-2014			Twitter US Airline		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Positive	89	93	91	97	93	95	96	92	94	93	88	90	97	95	96
Negative	95	91	93	86	97	91	85	96	90	83	92	87	90	97	93

The proposed model, with its Transformer-based architecture, exhibits a slightly higher level of complexity compared to conventional deep neural networks, particularly due to the inherent complexities of the Transformer model, which can be analyzed in terms of big-O notation. The training time complexity of the proposed model can be expressed as $O(E \times T \times P)$, where E is the number of epochs, T is the number of training samples, and P is the number of parameters in the Transformer-based architecture.

However, in comparison to a similar classification work handling two tasks by Yunitasari et al., the proposed model stands out for its simplicity, thanks to the application of a multi-task learning approach. Using the big-O notation specified above, Yunitasari et al. work exhibit twice the time complexity as compared to the proposed model due to the two explicitly framework used, as shown in Figure 4.7.

In terms of training time, the proposed model is compared with the model presented by Yunitasari et al., as illustrated in Figure 4.7, using the same hyperparameters employed in the proposed method. The model repetitively uses the same input and hidden layer settings to train the sentiment and sarcasm models, leading to redundant processes and requiring training to be done twice. Conversely, with the multi-task learning approach, both the sentiment and sarcasm classifiers can be trained in a single epoch, resulting in a significant reduction in training time. Specifically, the proposed method takes 84 minutes per epoch for training, whereas training both sentiment and sarcasm classifiers explicitly as in Yunitasari et al. work would require 168 minutes per epoch.

Furthermore, in terms of computation power, the proposed method requires fewer computational resources compared to Yunitasari et al. work due to the smaller size of the neural network. Our goal in designing the neural network was to shrink its size to eliminate unnecessary processes and mathematical operations. As a result, data is passed

through the hidden layer only once, as opposed to twice in Yunitasari et al. work, significantly reducing unwanted processes and computation power consumption.

Overall, the model analysis demonstrates that the proposed method, incorporating multitask learning and sarcasm detection, presents a more efficient and streamlined approach to sentiment analysis. Despite the slight increase in complexity compared to conventional deep neural networks, the benefits of multi-task learning in terms of training time reduction and computational efficiency are evident. By optimizing the neural network size and adopting the Transformer-based architecture, the proposed method successfully harnessed the advantages of sarcasm detection in sentiment analysis while maintaining efficiency and computational effectiveness.

