Sentiment Analysis in Political Text

Kolawole Adeniyi Adeniran

Submitted to the
Institute of Graduate Studies and Research
in partial fulfilment of the requirements for the degree of

Master of Technology
in
Information Technology

Eastern Mediterranean University
September 2023
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Ali Hakan Ulusoy
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Technology in Information Technology.

Asst. Prof. Dr. Ece Çelik
Director, School of Computing and Technology

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Master of Technology in Information Technology.

Prof. Dr. Ahmet Rizaner
Co-Supervisor

Prof. Dr. Ali Hakan Ulusoy
Supervisor

Examining Committee

1. Prof. Dr. Ali Hakan Ulusoy
2. Assoc. Prof. Dr. Kamil Yurtkan
3. Asst. Prof. Dr. Hüsnü Bayramoğlu

ABSTRACT

Political discourse in the digital age takes many forms, including news articles, social media posts, political speeches, and public opinion polls. Understanding popular sentiments, analyzing political trends, and evaluating the emotional landscape surrounding political problems all need a careful examination of the attitudes conveyed in various textual sources. However, the intrinsic complexities of language, contextual complexities, and the nuanced interaction of multiple emotions provide severe hurdles to sentiment analysis in political works. Traditional sentiment analysis approaches frequently fail to capture these subtle sentiments properly. The core research question is concerned with the adaption and optimization of transformer models to provide precise sentiment analysis across a wide range of political text forms.

Unlike many domains of NLP, the political arena contains nuances that are not easy to identify by machines, thus this study aims to use three transformer models, BERT, RoBERTa and GPT-3 to investigate which one of the models is best to encapsulate the complex scenario of the political arena based on the NewsMTSC dataset, which is curated from two distinct datasets, POLUSA and BiasFlipper. The results show that the best performing of these models is RoBERTa with an average accuracy of 84.97% and F1-score of 86.98%. While BERT scores are closer to RoBERTa with average accuracy and F1-score of 81.72% and 81.09% respectively, GPT-3 had the worst performance with an average and F1-score of 79.40% and 79.51%.

These results show that with advancements in technology and machine learning, accurately classifying data such as in politics will not always give the best results due to the complexity of human emotions in such cases.

Keywords: BERT, RoBERTa, GPT-3, Natural Language Processing, Fine-Tuning, Text Mining, NewsMTC Dataset, Political Discourse, Transformer-Based Models, Sentiment Analysis.

ÖZ

Dijital çağda siyasi söylem, haber makaleleri, sosyal medya gönderileri, siyasi konuşmalar ve kamuoyu yoklamaları da dahil olmak üzere birçok farklı biçimde ortaya çıkıyor. Popüler duyguları anlama, siyasi eğilimleri analiz etme ve siyasi sorunları çevreleyen duygusal manzarayı değerlendirme, çeşitli metin kaynaklarında iletilen tutumların dikkatli bir şekilde incelenmesini gerektirir. Ancak dilin karmaşıklığı, bağlamsal incelikler ve çeşitli duyguların karmaşık etkileşimi, siyasi eserlerde duygu analizi konusunda ciddi engeller oluşturuyor. Geleneksel duygu analizi yaklaşımları, bu ince duyguları etkili bir şekilde yakalamada sıklıkla başarısız oluyor. Temel araştırma sorusu, bu dönüştürücü modellerin siyasi metin türlerinin geniş yelpazesinde hassas duygu analizi sunmak için nasıl uyarlandığı ve optimize edildiği ile ilgilidir.

NLP'nin birçok alanının aksine, siyasi saha, makineler tarafından tanımlanması kolay olmayan incelikleri içerir; bu nedenle bu çalışma, POLUSA ve BiasFlipper adlı iki farklı veri kümesinden derlenen NewsMTSC veri kümesine dayalı olarak siyasi sahanın karmaşık senaryosunu hangi modelin en iyi şekilde özetlediğini araştırmak için BERT, RoBERTa ve GPT-3 olmak üzere üç dönüştürücü model kullanmayı amaçlar. Sonuçlar, bu modellerin en iyi performans göstereninin RoBERTa olduğunu gösteriyor; ortalama doğruluk oranı %84.97 ve F1 puanı %86.98. BERT, RoBERTa'ya daha yakın skorlar elde ederken ortalama doğruluk ve F1 puanları sırasıyla %81.72 ve %81.09'dur. GPT-3 ise ortalama ve F1 puanı %79.40 ve %79.51 ile en düşük performansa sahiptir.

Bu sonuçlar, teknoloji ve makine öğrenmesindeki ilerlemelerle, siyaset gibi verilerin doğru bir şekilde sınıflandırılmasının, insan duygularının karmaşıklığı nedeniyle her zaman en iyi sonuçları vermeyeceğini göstermektedir.

Anahtar Kelimeler: BERT, RoBERTa, GPT-3, Doğal Dil İşleme, İnce Ayar, Metin Madenciliği, NewsMTSC Veri Seti, Siyasi Tartışmalar, Dönüştürücü Tabanlı Modeller, Duygu Analizi.

ACKNOWLEDGMENT

I would want to offer my heartfelt gratitude and appreciation to everyone who helped and contributed to the accomplishment of this thesis.

First and foremost, I would like to express my gratitude to my supervisors, Prof. Dr. Ali Hakan Ulusoy and Prof. Dr. Ahmet Rizaner, for their invaluable guidance, constant support, and insightful input and feedback during the research process. Your expertise and mentorship have been invaluable in shaping this thesis.

To my parents Prof. Solomon A. Adeniran and Mrs. Patience O. Adeniran, for their unending encouragement and support through my studies, could not have wished for better parents than you, you rock and to my siblings, Kike, Kemi, and Ayo, your words of encouragement are invaluable, and hopefully, I will make you all proud someday.

TABLE OF CONTENTS

ABSTRACT iii
ÖZ v
ACKNOWLEDGMENT vii
LIST OF TABLES x
LIST OF FIGURES xi
LIST OF ABBREVIATIONS xii
1 INTRODUCTION 1
1.1 Statement of Research Problem 2
1.2 Objective of Research 2
1.3 Research Question 3
1.4 Definitions of Key Terms 4
1.5 Assumptions 4
1.6 Delimitations and Limitations 5
1.6.1 Delimitations 5
1.6.2 Limitations 5
1.7 Importance of the Study 6
1.8 Thesis Organization 6
2 LITERATURE REVIEW 8
2.1 Text Mining 9
2.2 Text Classification 13
2.3 Sentiment Analysis 16
2.3.1 Sentiment Analysis in Politics 18
2.4 Transformer-Based Language Models 19

3 METHODOLOGY 22
3.1 Data Collection 22
3.2 Data Splitting 24
3.3 Data Pre-processing 24
3.4 Model Selection and Training 25
3.4.1 BERT 28
3.4.2 RoBERTa 30
3.4.3 GPT-3 31
3.5 Evaluation 32
4 RESULTS AND ANALYSIS 34
4.1 Baseline 34
4.2 Training of Dataset 34
4.3 Results 37
5 CONCLUSION AND FUTURE WORK 43
5.1 Conclusion 43
5.2 Future Works 44
REFERENCES 45

LIST OF TABLES

Table 1: NewsMTSC data distribution.. 24

Table 2: Data testing results 37

Table 3: Comparison of best averages of the models between my result and baseline results. 39

LIST OF FIGURES

Figure 1: Transformer model general architecture.. 26

Figure 2: BERT architecture at training layer.. 29

Figure 3: GPT-3 architecture.. 32

Figure 4: Bar chart showing average model performances....................................... 40

Figure 5: Confusion matrix for best performing RoBERTa...................................... 41

Figure 6: Confusion matrix showing for best performing GPT-3.............................. 42

LIST OF ABBREVIATIONS

BERT	Bidirectional Encoder Representations for Transformers
CNN	Convolutional Neural Network
GNN	Graph Neural Network
GPT-3	Generative Pre-trained Transformer 3
LSTM	Long Short-Term Memory
MLM	Masked Language Model
NLP	Natural Language Processing
NSP	Next Sentence Prediction
RNN	Recurrent Neural Network
RoBERTA	Robustly Optimized BERT Pretraining Approach
TSC	Target-dependent Sentiment Classification

Chapter 1

INTRODUCTION

The digital age of the modern political platform consists of opinions communicated through online texts on social media or news sites. Sentiment analysis, a subclass of Natural Language Processing (NLP), helped facilitate separating real political articles from fake political articles, and the ability to categorize articles based on the bias of the writer. This thesis is based on the process of categorizing the sentiments in political texts from established news sites.

As the popularity of text classification increases, researchers have explored additional strategies to increase the accuracy and efficiency of classification. Machine learning breakthrough have opened new avenues for autonomous text classification. Various machine learning approaches for text classification have been researched in different forms. For example, researchers [1] used text classification in conjunction with text mining and natural language processes to evaluate incidents.

The majority of the works that have been done on sentiment analysis in the political discourse are entirely based on texts gathered from social networking sites especially Twitter as well as Facebook. Target-dependent Sentiment Classification (TSC) or aspect-term sentiment classification and other similar tasks such as aspect-based sentiment classification and stance detection have mainly focused on data in which the author(s) explicitly expressed their opinions [2].

1.1 Statement of Research Problem

Political articles are rich and diverse in the digital era, ranging from news articles and social media posts to political speeches and public opinion polls. Understanding popular sentiments, political trends and the broader emotional landscape surrounding political matters requires analyzing the feelings portrayed in these texts. However, due to the intricacy of language, contextual nuances, and the interplay of various emotions, sentiment analysis in political writings presents significant obstacles. Traditional sentiment analysis approaches frequently fail to capture these nuances of sentiment in such situations. As a result, this study aims to examine and construct an effective and robust sentiment analysis system based on transformer-based language models.

The primary research question is how transformer-based models can be adapted and optimized to accurately analyze sentiments in various forms of political texts, hence, providing valuable insights into public opinion and improving our understanding of political discourse in modern societies.

1.2 Objective of Research

The primary goal of this study is to compare various architectures of transformer-based language models of natural language processing based on their accuracies in successfully categorizing the sentiments in political texts. This includes a comparative analysis of transformer-based models, which includes conducting a deep study of various types of transformer-based models such as BERT, RoBERTa and GPT-3, to investigate which of the three architectures is best suited for sentiment analysis in political texts.

Another objective is fine-tuning the models to get both the best performance and accuracy from each model. This includes fine-tuning the pre-trained transformer models for allowing the sentiment analysis to adapt to political domain. This is because of domain-specific problems that arise in the sentiment analysis of political texts. These problems include political prejudice, sarcasm, irony, and subjectivity, which are all common in the political arena.

The final goal is to properly test the proposed sentiment analysis model on the dataset and quantify each model's performance by using conventional assessment criterial such as accuracy, precision, recall, and F1-score. This is crucial for determining the effectiveness of the architectures and finding potential areas for improvement.

1.3 Research Question

To perform this research titled sentiment classification in political texts, this section defines the precise research questions that will guide the study, offering an outline for future testing and evaluations and insights that will contribute to the overall goals.

a. What is the difference between sentiment analysis in political texts compared with other domains?
b. Which transformer-based language model architecture is more suitable for sentiment analysis in political texts?
c. How can common occurrences in daily languages such as sarcasm and irony and human factors such as biases affect the result of sentiment analysis in political texts?

d. How can the sentiment analysis model be made more accessible in areas such as political arenas to increase user trust and ensure transparency?

1.4 Definitions of Key Terms

Bidirectional Encoder Representation from Transformers (BERT) is built on the transformer architecture, which is a type of neural network architecture optimized for sequential data processing, such as text. The major innovation of BERT is its bidirectional training approach, which allows the model to learn word representations by considering the contexts of the phrases surrounding a word [3].

Robustly Optimized BERT Pre-training Approach (RoBERTa) is a type of BERT, which was launched in 2019. It is derived from BERT and utilizes the same training approach, but with modifications to improve performance. RoBERTa makes use of larger batch sizes, longer training cycles, and more diversified data, resulting in enhanced language understanding capabilities. It provides cutting-edge outcomes on variety of NLP tasks by fine-tuning its contextualized representations after pre-training [4].

Generative Pre-trained Transformer (GPT-3) is a deep neural network decoder-only model based on the transformer architecture. GPT-3, in contrast to prior architectures that relied on recurrence and convolution, utilizes attention mechanisms [5]. These attention mechanisms allow the model to concentrate on specific sections of the input texts that it considers to be the most important for its predictions [6].

1.5 Assumptions

As the training of this thesis is based on a third-party dataset, NewsMTSC, which will be discussed in Chapter 3, the data contained in the dataset is assumed to be adequate for this topic and the data are annotated accordingly with no partial judgement. Also, transformer-based models such as BERT, RoBERTa and GPT-3 are more suitable for this sentiment analysis as they capture the complex language patterns and contextual nuances relevant to political communications.

1.6 Delimitations and Limitations

1.6.1 Delimitations

This study is solely focused on English-based news articles and to be precise, they are based on the political scenes of the United States of America. This means texts in other languages are not considered in this thesis. Also, the dataset NewsMTSC only contains news from news articles and reputable news outlets. This implies other potential news sources such as social media posts or political speeches, will not be considered in this study.

The final delimitation of this thesis will be the processing of the dataset. Three transformer models, GPT-3, RoBERTa and BERT will be implemented for this analysis and other traditional sentiment analysis models will not be explored.

1.6.2 Limitations

The sentiment analysis model's findings and conclusion may have limited generalizability beyond the specific political texts in the NewsMTSC dataset. Linguistic patterns and sentiment expressions may differ depending on the political setting or dataset. The performance of the sentiment analysis model may be influenced by data bias in the NewsMTSc dataset, which may reflect the thoughts and opinions of specific news sources or locations. Political texts frequently contain sophisticated vocabulary, sarcasm, irony, and ambiguity, which can present obstacles for sentiment analysis models and result in inferior performance. Also, the implementation and training of transformer-based models can be computationally intensive, imposing restrictions on the model and dataset size.

Finally, due to the limited computing power available at the time of the testing, there were limits to the tunning of the training parameters. These include using an epoch value of 3 because higher epoch numbers lead to memory overflow with time, increasing the epoch values can lead to better training time and better results. The same issue goes for the learning rate, which was set to 2e-5, 3e-5 and 5e-5. From the results, it is seen that the results values become better with the decrease in learning rate, but once I go down below 2e-5, the system becomes unstable, and I could not try lower learning rates.

1.7 Importance of the Study

Analyzing news agencies' moods in political texts can shed light on potential biases and subjectivity in their reporting. Understanding the sentiment expressed by various news sources can aid in determining their objective and reliability in political coverage. This study can investigate how news outlets' emotional portrayals of political events and people affect public perception. It can highlight the influence of the media in shaping public perceptions of specific political issues and individuals. This research can also provide a comparative analysis of how various media organizations shape political narratives and objectives by evaluating sentiment across different news channels. This can aid in the identification of patterns and variances in sentiment representations.

1.8 Thesis Organization

The thesis organization is as follows, Chapter 2 will be about the literatures review of the works that have been done in the field of NLP and transform models. Chapter 3 will discuss the methodology that will be used to perform the analysis, it will also contain details about the dataset being used and the transformer models selected for the experiment. Chapter 4 will be about discussing the results and finally, chapter 5 will be the conclusion, which. will includes the summary of my findings.

Chapter 2

LITERATURE REVIEW

Natural Language Processing (NLP) has been around for decades, with early work dating back to the 1980s [7]. Until 2005, traditional techniques such as fully connected methods will be typically outperformed by shallower architectures that relied on feature engineering. However, [8] introduced a new approach for pre-training deep neural networks and machine learning, which marked a significant turning point in the field. This innovation enabled the use of deeper architectures which has led to big leaps in the improvements of performances of NLP tasks.

To capture text representations, a variety of neural network architectures, including Convolutional Neural Networks (CNNs) [9] and Recurrent Neural Networks (RNNs) [10], are being utilized considering recent advances in deep learning. Graph Neural Networks (GNNs) have recently gained attention as a promising new approach in the field of NLP [11]. GNNs originally appeared in 2009 [12] and has since been used for a range of NLP tasks including relational reasoning [13], text classifications [14], neural machine translation [15] and sequence labelling [16]. One specific variant of GNNs, Graph Convolutional Neural Networks (GCNNs) was first employed by Defferrard [14] in a text classification task and it was shown to outperform traditional CNN models.

2.1 Text Mining

The field of data mining has seen significant advancements in recent years, thanks in large part to the advancements of both software and hardware technologies. This has resulted in the accessibility of a vast range of data types. It is especially true for text data, as the rapid generation of large amounts of text content by a wide range of individuals has been made possible by the development of web and social media platforms. The internet enables the easy storage and processing of this text data, but as the amount of text data grows, there is a need for new algorithms to effectively learn patterns from this data in a scalable manner [17].

Managing text data, which often lacks structure, can be a challenge. One solution is to utilize search engines, which allow for quick retrieval of information using keyword queries. Improving the effectiveness and reliability of search engines has garnered significant attention in the realm of information retrieval research [18]. Studies in this area also delve into related topics such as text clustering, categorization, summarization, and recommender systems [19][20].

Research in the field of information retrieval has primarily been geared towards providing users with easy access to information, rather than analyzing the data to uncover patterns and insights. This contrasts with text mining, which goes beyond information access to aid in data analysis and decision-making. Traditional information retrieval has not placed a strong emphasis on text processing or transformation, whereas text mining endeavours to uncover meaningful patterns and trends in the data, even when a query is not relevant or necessary [20].

Text mining techniques have been used by researchers in a variety of disciplines to identify latent information. There is a growing interest in using text mining to analyze project contracts and communication interactions in the field of construction projects. The capacity of text mining to allow efficient contract analysis, discover critical phrases, and reveal feelings hidden within project-related texts drives its inclusion into this environment. Previous research has recognized the benefits of text mining, demonstrating its ability to improve project comprehension, reduce contractual analysis efforts, and highlight subtle communication dynamics amongst project stakeholders. Moreover, the use of Building Information Modeling (BIM) as a data modelling architecture for text analysis has grown in popularity, allowing for the development of sophisticated analytical models [21]. This allows the ability of the participants in the project to visually examine construction project contracts and project correspondence.

Text mining in the context of tourism, according to their research [22], the authors explore diverse text representation methods and text-based NLP techniques for tasks like sentiment analysis and topic extraction. They also investigate how these techniques are used in areas such as market demand analysis and tourist profiling. Researchers are currently looking into ways to tailor the newly established BERT to biological datasets. Well cited articles in fields have showcased the expansive applications of data mining, underscoring its significance [23].

The research presented in [24] provides useful insights into previously unknow research shifts in data mining. Furthermore, it emphasizes how these shifts help to improve understanding the most influential scholarly works in the discipline of text analytics. Researchers in [25] intends examining the modern level of text mining research by assessing changes in published literature over the last few years. Their purpose is to provide significant insights to both experts and scholars on the latest developments, methodology, and practical implications in text-mining research. The goal of the study in [26] is to offer a comprehensive overview of data sources, computational techniques, and applications of text data within the context of the food business.

In regards the significant quantity of labelled training data necessary for effective NLP models, the application of deep learning in financial text analysis frequently faces challenges as a consequence of limited availability of labelled datasets withing the banking industry. Addressing these challenges, the study in [27] BERT for Financial Text Mining (FinBERT), a specialized language model pre-trained on extensive banking text data. Additionally, the investigation discussed in [28] conducts a methodical review of literature pertaining to the utilization of text mining in the context of service management.

Researchers in [29] believe search engines specifically assist in opening up various sources of information and making them usable for administrative and research reasons. Considering this, the goal of their study was to present an overview of text data mining techniques and successful data quality management for Research Information Systems (RIS) in the context of open data and open research in scientific institutions and libraries, as well as proposals for their application. The authors in [30] examined the performance of text categorization on English texts using various classifiers, an ensemble of classifiers, and a neural probabilistic representation model termed word2vec. The goal of [31] was to provide a novel methodology for improving the auto-indexing of Arabic texts. This essay worked empirically to isolate several technical issues, logical flaws, and conceptual flaws in multiple categories [32].

The study in [33] aimed to use text mining to evaluate construction project contracts and visually examine project correspondences. Researchers in [34] presented an innovative procedure for extracting relevant fundamental characteristics or principles for analysing causes of accidents from incident narratives contained in unstructured texts and categorical data using Association Rule Mining (ARM). In the study conducted by researchers [35], they examined text mining activities involving the complete publications and abstracts published by postgraduate students from various Nigerian institutions in their study. The studied took a survey-based method, with questionnaire serving as the data gathering instrument. The study included 357 postgraduate students who were chosen using the Raosoft sample size calculator.

The authors in [36] focused on the extraction of features from clinical routine data through the application of text mining techniques. A retrospective study was conducted to determine how far automated extraction could extract interpretable information from clinical works of literature. Extraction of hidden insights through vast amounts of language input is a critical act in assisting analyst in the development and improvement of Intelligent Virtual Assistants (IVAs) for customer service. Verint Intent Manager (VIM) is an analytical tool which mixes two algorithms, unsupervised and semi-supervised, to aid researchers discover and structure concerning users' intentions in texts [37]. The authors in [38] suggest potential direction to fill the holes between general text mining technologies and the requirements associated with working on historical and location-specific texts.

2.2 Text Classification

Text classification, a fundamental topic within NLP, holds extensive practical implications across various domains. Some notable applications are media filtering and the detection of junk mail (SPAM) [39][40]. Text representation learning is a critical component of text classification.

Text classification is a frequently studied subject in databases, data mining, and information retrieval. It is defined as assigning a class label to a given data record based on its attributes [41]. The data records in text classification are text documents and the class labels are discrete class values. The training data is employed to construct a model that establishes a connection between features extracted from text documents and corresponding class labels. This model is then utilized to predict class labels for text documents that have not been previously seen. This problem can be tackled in both hard and soft forms, with the hard version assigning a specific label to a text document and the soft form assigning a probability value to the text document for each class.

It is worth noting that while continuous values can also be used as labels, categorical values are mostly used. While the problem of categorizing records with set-valued attributes is closely related to text classification, text classification must consider the frequency of words and the larger domain size of text data as opposed to standard set-valued classification tasks [42].

The authors in [43] propose \emph{Text Revealer}, the first model inversion attack for text reconstruction against text categorization using transformers. Because of the variety of Chinese characters in glyphs, spammers frequently enclose spam content within visually similar text to mislead the model, while ensuring that the users grasp the context. The authors in [44] propose integrating the fundamental features of human cognition related to these malicious texts into spam text detection models. They propose achieving this by using pre-trained model that thoroughly learns the structural syntax of Chinese letters and numbers, producing specific meanings right from the foundation level. The model that was suggested takes vectors as input to a dual-channel neural network form, employing a few CNNs to extract N-Gram information from a variety of word windows and improve the local feature representation via the concatenation operation, and the Bidirectional Long Short-Term Memory (BiLSTM) network to extract semantic association information from the context to obtain the high-level feature representation at the sentence level [45].

Huang in [46] investigates the evaluation of an active learning query approach for text classification. The evaluation of active learning's probability-based query approach for text classification was demonstrated. Because such models are unfamiliar with the job at hand, they may experience instability and poor performance. The authors in [47] suggest a plug-and-play strategy for bridging the gap using a simple self-training method that solely necessitates the class names and an unlabelled dataset and does not require domain expertise or trial and error.

The authors in [48] introduce ContrastNet, a contrastive learning system designed to tackle challenges related to biased modelling and excessive training in few shot text classification scenarios. The authors in [49] investigate an alternate unsupervised approach that deduces ideological affiliation based on patterns of social media image propagation. General topic modelling methods, on the other hand, have the edge of evaluating documents that capture significant word structure without the necessity for pre-training [50].