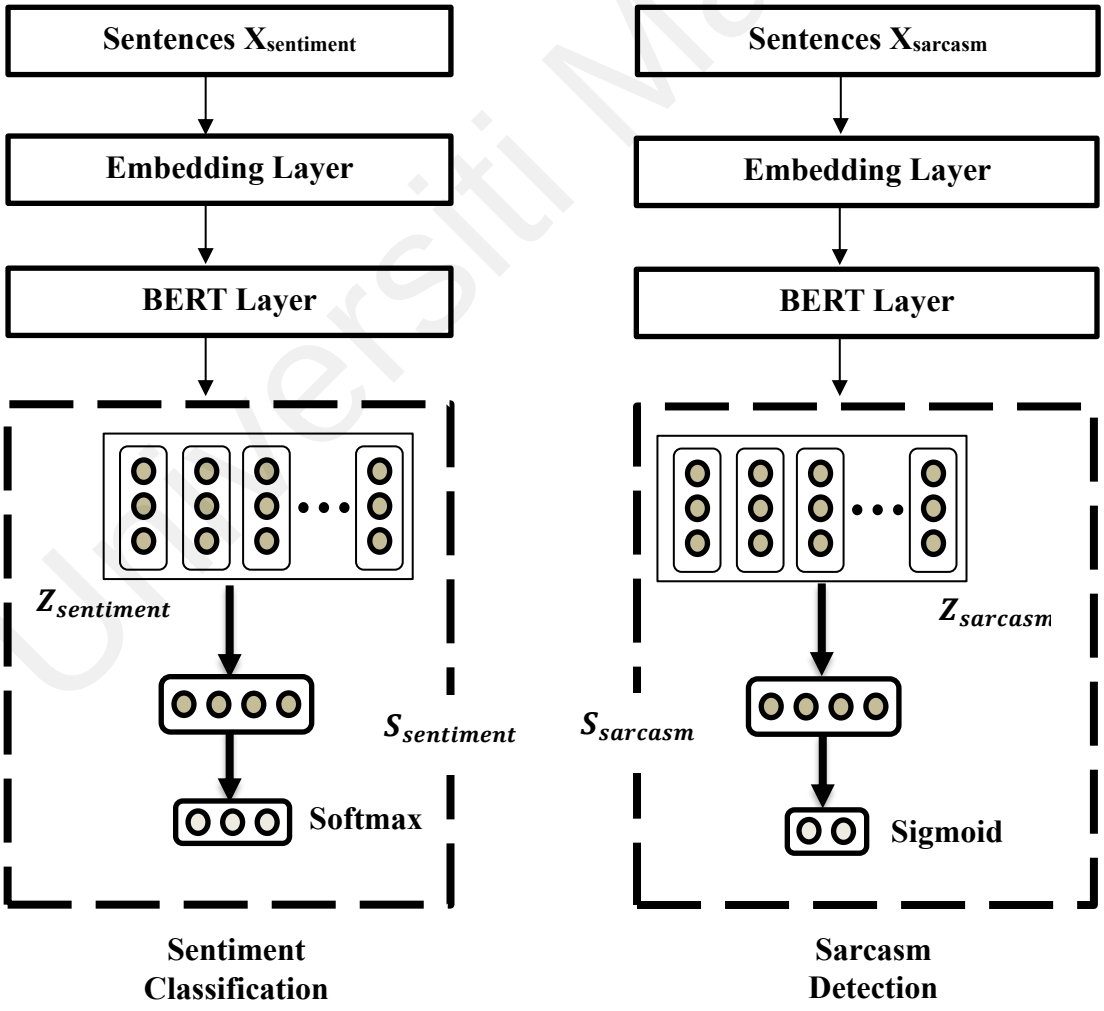


Figure 4.7: Architecture of Two Explicitly Sentiment and Sarcasm models

CHAPTER 5: CONCLUSION AND FUTURE WORK

5.1 Conclusion

Sentiment analysis plays a crucial role in extracting valuable insights from the vast data available on the internet, particularly on social media platforms. However, accurately analyzing sentiment is often hindered by the presence of sarcasm in sentences. To overcome this challenge, a multi-task learning framework that simultaneously performs sentiment analysis and sarcasm detection is proposed in this thesis.

The proposed framework leverages the power of the cutting-edge transformer-based BERT model, known for its ability to capture complex language structures and nuances. By incorporating sarcasm detection into the sentiment analysis task, the proposed method aims to enhance the accuracy of sentiment classification. The utilization of BERT allows us to improve the performance of the classification task and obtain more precise sentiment analysis results.

To evaluate the effectiveness of the approach, five experiments using different scenarios were conducted. The results of experiment 2 revealed limitations in the MTL-Bi-LSTM model, particularly in cases involving unseen datasets and neutral sentiments, where the removal of neutral words during pre-processing resulted in lower recall and F1 scores. To address this issue, the BERT model was selected, which demonstrated an immediate improvement in the method.

Consequently, the subsequent findings reveal that the proposed method consistently achieve competitive performance on multi-domain datasets, achieving F1-scores ranging from 90% to 91.2%, outperforming other state-of-the-art methods on various multi-domain datasets in experiment 4. Additionally, experiment 5 demonstrated that the inclusion of an efficient sarcasm detection model improves sentiment analysis

performance by 2.5 to 6.5 percentage points, especially on datasets containing sarcasm sentences that closely mimic real-life scenarios.

In conclusion, the proposed multi-task learning framework, combined with the BERT model, shows promising results in sentiment analysis and sarcasm detection. By effectively addressing the challenges posed by sarcasm, the method aims to provide more accurate sentiment analysis in real-world applications. This research contributes valuable insights to the field and lays the foundation for future advancements in sentiment analysis systems. Furthermore, the reduction in model complexity and computational costs associated with the multi-task learning approach makes it a practical and cost-effective solution. The combination of improved performance and reduced complexity makes the framework a promising approach for sentiment analysis tasks in various domains.

5.2 Future work

Experiment 5 underscores the fact that an efficient sarcasm detection model significantly enhances sentiment analysis performance. In light of this, for future work, we aim to design a multimodal sarcasm detection model, incorporating both speech and text inputs to analyze user tone to further enhance sarcasm recognition capabilities. This way, the model able to differentiate between positive and negative tone which tremendously help in deducing sentiment with underlying sarcastic context. The multi-task learning framework proves especially advantage here, as the addition of another tone recognition task incurs minimal model cost and complexity. This approach yields a model that accurately identifies sarcasm and adeptly classifies sentiment within sarcastic contexts.

REFERENCES

"Sarcasm". (n.d.). Retrieved from Cambridge Dictionary: https://dictionary.cambridge.org/dictionary/english/sarcasm

A Andrade-Segarra1, D., & A. Leon-Paredes, G. (2021). Deep Learning-based Natural Language Processing Methods Comparison for Presumptive Detection of Cyberbullying in Social Networks. *International Journal of Advanced Computer Science and Applications,, 12*(5), 796-803.