Although capsule networks are useful for image classifications, their applicability to text domains remains unexplored. Researchers in [51] demonstrate that the adaptability of capsule networks to text classification and reveal advantages they offer in comparison to CNNs. The study in [52] investigates capsule networks for text classification using dynamic routing. Due to data scarcity, many classification models perform poorly on short texts. In response to this challenge, Zeng et al. [53] propose a solution called topic memory networks for short text-classification. This approach utilizes an innovative topic memory technique to encode latent topic representations that provide insights into class label. The study in [54] presents a convolutional recurrent neural network for text classification that combines the benefits of convolutional neural networks for extracting local characteristics from text with those of RNNs in memory for connecting the extracted features. For text classification, a graph convolutional network was proposed [55].

Despite their significance, deep models often overlook nuanced classification cues stemming from the intricate relationships between words and classes, as these models heavily depend on text-level representations. To address this concern, a solution was introduced in [56] in the form of an interaction technique. This technique aims to integrate word-level matching signals into text classification process, thus mitigating the issue. To address these issues, the study in [57] offers a new GNN-based model that constructs graphs for each input text while sharing global parameters, rather than a single graph for the entire corpus.

Researchers in [58] compare the logistic regression, random forest, and KNN models for text classification. However, a classification system was devised for categorizing BBC news articles. The investigation in [59] provides an exhaustive assessment of more than 150 text classification models based on deep learning, which have been constructed in recent times. This study delves into their technical advancements, shared features, and notable strengths.

2.3 Sentiment Analysis

Sentiment analysis involves the process of collecting and identifying the varying range of emotions express in texts to comprehend different points of views or judgments pertaining to specific subjects [60]. An entity can be any distinguishable or distinct item, such as humans, events, organizations, systems, or products [61]. The aim of sentiment analysis is to determine biases, as well as feeling or emotion, expressed in a text towards a specific entity. The primary purpose is to comprehend users' opinions, preferences, or evaluations of certain things using text analysis.

In recent years, sentiment analysis has grown in prominence in both research and industry. This is due to the vast amount of opinionated text that is generated by internet users daily [62]. This large amount of data provides an opportunity for researchers and businesses to gain insights into people's opinions, preferences, and evaluations of different subjects, products, or services. Sentiment analysis allows one to extract this information, classify it, and use it for decision-making or further analysis.

Research in [63] presents a Sentiment Analyser (SA) that extracts sentiments (or opinions) on a subject from text documents found online. The study in [64] employed manually tagged Twitter postings that have been rated for the sentiment. The goal of [65] is to address one of the core difficulties of sentiment analysis: sentiment polarity classification. As the field has expanded, many subareas have emerged, each targeting a particular level of analysis or research questions. The study in [66] concentrates on analysing sentiments on an aspect based level, to locate and consolidate sentiments related to subjects contained in the document, and also specific aspect in those entities. The authors in [67] define multimodal sentiment analysis as modelling intra- and inter-modality dynamics.

Ignoring higher-level abstractions could potentially undermine the acquisition of sentiment-related characteristics from text and further diminish the effectiveness in classifying of sentiments. As a response to these challenges, a new architecture for text sentiment classification called AEC-LSTM is presented, which tries to improve the Long Short-Term Memory (LSTM) network through incorporating Emotional Intelligence (EI) and an attention mechanism [68]. To solve the sentiment analysis challenge, the study in [69] provides a novel technique that comprises data collection, pre-processing, feature encoding, and classification using three long short-term memory variations. The authors in [70] identified bias and motivation classes contained in tweets about tourism in some cities in Thailand.

Research in [71] offers a multimodal sentiment knowledge-sharing framework (UniMSE) that integrates MSA and ERC tasks based on features, labels, and models. The study in [72] presents a sentiment analysis methodology in a multi-label context that follows Plutchik's wheel of emotions. Word2Vec, Glove, and bidirectional encoder representations from transformers (BERT), for example, build vectors by evaluating word distances, similarities, and occurrences while neglecting other factors such as word sentiment orientation. To address these constraints, the authors in [73] develop LeBERT, a sentiment classification model that combines sentiment lexicon, N-grams, BERT, and CNN. It employs the techniques of data mining procedures to analyse the subjective opinions present within a document or collection of documents [74].

2.3.1 Sentiment Analysis in Politics

Researchers in [75] present an approach based on topic modelling, specifically involving item identification and text sentiment analysis, to analyse social media posts. This method is then used to visualize the progression impact of COVID-19 pandemic. The study in [76] contends that the term “fake news” understates the Russian misinformation campaign. Along with its numerous advantages, the literature indicates that machine learning methods might inject human bias into the outcomes they derive from training data. The authors in [77] investigate sentiment analysis algorithms for human influence by modifying characteristics of humans by using an established template to see how it affect sentiments.

Chakraborty et al. [78] investigate the task of analysing a vast collection of internet news articles to determine the prevailing sentiments in the circumstances of the COVID-19 pandemic. They also do statistical research to look at the relationship within the actual impact of COVID-19 and the opinions conveyed through internet news. The study in [79] emphasizes the challenges that still need to be investigated to advance the understanding of subjective phrases. The authors in [80] make a contribution by scrutinizing the term patterns that were present as the headlines of Daily Star, a popular English tabloid published daily in Bangladesh.

While sentiment analysis can effectively encapsulate the general tone of a document, what is more important in the context of political analysis is the document's position on certain subjects. This refers to how the document defines a specific idea, entity, or society as positive or negative, reflecting the author's intrinsic political opinions. The authors in [81] dispute the usefulness of approximating the author's opinion in the study of texts associated with politics using sentiment scoring and they emphasize for a stronger focus on recognizing the theoretical difference between a bias and opinion in a document. The study in [82] tackles the job of Targeted Sentiment Analysis for news headlines issued by major media outlets during the 2019 Argentine Presidential Elections.

2.4 Transformer-Based Language Models

In terms of computational efficiency, the transformer design outperforms RNN-based models. BERT and GPT recently demonstrated effectiveness of transformer models on a variety of NLP tasks by utilizing pre-trained language models on large-scale datasets. These transformer systems perform poorly in terms of the language model itself. Neither self-attention nor the transformer's positional encoding can efficiently incorporate the word-level sequential context required for language modelling [83].

Researchers in [83] investigated optimal transformer architectures for language models, such as adding additional LSTM layers to better capture the sequential context while being computationally economical. The study in [84] investigates ctrl: a controllable generation conditional transformer language model. Although large-scale language models demonstrate potential text-generating capabilities, users cannot readily modify specific features of the generated text. A straightforward, efficient intra-layer model parallel technique to train transformer models with billions of parameters was implemented [85].

Researchers in [86] present a novel self-attention model (specifically multi-linear attention) using Block-term Tensor Decomposition (BTD) based on the notions of tensor decomposition and parameter sharing. To address the issue, the study in [87] presents a novel model known as textbfExplicit Sparse Transformer. The authors in [88] introduce an end-to-end neural transformer-based Spoken Language Understanding (SLU) model capable of predicting the variable-length domain, intent, and slots vectors inherent in an audio signal without the use of an intermediate token prediction architecture. In contrast, the study in [89] presents VD-BERT, as a simple yet effective unified vision-dialogue transformer system that leverages pre-trained BERT language models for visual dialogue tasks.

Research in [90] shows how to combine convolution neural networks with transformers to model the local as well as global dependencies of an audio sequence in a parameter-efficient manner. Embedding knowledge awareness in language model pre-training is achieved without modifying the transformer design, introducing explicit knowledge layers, or storing semantic information externally [91]. To demonstrate the generality of ideas, the study in [92] creates the Relationship-Sensitive Transformer Network (RSTNet) for image captioning using the two modules and the vanilla transformer architecture.

Researchers in [3] proposed a new language representation paradigm known as BERT, that can be fine-tuned with just one additional output layer to create cutting-edge models for a variety of tasks. Liu in [93] describes BERTSUM, a straightforward variation of BERT for extractive summarization. Simple BERT-based models for relation extraction and semantic role labelling are presented by Shi et al. [94]. Extensive experiments were undertaken in [95] to evaluate various BERT fine-tuning strategies on text classification tasks and present a generic solution for BERT fine-tuning. This finding suggests that BERT networks may capture structural information about language. This was supported by [96] conducting a series of studies to unpack the elements of English language structure taught by BERT. The authors in [97] offer Visual-Linguistic BERT (VL-BERT), a new pre-trained generic representation for visual-linguistic tasks. According to research in [98], Sentence-BERT (SBERT) is a pre-trained BERT network modification that uses siamese and triplet network architectures to create semantically relevant sentence embeddings that can be compared using cosine similarities. When reading a domain text, experts draw conclusions based on their relevant knowledge.

To achieve this capability, the study in [99] proposes a knowledge-enabled language representation model (K-BERT) using Knowledge Graphs (KGs), in which triples are injected as domain knowledge into phrases. Sun et al. [100] suggest MobileBERT as a method for compressing and speeding the widely used BERT model. To speed up inference and minimize model size while retaining accuracy, the study in [101] offers a unique transformer distillation method specifically developed for knowledge distillation of transformer-based models.

Chapter 3

METHODOLOGY

This thesis adopts an empirical research approach to investigate the complex terrain of sentiment within political discourse. Empirical research is distinguished by its emphasis on observable, measurable data to reach conclusions based on tangible outcomes. In this context, sentiment analysis is an effective method for uncovering the complex emotional components hidden in political texts. This chapter outlines the methodological framework that was performed to attain the research objectives.

3.1 Data Collection

The dataset employed in this thesis is open-source data called NewsMTSC: Sentiment in English News Articles. The included news articles were collected and annotated by researchers [102]. It includes more than eleven thousand hand-annotated news articles in USA.

The original sources of the data in the corpus were from two separate data sets, POLUSA [103] and Bias Flipper 2018 [104]. These datasets were gathered by researchers using web crawlers on news aggregator sites daily between January 2017 and August 2019. The POLUSA dataset contains nine hundred thousand news articles while the Bias Flipper contains news articles that are related to 2781 political events in the United States of America.

Articles are categorized as negative (-1), positive (1), or neutral (0) to indicate the political alignment of the average reader in the USA. A negative classification suggests that the article is left-leaning, while a positive classification suggests that it is right-leaning. A neutral classification indicates that no sentiment was expressed in the article.

The annotated dataset includes three different categories, -1, 0, and 1 with articles containing statements such as “Hours later, Spicer apologized, although he made a number of mistakes during his apology, including mispronouncing Syrian President Bashar Al-Assad’s name” as -1, which means the polarity of -1 is left-leaning, “Trump's ability to bring new white, working-class voters into the GOP fold allowed him to compete and win in more territory than Republicans are accustomed to (think Pennsylvania, Michigan and Wisconsin).” as 1 and “Hillary Clinton blamed the Democratic National Committee, Facebook, and conspiracy site Infowars Wednesday for her election defeat during an interview in which she pointed at a total of 18 alleged guilty parties for her big loss.” as 0. These excerpts show that the article annotated as -1 is left-leaning because it focuses on the minute errors made during the press conference of the then-press secretary, Sean Spicer (a Republican) rather than reporting on the real details of the presser. On the other hand, the excerpt annotated as 1 represents the opinion of the writer about then-president, Donald J. Trump, which shows the article as right-leaning and the article annotated as 0, reports directly what was said by the runner-up of the 2016 United States of America presidential election, with no bias or addition as centric.

3.2 Data Splitting

The dataset was partitioned into 3 sets as the training set, the verification set and the testing set. The data distribution is shown in Table 1.

Table 1: NewsMTSC data distribution

Set	**Total**	**Positive**	**Neutral**	**Negative**
Train	8739	2395	3028	3316
Validate	1476	246	784	482
Test	1146	361	587	624

NewsMTSC, which consists of 11361 articles, and is divided into 3 subsets, training, testing, and validation was merged into a single set because validation was not implemented in the models. This allows the consolidated dataset to be divided into a 70 to 30 ratio with 70% for the training and the remaining 30% for the testing of the models.

3.3 Data Pre-processing

The following data cleaning was made to make the dataset suitable for conducting the required training and testing of the models to be used.

- Lowercasing: To preserve uniformity in letter casing, all text data is transformed to lowercase, preventing the model from treating differently cased words as distinct.
- Tokenization: This divides the texts into distinct words or sub word units. Tokenization is a prerequisite for further analysis.

- Punctuation and special character removal: Punctuation marks and special characters, such as commas, periods, and symbols, that do not contribute to sentiment analysis are eliminated.
- Stop word removal: Articles, prepositions, and conjunctions are common stop words that are eliminated to reduce noise and increase computing performance.
- Stemming: This is used as a strategy for reducing an inflected word to its root. For instance, "information", "informative", "informing" can all be reduced to one single word "inform".
- Handling numeric and non-textual data: Any numerical figures, dates, or non-textual components found in the data are handled appropriately or eliminated because they do not contribute to the analysis.
- Handling HTML tags and URLs: If the data contains HTML tags or URLs, they are stripped or altered so that the analysis is simply focused on the textual content.

3.4 Model Selection and Training

Transformer-based language models will be used to conduct the sentiment analysis because of their exceptional ability to capture intricate patterns in language and contextual nuances. The transformer's main novelty is its "self-attention" mechanism, which allows the model to weigh the value of individual words in sequence in relation to each other. This approach allows the model to incorporate contextual links while avoiding the need for explicit sequential processing.

Three well-known models, BERT, RoBERTa and GPT-3, will be used, each with distinct capabilities that correspond with the complexities of political discourse. BERT excels at understanding context because of its bidirectional architecture, whereas RoBERTa focuses on training efficiency. GPT-3, known for its generative abilities, offers potential insight into the impact of sentiment on language generation.

These three selected models share the same basic architecture for the transformer model which is an encoder-decoder model. The transformer architecture is shown in Figure 1 below [105].

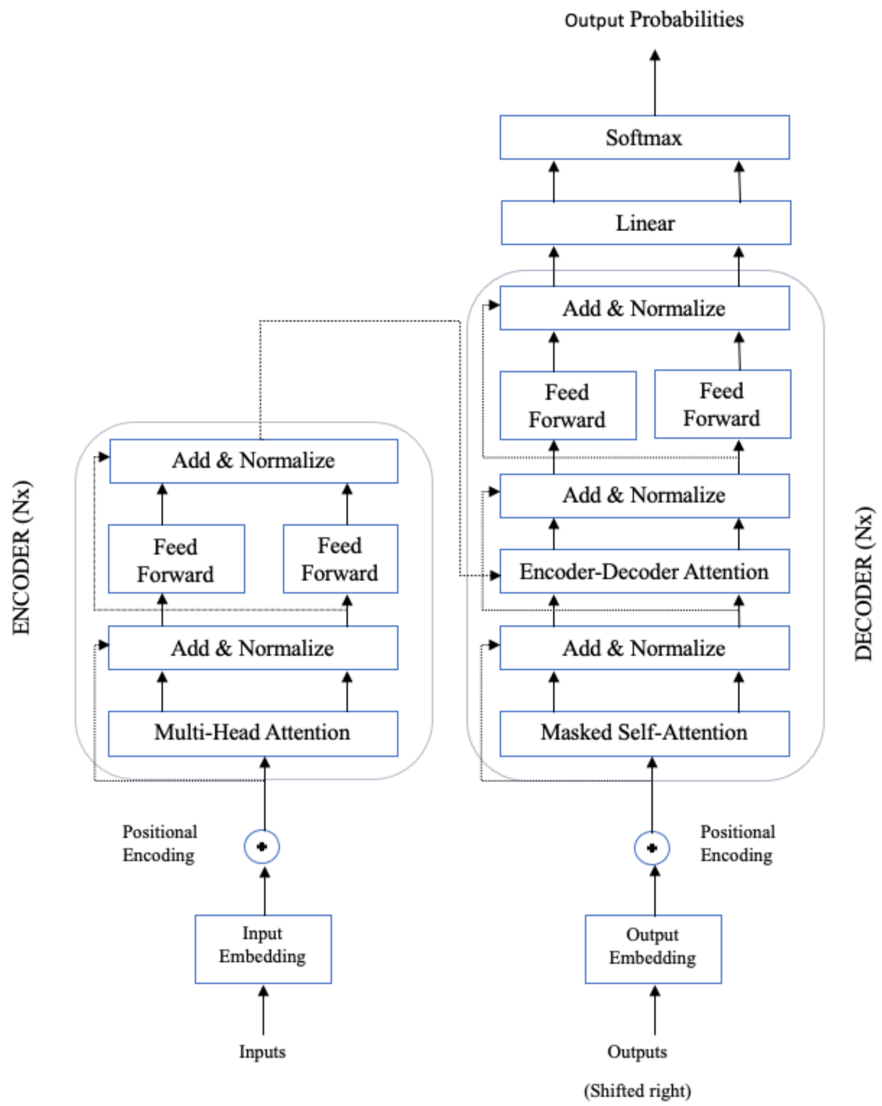

Figure 1: Transformer model general architecture

The transformer architecture can be broken down into different modules.

- Input embeddings: The input sequence is initially embedded in a vector space with dense vectors representing each word or token.
- Encoder and decoder stacks: The transformer model is made up of 2 parts: an encoder and a decoder. The encoder is responsible for processing the input sequence, while the decoder is responsible for producing the output sequence for a specific task such as classification.
- Self-attention mechanism: Transformer model uses self-attention to weigh the relevance of different words in a sequence based on their connections. It computes attention scores for each word in the sequence by comparing it to the other words in the sequence.
- Multi-head attention: Transformer model employs several “heads” of attention rather than a single attention mechanism. This allows the model to learn various correlations at various positions and scales.
- Positional encoding: Because the model does not grasp word order natively, positional encodings are appended to the input embeddings to offer information about word placements.
- Feedforward neural network: Transformer model uses feedforward neural networks after the attention layer for additional input processing. These networks are applied to each point in the sequence independently.
- Residual connections and layer normalization: The residual connections are utilized to mitigate the vanishing gradient problem while the layer normalizations are used to stabilize training. The residual connections and layer normalization are shown as “Add & Normalized” in Figure 1.

- Encoder-decoder attention: An additional layer of attention is provided in the decoder, allowing the model to focus on the input sequence while generating the output sequence.
- Masking: Masking is used during training to prevent the decoder from peeking at future positions, ensuring that predictions are exclusively based on previous positions. This is included in the self-attention layer of the decoder side.

3.4.1 BERT

BERT is an NLP model that was designed by Google researchers in 2018 to achieve high-level accuracy on many NLP applications. BERT's architecture is based on the transformer model which has proven more effective in NLP applications than traditional models [3]. BERT uses encoders in a transformer as a sub-structure to pre-train models from NLP tasks such as text classifications, text summarization and Stanford Q/A. Its execution of these tasks is divided into 2 phases which are the pre-training for language comprehension and fine-tuning for a specific activity [105].

BERT uses the encoder part of the transformer model but not the decoder part. Instead, BERT uses its innovative training layer. BERT was developed to understand languages by training on the Masked Language Model (MLM) and Next Sentence Prediction (NSP) methods. It uses MLM as a mask to learn bidirectional contexts in sentences. As a result, sentences are selected at random as input, masks some of the sentences, and then reconstructs the masked words from the preceding and succeeding texts to generate its output. It achieves NSP due to its capacity to input two sentences at once and detect whether the second sentence follows the first.

BERT is available in two sizes, the $\text{BERT}_{\text{Base}}$ and $\text{BERT}_{\text{Large}}$. $\text{BERT}_{\text{Base}}$ consists of 12-layered transformer encoder blocks, in which each block contains 12-head self-attention layers and 768 hidden levels with a total of 110 million trainable parameters, while the $BERT_{Large}$ consists of 24-layered transformer encoder blocks and each block contains 24-head self-attention-layers and 1024 hidden layers, producing an approximate value of 340 million trainable parameters. With these trainable parameters, $BERT_{Large}$ produces higher accuracy when compared to $BERT_{Base}$ but at the cost of high computing power. Due to this reason, $BERT_{Base}$ will be used for training due to the lack of an efficient computer to handle the resources required for $BERT_{Large}$.

BERT's basic architecture at the training layer is shown in Figure 2 [105].

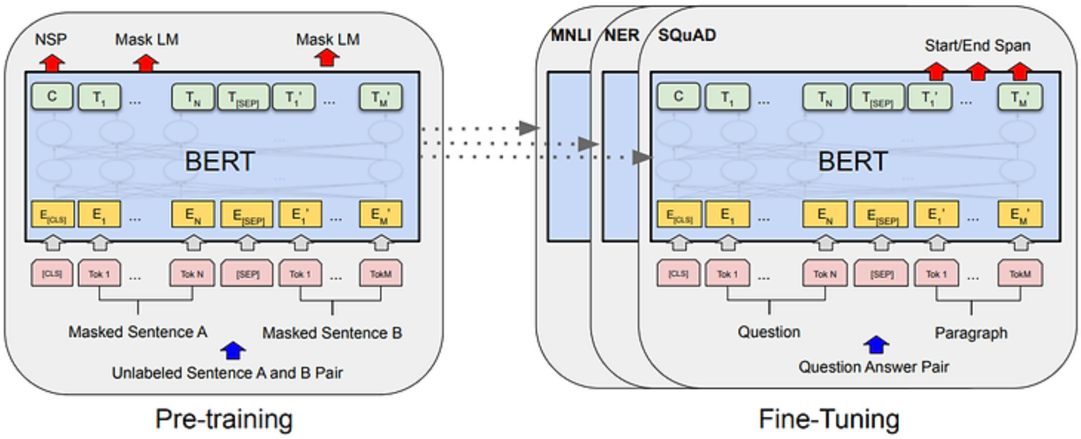

Figure 2: BERT architecture at training layer

The fine-tuning layer consists of 3 types which are Stanford Question and Dataset (SQuAD) v2.0 for question-and-answer tasks, Named Entity Recognition (NER) for entity identification and Multi-genre Natural Language Inference (MNLI) for classifications. The MNLI will be used for fine-tuning in this thesis.

MNLI is often used to fine-tune BERT because of its benchmark dataset for evaluating model's reasoning abilities and generalization across different fields.

Therefore, BERT fine-tuning with the MNLI dataset leads to a better classification performance when compared to training a specific dataset with the default BERT model.

3.4.2 RoBERTa

RoBERTa is a BERT model variant that is intended to improve the performance of pre-training language models for various NLP applications. RoBERTa introduces several modifications to the pre-training process of BERT which led to improved model performance. These modifications are:

- Larger batch sizes: RoBERTa uses larger batches of data for pre-training, facilitating improved hardware utilization and convergence.
- Dynamic masking: During pre-training, BERT masks out words at random to predict them. RoBERTa alters this procedure by employing a dynamic masking strategy that processes each sentence many times with various masked terms. This exposes the model to broader range of contexts for each word.
- More training data: When compared to BERT, RoBERTa is trained on more data over a longer period. It is pre-trained on 160 GB of text while BERT was trained on 16 GB to text.
- No next sentence prediction task: In the decoding part of BERT's architecture, BERT uses NSP to predict if two sentences in a text were connected. RoBERTa eliminates this process in favour of focusing entirely on the masked word prediction task.
- No sentence order prediction: The arrangement of sentences in a document was predicted by BERT but RoBERTa abandons this task, making the pre-training procedure easier.

- Hyperparameter tuning: To discover the best settings for model training, RoBERTa performs a more thorough search for hyperparameters.

Because of these changes, RoBERTa delivers a cutting-edge performance on a variety of benchmark NLP tasks, frequently surpassing BERT and other BERT variants. Its advancements are largely because it explores new ways to exploit more data, train longer, and successfully modify hyperparameters. RoBERTa shares the same architecture with BERT that is presented in Figure 2.