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classificaition in web forums. *ACM Transactions on Information Systems (TOIS)*, 12.

Abdelgwad, M. M., Soliman, T. H., & Taloba, A. I. (2022). Arabic aspect based sentiment classification using BERT. *Journal of Big Data*.

Akhand, M., Roy, A., Dhar, A. C., & Kamal, M. A. (2021). Recent Progress, Emerging Techniques, and Future Research Prospects of Bangla Machine Translation: A Systematic Review. *International Journal of Advanced Computer Science and Applications, 12*(9).

Arunachalam, R., & Sarkar, S. (2013). *The new eye of government: Citizen sentiment analysis in social media.* Nagoya, Japan: Proceedings of the IJCNLP 2013 workshop on natural language processing for social media (SocialNLP).

Basiri, M. E., Nemati, S., Abdar, M., Cambria, E., & Acharya, U. R. (2021). ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis. *Future Generation Computer Systems, 115*, 279-294.

Batra, H., Punn, N. S., Sonbhadra, S. K., & Agarwal, S. (2021). Bert-based sentiment analysis: A software engineering perspective. *Database and Expert Systems Applications: 32nd International Conference, DEXA 2021, Virtual Event, September 27--30, 2021, Proceedings, Part I 32.*

Bespalov, D., Bai, B., Qi, Y., & Shokoufandeh, A. (2011). Sentiment classification based on supervised latent n-gram analysis. *Proceedings of the 20th ACM international conference on Information and knowledge managemnt, Glasgow, UK.*, 373-382.

Chaithanya, K. A. (2020). *Twitter and Reddit Sentimental analysis Dataset*. Retrieved from Kaggle: https://www.kaggle.com/cosmos98/twitter-and-reddit-sentimental-analysis-dataset

Chen, T., Xu, R., He, Y., & Wang, X. (2017). Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Systems with Applications, 72*, 221-230.

Crawshaw, M. (2020). Multi-task learning with deep neural networks: A survey. *ArXiv, abs/2009.09796*, 43.

Devlin, Jacob, Ming-Wei, C., Kenton, L., Kristina, & Toutanova. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).*

Geetha, M., & Renuka, D. K. (2021). Improving the performance of aspect based sentiment analysis using fine-tuned Bert Base Uncased model. *International Journal of Intelligent Networks, 2*, 64-69.

H. V Phan, T., & Do, P. (2020). BERT+vnKG: Using Deep Learning and Knowledge. *International Journal of Advanced Computer Science and Applications, 11*(7).

Hongchan, L., Yu, M., Zishuai, M., & Haodong, Z. (2021). Weibo text sentiment analysis based on bert and deep learning. *Applied Sciences, 11*(22), 10774.

Huang, H., Jin, Y., & Rao, R. (2020). Sentiment-aware transformer using joint training. *IEEE*.

Karimi, A., Rossi, L., & Prati, A. (2020). Improving bert performance for aspect-based sentiment analysis. *arXiv preprint arXiv:2010.11731*.

Khan, J., Ahmad, N., Khalid, S., Ali, F., & Lee, Y. (2023). Sentiment and Context-Aware Hybrid DNN With Attention for Text Sentiment Classification. *IEEE Access, 11*, 28162-28179.

Lei, Z., Yang, Y., & Yang, M. (2018). SAAN: A sentiment-aware attention network for sentiment analysis. *Association for Computing Machinery.* New York, USA.

Li, W., Qi, F., Tang, M., & Yu, Z. (2020). Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification. *Neurocomputing, 387*, 63-77.

Liu, G., & Guo, J. (2019). Bidirectional LSTM with attention mechanism and convolutional layer for text classification. *Neurocomputing, 337*, 325-338.

Maas, A. L., Daly, R. E., Pham. P. T, Huang, D., Y., N. A., & Potts, C. (2011). Learning word vectors for sentiment analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142-150.

Mahdaouy, A. E., Mekki, A. E., Essefar, K., Mamoun, N. E., Berrada, I., & Khoumsi, A. (2021). Deep multi-task model for sarcasm detection and sentiment analysis in Arabic language. *arXiv preprint arXiv:2106.12488*.

Matsumoto, S., Takamura, H., & Okumura, M. (2005). Sentiment classification using word sub-sequences and Dependency Sub-trees. *Pacific-Asia conference on knowledge discovery and data mining*, (pp. 301-311). Berlin, Heidelberg.

Misra, R. (2019). *News Headlines Dataset For Sarcasm Detection*. Retrieved from Kaggle: https://www.kaggle.com/rmisra/news-headlines-dataset-for-sarcasm-detection

Naderalvojoud, B., & Sezer, E. A. (2020). Sentiment aware word embeddings using refinement and senti-contextualized learning approach. *Neurocomputing, 405*, 149-160.