3.4.3 GPT-3

Open AI's GPT-3 is an advanced language processing AI model which is based on the transformer model. It is the third version of GPT series and is one of the most powerful and adaptable language models available. It is trained on large number of text data to produce coherent and contextually appropriate human-like text. Although GPT-3 is primarily intended for natural language generation tasks, it allows for modification to be used for other NLP tasks such as text classifications tasks.

GPT-3 contains 12 layers and each layers contains 768 hidden layers which allows for 125 million trainable parameters. The architecture of GPT-3 is shown in Figure 3 [106].

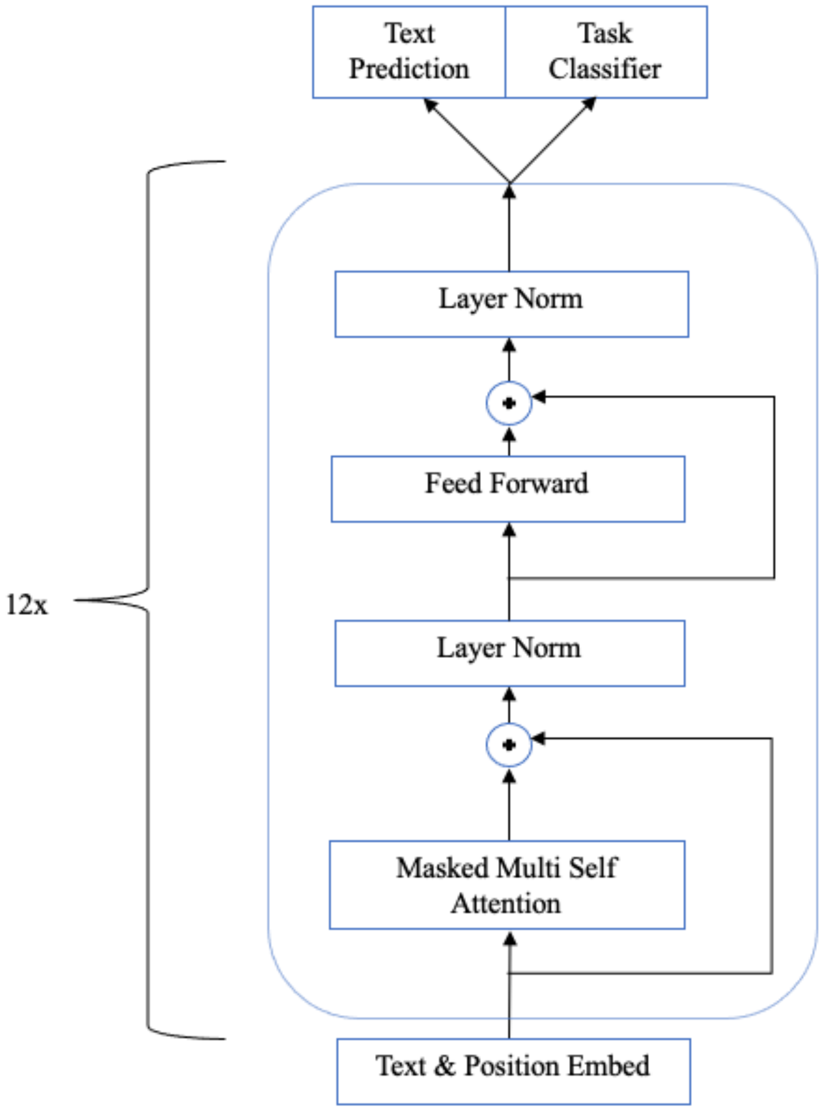

Figure 3: GPT-3 architecture

3.5 Evaluation

For the performance evaluation of the classification, accuracy, precision, recall and F1-score values will be used. These metrics are based on the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) of the testing result.

- Accuracy: This is the most straightforward metric, which is the ratio of TP to total instances in the testing dataset.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (1)$$

- Precision: Precision is the ratio of TP predictions to all the positive predictions (TP and FP)

  $$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (2)$$

- Recall: This is the ratio of TP to all actual positive cases (TP and FN). It assesses the model's capacity to catch all positive cases, which is especially important when FN values can be costly.

  $$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

- F1-score: This is the harmonic mean of precision and recall. It provides a balanced assessment of a model's performance, particularly when the testing dataset contains imbalanced classifications.

  $$\text{F1-score} = \frac{2*(\text{precision}*\text{recall})}{\text{precision}+\text{recall}} \quad (4)$$

Chapter 4

RESULTS AND ANALYSIS

This chapter focuses on data training and testing and the outcomes. Three different runs were made for each transformer model while varying the learning rate at each run. The detailed results will be discussed.

4.1 Baseline

The baseline for the result will be the study by researchers in [102]. The research consisted of 2 transformer models, BERT and RoBERTa with four different tokenization architectures which are Sentence Pair Classification BERT (SPC-BERT) [4], Target-Dependent sentiment classification with BERT (TD-BERT) [107], a Local Context Focus (LFC) mechanism for aspect-based sentiment classification [108] and Gated Recurrent Unit for (GRU) for TSC, which was proposed in [4]. These tokenization architectures were also applied to their RoBERTa test. Their best performances were SPC-BERT with an F1-score of 80.1% and GRU-RoBERTa with an F1-score of 83.1%.

4.2 Training of Dataset

For each transformer architecture that was used in this thesis (BERT, RoBERTa and GPT-3), three experiments were conducted on each with different learning rate values of 2e-5, 3e-5 and 5e-5. The tokenizer that was used for BERT was the $BERT_{Base}$ (BERT-base), which is the vanilla tokenizer for BERT and for RoBERTa was $RoBERT_{Base}$ (RoBERTa-base). GPT-3 on the other hand does not use a tokenizer because it is baked into the core program. The epoch value remained constant throughout the experiments and was set to 3. This is due to the constrained computing resources available at the time of the experiment. For the optimization, AdamW [109] was used. The batch size was set to 32 with dropout rate of 0.1. These values can be changed when fine-tuning the model but for the process of this training, only the learning rate was changed while other parameters remained constant.

The learning rate is a critical hyperparameter in training machine learning models with significant effects on the training process and model performance. Learning rate determines how fast a model converges, with a higher learning rate leading to faster convergence, and a lower learning rate leading to slower model convergence. The major drawback with a lower learning rate is the time it takes to converge, and it can lead to overfitting of the model whereas the higher learning rate can lead to poor results, overshoot the optimal solution or lead to divergence of the model. The learning rates used in this thesis from lowest to highest are 2e-5, 3e-5 and 5e-5. Lowering the learning rate to 1e-5 caused instability and crashes during the testing so I made 2e-5 the lowest.

The input length was set to 160 and batch size was set to 32 for the models. GPT-3 does not have a strict limit for input length whereas both BERT and RoBERTa have a maximum input length of 512 tokens, which means texts with a larger number of tokens will be truncated. To avoid this, the texts were segmented into 160 tokens each. Also, larger tokens might lead the models to lose context in longer texts, which can defeat the purpose of the training. Higher tokens also require more memory and computing resources, impacting both processing time and efficiency. As I have limited processing power, and the average length of texts in NewsMTSC dataset is

152 [102], 160 tokens input length was selected for the three models. The selected input length can lead to texts losing their contexts if they are longer than 160 tokens, or if important sentiment-bearing words or phrases are truncated due to the limited length.

The feature extractions in these models include tokenization, special tokens, padding and truncation, conversion to input tensors and label transformation.

- Tokenization involves breaking down text into smaller parts known as tokens, WordPiece tokenizer will be implemented for the models.
- Special tokens: These tokens are used to specify the beginning and ending of each sentence. The CLS token shows the beginning of a sentence, and the SEP token shows the ending of the sentence.
- Padding and Truncation: The selected models process inputs in batch sequences and they require all batches to have the same input length. To achieve this, sequences are padded if the length of the batch is less than the input length with special padding tokens and are truncated if the length of the batch is greater than the input length.
- Conversion to Input Tensors: The tokenized inputs are converted into numerical tensors. This includes input ids, which are integer identifiers associated with each token in an input sequence and attention masks, which are binary masks that indicate which of the tokens are part of the batch (1) and which of the tokens are paddings (0). This helps the models to focus on the real tokens alone.
- Label Transformation: The polarities of the datasets -1 (negative), 0 (neutral), and 1 (positive) are transformed to more suitable formats for the models' classification outputs, 0 for neutral, 1 for negative and 2 for positive.

Text summarization is not implemented in the systems. Text summarization is a challenging NLP task, and depending on the algorithms implemented, it can either have a positive or negative effect on the overall results. Text summarization helps with summarizing lengthy texts into a condensed version to reduce the data, which can lead to some contexts being lost.

4.3 Results

The NewsMTSC dataset was split into randomly 70% for training and 30% for testing after combining all the datasets (train, validate, and test) into one. The results of the runs are shown in Table 2.

Table 2: Data testing results

Model	**Run**	**Learning rate**	**Accuracy (%)**	**Precision (%)**	**F1-score (%)**
$\text{BERT}_{\text{Base}}$	1	2e-5	82.62	83.42	80.90
	2	3e-5	81.45	80.98	80.94
	3	5e-5	81.11	78.71	81.42
	Average		81.72	81.04	81.09
$\text{RoBERTa}_{\text{Base}}$	1	2e-5	**86.09**	**87.98**	**86.98**
	2	3e-5	84.64	87.85	84.90
	3	5e-5	84.17	87.19	83.50
	Average		**84.97**	**87.67**	**85.13**
GPT-3	1	2e-5	79.62	80.81	80.70
	2	3e-5	79.49	82.81	79.77
	3	5e-5	79.10	77.05	78.06
	Average		79.40	80.22	79.51

From Table 2, RoBERTa resulted in better performance on average than both BERT and GPT-3. This can be attributed to the fact that $\text{RoBERTa}_{\text{Base}}$ was pre-trained on news articles whereas $\text{BERT}_{\text{Base}}$ was pre-trained on Toronto BookCorpus [110] and English Wikipedia. When compared to the results from the study in [103], RoBERTa model performed higher than BERT in both, with $\text{RoBERTa}_{\text{Base,}}$ which is the original pre-trained RoBERTa model, performing better than all the other versions with an average accuracy score of 84.97% and average F1-score of 85.13% compared to average accuracy of 83.1% (GRU-RoBERTa) and average F1-score of 82.5% (GRU-RoBERTa).

From the results in Table 2, GPT-3 have the worst performance of the three transformer models with an average accuracy of 79.40% and an average F1-score of 79.51%. This can be attributed to the fact that GPT-3, though a transformer model, is designed purposefully for text generation and not for text classification.

From the results in Table 2, it is shown that the lowest learning rates (2e-5) gave better results in each of the three models used. This can be attributed to lower learning rates often leading to better generalization, as they allow the model to fine-tune its weights more carefully.

Table 3: Comparison of best averages of the models between my result and baseline results.

Result			Baseline Result		
Model	Accuracy (%)	F1-score (%)	Model	Accuracy (%)	F1-score (%)
BERTbase	81.72	81.09	GRU-BERT	81.1	80.2
RoBERTabase	84.97	85.13	RoBERTa-GRU	83.8	83.1

Table 3 above shows the comparisons between the average F1-scores and accuracies of Bertbase and RoBERTabase models used in this thesis with that of the baseline result [102]. In both cases, the RoBERTa models performed better than the BERT models with the F1-score of 83.1 and accuracy of 83.8 of the RoBERTa-GRU model of the baseline result compared to the F1-score of 80.2 and accuracy of 81.1 in the GRU-BERT model of the baseline result. This trend is also reflected in the result in Table 2 with RoBERTabase model having an F1-score and accuracy of 84.97 and 81.72 respectively and BERTbase model having an F1-score and accuracy of 81.72 and 81.09 respectively. This trend shows that of the two models, RoBERTa and BERT, irrespective of the model version, RoBERTa performs better than BERT specifically in the domain of political news classifications.

When the best average results from the result I got compared to the baseline result, RoBERTabase performs better than RoBERTa-GRU with an average accuracy of 84.97 and F1-score of 85.13 compared to the average accuracy of 83.8 and F1-score of 83.1 of the RoBERTa-GRU.

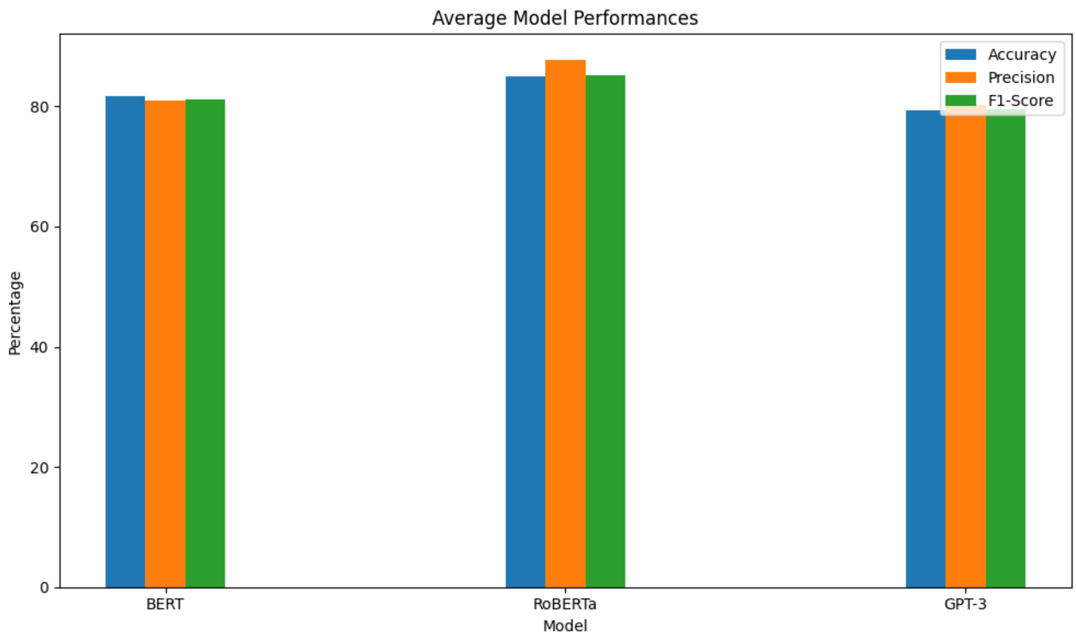

Figure 4: Bar chart showing average model performances.

Figure 4 shows the bar chart comparing the different metrics results for the average performances of each model tested. As evident in the chart, the average precision values for both RoBERTa and GPT-3 models are slightly higher than their respective average accuracies whereas BERT's average precision value is lower than its average precision value. This shows both RoBERTa and GPT-3 model have a higher rate of predicting the true positive classes i.e., negative as negative, positive as positive and neutral as neutral, rather than successfully classifying the sentiments but for BERT, it shows that it is better at classifying sentiments better than the positive classes only. In this context, the distinction between accuracy and precision becomes apparent. While each model can correctly classify classes, RoBERTa and GPT-3 have higher performance when classifying the positive values.

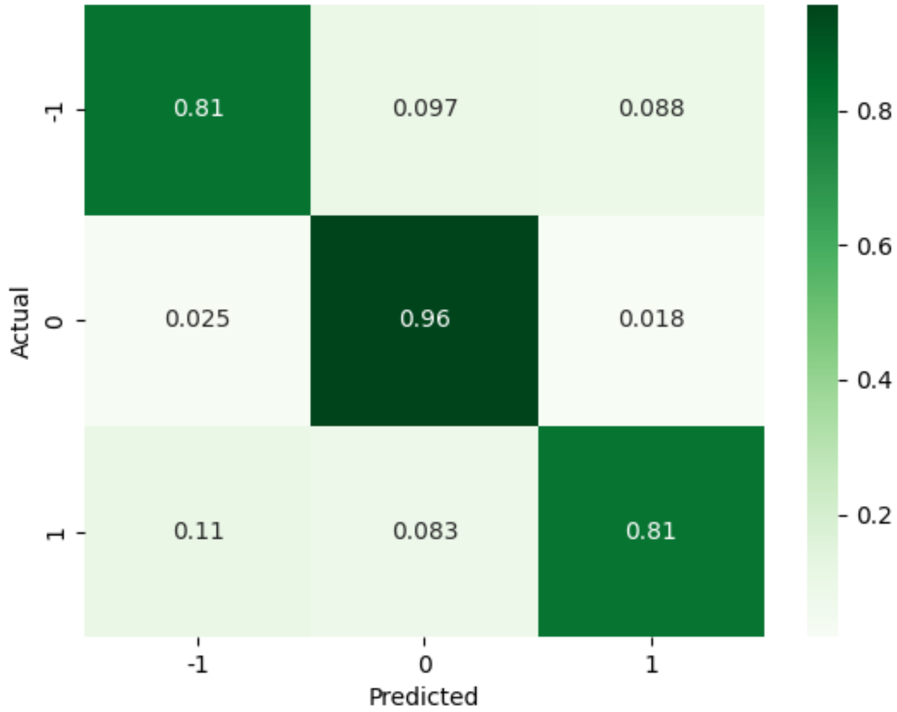

Figure 5: Confusion matrix for best performing RoBERTa

Figure 5 shows the confusion matrix for the best result from the RoBERTa model. It can be seen from the result that when it comes to predicting a no-sentiment article, RoBERTa has an accuracy value of 96%, but the shortcomings are from analysing the news with a sentiment (-1 and 1), classifying each of the sentiments correctly at 81%. From the heatmap, it can also be seen that most of the wrong classifications come from classifying sentiment news as non-sentiment (0), with 9.7% of the negatively annotated news sentiment classified as neutral and 8.3% of the positively annotated news also classified as neutral.

Figure 6, which shows the confusion matrix for the best-performing GPT-3 model, when compared to RoBERTa, GPT-3 classified neutral annotated news rightly with 97% accuracy, but it has difficulty classifying both the positive and negative annotated news correctly with accuracy of 67% for the negatively annotated and 65% for the positively annotated news. Also, GPT-3 misclassified 20% of the negative labels as neutral and 27% of the positive labels also as neutral. This can be alluded to the lack of classification supports for the GPT-3 model.

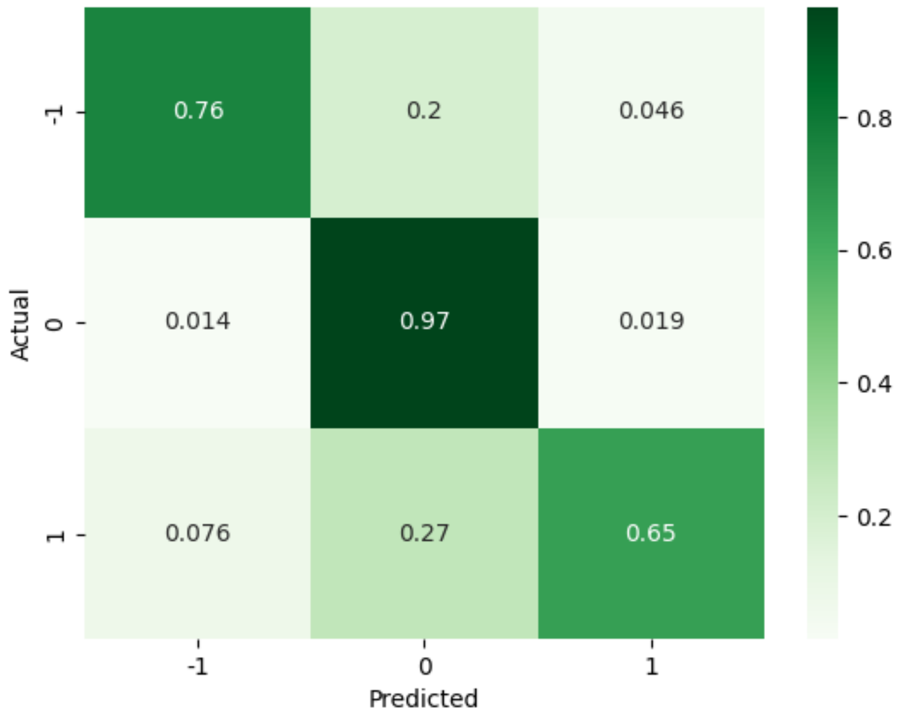

Figure 6: Confusion matrix showing for best performing GPT-3

Chapter 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

The basis of this thesis was outlined by five research questions which are:

- What is the difference between sentiment analysis in political texts compared with other domains?
- Which transformer-based language model architecture is more suitable for sentiment analysis in political texts?
- How can common occurrences in daily languages such as sarcasm and irony and human factors such as biases affect the result of sentiment analysis in political texts?
- How can the sentiment analysis model be made more accessible in areas such as political arenas to increase user trust and ensure transparency?

As seen by the results, using transformer-based classification models can provide acceptable accuracies for classifications in political texts (86.09%), but when compared to other domains such as reviews and customers satisfactions where higher accuracies are attenable. This can be attributed to the fact that when customers are reviewing products for example, they tend to leave direct comments without any complexity which can then lead to better text classification. Because of the nuances in political texts and news in general, this phenomenon can lead to lower classification accuracies.

The best transformer-based language model architecture that is more suitable for political text classification is the RoBERTa model. This can be traced to the fact the RoBERTa was pre-trained on news articles, unlike BERT which was pre-trained on English Wikipedia and Toronto BookCorpus [109]. GPT-3 performance proves that though it can be modified for text classification, its training data is mostly suitable for text generation.

The lower accuracy values can be attributed to the nuances found in everyday speech. This includes ironies, metaphors, and touches of sarcasm. As an NLP, transformer-based models are limited to their training data, and as such, detecting these nuances cannot be achieved except if they are pre-trained by such a dataset. Finally, for sentiment analysis to be adaptable for public consumption, there needs to be another classification model that can handle the tiny nuances in human interactions and communications.

5.2 Future Works

Future studies could broaden the scope of the analysis beyond a single language to include many languages and political settings. This approach would give information on how sentiment analysis models operate across linguistic and cultural variances, hence enhancing cross-cultural insights. Also, future research could improve sentiment analysis models' ability to identify and analyse figurative language, such as sarcasm and irony, which are common in political debate. This development would allow for a more accurate representation of sentiments.

REFERENCES

[1] F. Zhang, H. Fleyeh, X. Wang, and M. Lu, “Construction site accident analysis using text mining and natural language processing techniques,” *Automation in Construction*, vol. 99, pp. 238–248, Mar. 2019, doi: https://doi.org/10.1016/j.autcon.2018.12.016.

[2] Marjan Hosseinia, E. C. Dragut, and A. Mukherjee, “Stance Prediction for Contemporary Issues: Data and Experiments,” Jan. 2020, doi: https://doi.org/10.18653/v1/2020.socialnlp-1.5.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *Proceedings of the 2019 Conference of the North*, vol. 1, 2019, doi: https://doi.org/10.18653/v1/n19-1423.

[4] Y. Liu *et al.*, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” *arXiv (Cornell University)*, Jul. 2019, doi: https://doi.org/10.48550/arxiv.1907.11692.

[5] Ashish Vaswani *et al.*, “Attention is All you Need,” *Nips.cc*, pp. 5998–6008, 2017, Accessed: Dec. 12, 2019. [Online]. Available: https://papers.nips.cc/paper/7181-attention-is-all-you-need

[6] Dzmitry Bahdanau, K. Cho, and Yoshua Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *arXiv (Cornell University)*, Sep. 2014, doi: https://doi.org/10.48550/arxiv.1409.0473.

[7] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: https://doi.org/10.1038/323533a0.