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., & Wilson, T. (2013). SemEval-2013 Task 2: Sentiment Analysis in Twitter. *Association for Computational Linguistics.* Atlanta, Georgia, USA.

Nguyen, Quoc, D., Vu, T., & Nguyen, A. T. (2020, 11 16-20). BERTweet: A pre-trained language model for English Tweets. *In Proceedings of EMNLP 2020: System Demonstrations*, pp. 9-14.

Onan, A. (2022). Bidirectional convolutional recurrent neural network architecture with group-wise enhancement mechanism for text sentiment classification. *Journal of King Saud University-Computer and Information Sciences, 34*(5), 2098-2117.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of the 42nd*

annual meeting on Association for Computational Linguistics, Barcelona, Spain, 271-279.

Pang, B., & Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *arXiv preprint cs/0506075*.

Pang, G., Lu, K., Zhu, X., He, J., Mo, Z., Peng, Z., & Pu, B. (2021). Aspect-level sentiment analysis approach via BERT and aspect feature location model. *Wireless communications and mobile computing, 2021*, 1-13.

Podium. (2017). *Online Reviews Stats & Insights.* Podium.

Pota, M., Ventura, M., Catelli, R., & Esposito, M. (2020). An effective BERT-based pipeline for Twitter sentiment analysis: A case study in Italian. *Sensors, 31*(1), 133.

Rezaeinia, S. M., Rahmani, R., Ghodsi, A., & Veisi, H. (2019). Sentiment analysis based on improved pre-trained word embeddings. *Expert Systems with Applications, 117*, 139-147.

Rosenthal, S., Ritter, A., Nakov, P., & Stoyanov, V. (2014). SemEval-2014 Task 9: Sentiment Analysis in Twitter. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)* (pp. 73-80). Dublin, Ireland: Association for Computational Linguistics.

Rosenthal, S., Ritter, A., Nakov, R., & Stoyanov, V. (2014). SemEval-2014 Task 9: Sentiment Analysis in Twitter. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval2014)* (pp. 73-80). Dublin: Association for Computational Linguistics.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality. *Proceedings of the 2013 conference on empirical methods in natural language processing.*

Statista. (n.d.). *Global digital population as of April 2022*. Retrieved 7 27, 2022, from https://www.statista.com/statistics/617136/digital-population-worldwide/

Usama, M., Ahmad, B., Song, E., Hossain, M. S., Alrashoud, M., & Muhammad, G. (2020). Attention-based sentiment analysis using convolutional and recurrent neural network. *Future Generation Computer Systems, 113*, 571-578.

Vaswani, A., Shazeer, N., Pamar, N., Uszkoreit, J., Jones, L., N. Gomez, A., . . . Polosukhin, I. (2017). Attention is All you Need. *Advances in neural information processing systems, 30*, 30.

Wan, Y., & Gao, Q. (2015). An ensemble sentiment classification system of twitter data for airline services analysis. *2015 IEEE international conference on data mining workshop (ICDMW).*

Wang, J., Zhang, Y., Yu, L.-C., & Zhang, X. (2022). Contextual sentiment embeddings via bi-directional GRU language model. *Knowledge-Based Systems, 235*, 107663.

Xu, H., Shu, L., Yu, P. S., & Liu, B. (2020, 12). Understanding pre-trained bert for aspect-based sentiment analysis. *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 244-250). Barcelona: International Committee on Computational Linguistics.

Yafoz, A., & Mouhoub, M. (2021). *Sentiment Analysis in Arabic Social Media Using Deep Learning Models.* 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC).

Yanagimoto, H., Shimada, M., & Yoshimura, A. (2013). Document similarity estimation for sentiment analysis using neural network. *estimation for sentiment analysis using neural network, 2013 IEEE/ACIS12th Int. Conf. Comput. Inf. Sci*.

Yousif, A., Niu, Z., Chambua, J., & Younas, K. Z. (2019). Multi-task learning model based on recurrent convolutional neural networks for citation sentiment and purpose classification. *Neurocomputing*, 195-205.

Yunitasari, Y., Musdholifah, A., & Sari, A. K. (2019). Sarcasm Detection For Sentiment Analysis in Indonesian Tweets. *Indonesian Journal of Computing and Cybernetics Systems*, 53-62.

Zhang, Y., Wang, J., & Zhang, X. (2021). Learning sentiment sentence representation with multiview attention model. *Information Sciences, 571*, 459-474.

Zhao, J., Liu, K., & Xu, L. (2016). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions.* MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info.

Zhao, W., Ye, J., Yang, M., Lei, Z., Zhang, S., & Zhao, Z. (2018). Investigating capsule networks with dynamic routing for text classification. *arXiv preprint arXiv:1804.00538*.