[8] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal Neural Language Models," *proceedings.mlr.press*, Jun. 18, 2014. https://proceedings.mlr.press/v32/kiros14.html

[9] Y. Kim, "Convolutional Neural Networks for Sentence Classification," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, doi: https://doi.org/10.3115/v1/d14-1181.

[10] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[11] P. W. Battaglia, R. Pascanu, M. Lai, D. Rezende, and K. Kavukcuoglu, "Interaction Networks for Learning about Objects, Relations and Physics," *arxiv.org*, Dec. 2016, Available: http://arxiv.org/abs/1612.00222

[12] F. Scarselli, M. Gori, Ah Chung Tsoi, M. Hagenbuchner, and G. Monfardini, "The Graph Neural Network Model," *IEEE Transactions on Neural*

Networks, vol. 20, no. 1, pp. 61–80, Jan. 2009, doi: https://doi.org/10.1109/tnn.2008.2005605.

[13] P. W. Battaglia, R. Pascanu, M. Lai, D. Rezende, and K. Kavukcuoglu, "Interaction Networks for Learning about Objects, Relations and Physics," *arxiv.org*, Dec. 2016, Available: http://arxiv.org/abs/1612.00222

[14] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering," 2016. Available: https://proceedings.neurips.cc/paper/2016/file/04df4d434d481c5bb723be1b6df1ee65-Paper.pdf

[15] J. Bastings, I. Titov, W. Aziz, D. Marcheggiani, and K. Simaan, "Graph Convolutional Encoders for Syntax-aware Neural Machine Translation," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, doi: https://doi.org/10.18653/v1/d17-1209.

[16] J. Zhou *et al.*, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020, doi: https://doi.org/10.1016/j.aiopen.2021.01.001.

[17] J. Han, M. Kamber, and J. Pei, *Data Mining: concepts and techniques*. Elsevier, 2012. doi: https://doi.org/10.1016/c2009-0-61819-5.

[18] B. Croft, D. Metzler, and T. Strohman, *Search Engines*. Pearson Higher Ed, 2011.

[19] R Baeza-Yates and Berthier Ribeiro-Neto, *Modern information retrieval : the concepts and technology behind search*. New York: Addison Wesley, 2011.

[20] K. S. Jones, *Readings in information retrieval*. Amsterdam U.A: Morgan Kaufmann, 2006. Available: https://dl.acm.org/citation.cfm?id=275543

[21] M. Marzouk and M. Enaba, "Text analytics to analyze and monitor construction project contract and correspondence," *Automation in Construction*, vol. 98, pp. 265–274, Feb. 2019, doi: https://doi.org/10.1016/j.autcon.2018.11.018.

[22] Q. Li, S. Li, S. Zhang, J. Hu, and J. Hu, "A Review of Text Corpus-Based Tourism Big Data Mining," *Applied Sciences*, vol. 9, no. 16, p. 3300, Jan. 2019, doi: https://doi.org/10.3390/app9163300.

[23] J. Lee *et al.*, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, Sep. 2019, doi: https://doi.org/10.1093/bioinformatics/btz682.

[24] H. Jung and B. G. Lee, "Research trends in text mining: Semantic network and main path analysis of selected journals," *Expert Systems with Applications*, vol. 162, p. 113851, Dec. 2020, doi: https://doi.org/10.1016/j.eswa.2020.113851.

[25] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, “Text Mining in Big Data Analytics,” *Big Data and Cognitive Computing*, vol. 4, no. 1, p. 1, Jan. 2020, doi: https://doi.org/10.3390/bdcc4010001.

[26] D. Tao, P. Yang, and H. Feng, “Utilization of text mining as a big data analysis tool for food science and nutrition,” *Comprehensive Reviews in Food Science and Food Safety*, Feb. 2020, doi: https://doi.org/10.1111/1541-4337.12540.

[27] Z. Liu, D. Huang, K. Huang, Z. Li, and J. Zhao, “FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining,” *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, Jul. 2020, doi: https://doi.org/10.24963/ijcai.2020/622.

[28] S. Kumar, A. K. Kar, and P. V. Ilavarasan, “Applications of text mining in services management: A systematic literature review,” *International Journal of Information Management Data Insights*, vol. 1, no. 1, p. 100008, Apr. 2021, doi: https://doi.org/10.1016/j.jjimei.2021.100008.

[29] Otmane Azeroual, G. Saake, M. Abuosba, and Joachim Schöpfel, “Text data mining and data quality management for research information systems in the context of open data and open science,” Dec. 2018, doi: https://doi.org/10.48550/arxiv.1812.04298.

[30] Z. H. Kilimci and S. Akyokuş, “The Analysis of Text Categorization Represented With Word Embeddings Using Homogeneous Classifiers,” *IEEE*

Xplore, Jul. 01, 2019. https://ieeexplore.ieee.org/document/8778329 (accessed Jul. 25, 2022).

[31] R. A. Haraty and R. Nasrallah, "Indexing Arabic texts using association rule data mining," *Library Hi Tech*, vol. 37, no. 1, pp. 101–117, Mar. 2019, doi: https://doi.org/10.1108/lht-07-2017-0147.

[32] N. Z. Da, "The Computational Case against Computational Literary Studies," *Critical Inquiry*, vol. 45, no. 3, pp. 601–639, Mar. 2019, doi: https://doi.org/10.1086/702594.

[33] M. Marzouk and M. Enaba, "Text analytics to analyze and monitor construction project contract and correspondence," *Automation in Construction*, vol. 98, pp. 265–274, Feb. 2019, doi: https://doi.org/10.1016/j.autcon.2018.11.018.

[34] S. Sarkar, S. Vinay, Chawki Djeddi, and J. Maiti, "Text Mining-Based Association Rule Mining for Incident Analysis: A Case Study of a Steel Plant in India," pp. 257–273, Jan. 2021, doi: https://doi.org/10.1007/978-3-030-71804-6_19.

[35] M. T. Ibrahim and A. Tella, "Analysis of Text Mining from Full-text Articles and Abstracts by Postgraduates Students in Selected Nigeria Universities," *International Journal of Higher Education*, vol. 9, no. 4, p. 169, Jun. 2020, doi: https://doi.org/10.5430/ijhe.v9n4p169.

[36] B. Grundel *et al.*, “Merkmalsextraktion aus klinischen Routinedaten mittels Text-Mining,” vol. 118, no. 3, pp. 264–272, Jul. 2020, doi: https://doi.org/10.1007/s00347-020-01177-4.

[37] X. Chen and I. Beaver, “A Semi-Supervised Deep Clustering Pipeline for Mining Intentions From Texts,” Feb. 2022, doi: https://doi.org/10.48550/arxiv.2202.00802.

[38] M. Brown and P. Shackel, “Text Mining Oral Histories in Historical Archaeology,” *International Journal of Historical Archaeology*, Jan. 2023, doi: https://doi.org/10.1007/s10761-022-00680-5.

[39] N. Jindal and B. Liu, “Review spam detection,” *Proceedings of the 16th international conference on World Wide Web - WWW ’07*, 2007, doi: https://doi.org/10.1145/1242572.1242759.

[40] C. C. Aggarwal and C. Zhai, “An Introduction to Text Mining,” *Mining Text Data*, pp. 1–10, 2012, doi: https://doi.org/10.1007/978-1-4614-3223-4_1.

[41] C. C. Aggarwal and C. Zhai, “A Survey of Text Classification Algorithms,” *Mining Text Data*, pp. 163–222, 2012, doi: https://doi.org/10.1007/978-1-4614-3223-4_6.

[42] W. W. Cohen, “Learning trees and rules with set-valued features,” *www.osti.gov*, Dec. 1996, Accessed: Jan. 16, 2022. [Online]. Available: https://www.osti.gov/biblio/430732

[43] R. Zhang, Seira Hidano, and Farinaz Koushanfar, “Text Revealer: Private Text Reconstruction via Model Inversion Attacks against Transformers,” Sep. 2022, doi: https://doi.org/10.48550/arxiv.2209.10505.

[44] K. Lai, Y. Long, B. Wu, Y. Liu, and B. Wang, “Semorph: A Morphology Semantic Enhanced Pre-trained Model for Chinese Spam Text Detection,” Oct. 2022, doi: https://doi.org/10.1145/3511808.3557448.

[45] C. Li, Q. Liu, and K. Ma, “DCCL: Dual-channel hybrid neural network combined with self-attention for text classification,” vol. 20, no. 2, pp. 1981–1992, Jan. 2022, doi: https://doi.org/10.3934/mbe.2023091.

[46] L. Huang, “Evaluation of query strategy of active learning for text classification,” *2nd International Conference on Artificial Intelligence, Automation, and High-Performance Computing (AIAHPC 2022),* Nov. 2022, doi: https://doi.org/10.1117/12.2641410.

[47] A. Gera, A. Halfon, E. Shnarch, Y. Perlitz, L. Ein-Dor, and N. Slonim, “Zero-Shot Text Classification with Self-Training,” *ACLWeb*, Dec. 01, 2022. https://aclanthology.org/2022.emnlp-main.73 (accessed Aug. 06, 2023).

[48] J. Chen, R. Zhang, Y. Mao, and J. Xu, “ContrastNet: A Contrastive Learning Framework for Few-Shot Text Classification,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 10492–10500, Jun. 2022, doi: https://doi.org/10.1609/aaai.v36i10.21292.

[49] X. Liu *et al.*, “Unsupervised Image Classification by Ideological Affiliation from User-Content Interaction Patterns,” May 2023, doi: https://doi.org/10.48550/arxiv.2305.14494.

[50] W. Xu, X. Jiang, J. Desai, B. Han, F. Yan, and F. Iannacci, “KDSTM: Neural Semi-supervised Topic Modeling with Knowledge Distillation,” Jul. 2023, doi: https://doi.org/10.48550/arxiv.2307.01878.

[51] J. Kim, S. Jang, E. Park, and S. Choi, “Text classification using capsules,” *Neurocomputing*, vol. 376, pp. 214–221, Feb. 2020, doi: https://doi.org/10.1016/j.neucom.2019.10.033.

[52] M. Yang, W. Zhao, J. Ye, Z. Lei, Z. Zhao, and S. Zhang, “Investigating Capsule Networks with Dynamic Routing for Text Classification,” *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, doi: https://doi.org/10.18653/v1/d18-1350.

[53] J. Zeng, J. Li, S. Yan, C. Gao, M. R. Lyu, and I. King, “Topic Memory Networks for Short Text Classification,” Jan. 2018, doi: https://doi.org/10.18653/v1/d18-1351.

[54] R. Wang, Z. Li, J. Cao, T. Chen, and L. Wang, “Convolutional Recurrent Neural Networks for Text Classification,” *IEEE Xplore*, Jul. 01, 2019. https://ieeexplore.ieee.org/document/8852406 (accessed Aug. 19, 2022).

[55] L. Yao, C. Mao, and Y. Luo, “Graph Convolutional Networks for Text Classification,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 7370–7377, Jul. 2019, doi: https://doi.org/10.1609/aaai.v33i01.33017370.

[56] C. Du, Z. Chen, F. Feng, L. Zhu, T. Gan, and L. Nie, “Explicit Interaction Model towards Text Classification,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 6359–6366, Jul. 2019, doi: https://doi.org/10.1609/aaai.v33i01.33016359.

[57] L. Huang, D. Ma, S. Li, X. Zhang, and H. Wang, “Text Level Graph Neural Network for Text Classification,” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3444–3450, Jan. 2019, doi: https://doi.org/10.18653/v1/d19-1345.

[58] K. Shah, H. Patel, D. Sanghvi, and M. Shah, “A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification,” *Augmented Human Research*, vol. 5, no. 1, Mar. 2020, doi: https://doi.org/10.1007/s41133-020-00032-0.

[59] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, “Deep Learning--based Text Classification,” *ACM Computing Surveys*, vol. 54, no. 3, pp. 1–40, May 2021, doi: https://doi.org/10.1145/3439726.

[60] B. Liu, “Sentiment Analysis and Opinion Mining,” *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, May 2012, doi: https://doi.org/10.2200/s00416ed1v01y201204hlt016.

[61] H. T. Phan, N. T. Nguyen, V. C. Tran, and D. Hwang, “An approach for a decision-making support system based on measuring the user satisfaction level on Twitter,” *Information Sciences*, vol. 561, pp. 243–273, Jun. 2021, doi: https://doi.org/10.1016/j.ins.2021.01.008.

[62] M. Hoang, O. A. Bihorac, and J. Rouces, “Aspect-Based Sentiment Analysis using BERT,” *ACLWeb*, Sep. 01, 2019. https://aclanthology.org/W19-6120

[63] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, “Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques,” *Third IEEE International Conference on Data Mining*, doi: https://doi.org/10.1109/icdm.2003.1250949.

[64] F. Å. Nielsen, “A new ANEW: Evaluation of a word list for sentiment analysis in microblogs,” *ResearchGate*, 2011. https://www.researchgate.net/publication/50378498_A_new_ANEW_Evaluation_of_a_word_list_for_sentiment_analysis_inmicroblogs

[65] X. Fang and J. Zhan, “Sentiment analysis using product review data,” *Journal of Big Data*, vol. 2, no. 1, Jun. 2015, doi: https://doi.org/10.1186/s40537-015-0015-2.

[66] K. Schouten and F. Frasincar, “Survey on Aspect-Level Sentiment Analysis,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 3, pp. 813–830, Mar. 2016, doi: https://doi.org/10.1109/tkde.2015.2485209.

[67] A. Zadeh, M. Chen, Soujanya Poria, E. Cambria, and L.-P. Morency, “Tensor Fusion Network for Multimodal Sentiment Analysis,” *Empirical Methods in Natural Language Processing*, pp. 1103–1114, Sep. 2017, doi: https://doi.org/10.18653/v1/d17-1115.

[68] F. Huang, X. Li, C. Yuan, S. Zhang, J. Zhang, and S. Qiao, “Attention-Emotion-Enhanced Convolutional LSTM for Sentiment Analysis,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2021, doi: https://doi.org/10.1109/tnnls.2021.3056664.

[69] A. Iqbal, R. Amin, J. Iqbal, R. Alroobaea, A. Binmahfoudh, and M. Hussain, “Sentiment Analysis of Consumer Reviews Using Deep Learning,” *Sustainability*, vol. 14, no. 17, p. 10844, Aug. 2022, doi: https://doi.org/10.3390/su141710844.

[70] N. Leelawat *et al.*, “Twitter data sentiment analysis of tourism in Thailand during the COVID-19 pandemic using machine learning,” *Heliyon*, vol. 8, no. 10, p. e10894, Oct. 2022, doi: https://doi.org/10.1016/j.heliyon.2022.e10894.

[71] G. Hu, T.-E. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li, “UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition,”

ACLWeb, Dec. 01, 2022. https://aclanthology.org/2022.emnlp-main.534 (accessed Aug. 07, 2023).

[72] S. F. Yilmaz, E. B. Kaynak, A. Koc, H. Dibeklioglu, and S. S. Kozat, “Multi-Label Sentiment Analysis on 100 Languages With Dynamic Weighting for Label Imbalance,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 1, pp. 1–13, 2021, doi: https://doi.org/10.1109/tnnls.2021.3094304.

[73] J. Mutinda, W. Mwangi, and G. Okeyo, “Sentiment Analysis of Text Reviews Using Lexicon-Enhanced Bert Embedding (LeBERT) Model with Convolutional Neural Network,” *Applied Sciences*, vol. 13, no. 3, p. 1445, Jan. 2023, doi: https://doi.org/10.3390/app13031445.

[74] B. Kaur, K. Srivastava, and Er. Priyanka, “Sentiment Analysis for Product Review,” *International Research Journal of Modernization in Engineering Technology and Science*, vol. 5, no. 4, Apr. 2023, doi: https://doi.org/10.56726/irjmets36776.

[75] T. de Melo and C. M. S. Figueiredo, “Comparing News Articles and Tweets About COVID-19 in Brazil: Sentiment Analysis and Topic Modeling Approach,” *JMIR Public Health and Surveillance*, vol. 7, no. 2, p. e24585, Feb. 2021, doi: https://doi.org/10.2196/24585.

[76] G. Alvarez, J. Choi, and S. Strover, “Good News, Bad News: A Sentiment Analysis of the 2016 Election Russian Facebook Ads,” *International Journal*

of Communication, vol. 14, no. 14, pp. 3027–3053, May 2020, doi: https://doi.org/10.26153/tsw/9852.

[77] L. Thurnay and T. J. Lampoltshammer, “Human biases in government algorithms,” *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*, Sep. 2020, doi: https://doi.org/10.1145/3428502.3428529.

[78] A. Chakraborty and S. Bose, “Around the world in 60 days: an exploratory study of impact of COVID-19 on online global news sentiment,” *Journal of Computational Social Science*, vol. 3, no. 2, pp. 367–400, Oct. 2020, doi: https://doi.org/10.1007/s42001-020-00088-3.

[79] J. K. Alwan, A. J. Hussain, D. H. Abd, A. T. Sadiq, M. Khalaf, and P. Liatsis, “Political Arabic Articles Orientation Using Rough Set Theory With Sentiment Lexicon,” *IEEE Access*, vol. 9, pp. 24475–24484, 2021, doi: https://doi.org/10.1109/access.2021.3054919.

[80] A. Hossain, M. Karimuzzaman, M. M. Hossain, and A. Rahman, “Text Mining and Sentiment Analysis of Newspaper Headlines,” *Information*, vol. 12, no. 10, p. 414, Oct. 2021, doi: https://doi.org/10.3390/info12100414.

[81] S. E. Bestvater and B. L. Monroe, “Sentiment is Not Stance: Target-Aware Opinion Classification for Political Text Analysis,” *Political Analysis*, vol. 31, no. 2, pp. 1–22, Apr. 2022, doi: https://doi.org/10.1017/pan.2022.10.

[82] T. A. Salgueiro, Emilio Recart Zapata, Damián Ariel Furman, J. Manuel, and P. Nicolás, "A Spanish dataset for Targeted Sentiment Analysis of political headlines," Aug. 2022, doi: https://doi.org/10.48550/arxiv.2208.13947.

[83] C. Wang, M. Li, and A. J. Smola, "Language Models with Transformers," Apr. 2019, doi: https://doi.org/10.48550/arxiv.1904.09408.

[84] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, "CTRL: A Conditional Transformer Language Model for Controllable Generation," *arXiv:1909.05858 [cs]*, Sep. 2019, Available: https://arxiv.org/abs/1909.05858

[85] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism," *arXiv:1909.08053 [cs]*, Mar. 2020, Available: https://arxiv.org/abs/1909.08053

[86] X. Ma *et al.*, "A Tensorized Transformer for Language Modeling," *Neural Information Processing Systems*, 2019. https://proceedings.neurips.cc/paper/2019/hash/dc960c46c38bd16e953d97cdeefdbc68-Abstract.html (accessed Aug. 07, 2023).

[87] G. Zhao, J. Lin, Z. Zhang, X. Ren, Q. Su, and X. Sun, "Explicit Sparse Transformer: Concentrated Attention Through Explicit Selection," *arXiv:1912.11637 [cs]*, Dec. 2019, Available: https://arxiv.org/abs/1912.11637

[88] M. Radfar, Athanasios Mouchtaris, and S. Kunzmann, "End-to-End Neural Transformer Based Spoken Language Understanding," Oct. 2020, doi: https://doi.org/10.21437/interspeech.2020-1963.

[89] Y. Wang, Shafiq Joty, M. R. Lyu, I. King, C. Xiong, and Steven, "VD-BERT: A Unified Vision and Dialog Transformer with BERT," *arXiv (Cornell University)*, Apr. 2020, doi: https://doi.org/10.48550/arxiv.2004.13278.

[90] A. Gulati *et al.*, "Conformer: Convolution-augmented Transformer for Speech Recognition," *arXiv (Cornell University)*, May 2020, doi: https://doi.org/10.48550/arxiv.2005.08100.

[91] C. Rosset, C. Xiong, M. Phan, X. Song, P. Bennett, and S. Tiwary, "Knowledge-Aware Language Model Pretraining," *arXiv.org*, Feb. 04, 2021. https://arxiv.org/abs/2007.00655 (accessed Aug. 07, 2023).

[92] X. Zhang *et al.*, "RSTNet: Captioning with Adaptive Attention on Visual and Non-Visual Words," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, doi: https://doi.org/10.1109/cvpr46437.2021.01521.

[93] Y. Liu, "Fine-tune BERT for Extractive Summarization," *arXiv:1903.10318 [cs]*, Sep. 2019, Available: https://arxiv.org/abs/1903.10318

[94] P. Shi and J. Lin, “Simple BERT Models for Relation Extraction and Semantic Role Labeling,” *arXiv.org*, Apr. 10, 2019. https://arxiv.org/abs/1904.05255 (accessed Aug. 07, 2023).

[95] C. Sun, X. Qiu, Y. Xu, and X. Huang, “How to Fine-Tune BERT for Text Classification?,” *Lecture Notes in Computer Science*, pp. 194–206, 2019, doi: https://doi.org/10.1007/978-3-030-32381-3_16.

[96] G. Jawahar, B. Sagot, and D. Seddah, “What Does BERT Learn about the Structure of Language?,” *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3651–3657, 2019, doi: https://doi.org/10.18653/v1/p19-1356.

[97] W. Su *et al.*, “VL-BERT: Pre-training of Generic Visual-Linguistic Representations,” *arXiv:1908.08530 [cs]*, Feb. 2020, Available: https://arxiv.org/abs/1908.08530

[98] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, doi: https://doi.org/10.18653/v1/d19-1410.

[99] W. Liu *et al.*, “K-BERT: Enabling Language Representation with Knowledge Graph,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol.

34, no. 03, pp. 2901–2908, Apr. 2020, doi: https://doi.org/10.1609/aaai.v34i03.5681.

[100] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, “MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices,” *arXiv (Cornell University)*, Apr. 2020, doi: https://doi.org/10.18653/v1/2020.acl-main.195.

[101] X. Jiao *et al.*, “TinyBERT: Distilling BERT for Natural Language Understanding,” *Empirical Methods in Natural Language Processing*, Nov. 2020, doi: https://doi.org/10.18653/v1/2020.findings-emnlp.372.

[102] F. Hamborg and K. Donnay, “NewsMTSC: A Dataset for (Multi-)Target-dependent Sentiment Classification in Political News Articles,” *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, doi: https://doi.org/10.18653/v1/2021.eacl-main.142.

[103] L. Gebhard and F. Hamborg, “The POLUSA Dataset: 0.9M Political News Articles Balanced by Time and Outlet Popularity,” *arXiv (Cornell University)*, Aug. 2020, doi: https://doi.org/10.1145/3383583.3398567.

[104] W.-F. Chen, H. Wachsmuth, Khalid Al Khatib, and B. Stein, “Learning to Flip the Bias of News Headlines,” *Proceedings of the 11th International Conference on Natural Language Generation*, pp. 79–88, Nov. 2018, doi: https://doi.org/10.18653/v1/w18-6509.

[105] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: a review of BERT-based approaches," *Artificial Intelligence Review*, Feb. 2021, doi: https://doi.org/10.1007/s10462-021-09958-2.

[106] Q. Zhu and J. Luo, "Generative Pre-Trained Transformer for Design Concept Generation: An Exploration," *Proceedings of the Design Society*, vol. 2, pp. 1825–1834, May 2022, doi: https://doi.org/10.1017/pds.2022.185.

[107] Z. Gao, A. Feng, X. Song, and X. Wu, "Target-Dependent Sentiment Classification With BERT," *IEEE Access*, vol. 7, pp. 154290–154299, 2019, doi: https://doi.org/10.1109/access.2019.2946594.

[108] B. Zeng, H. Yang, R. Xu, W. Zhou, and X. Han, "LCF: A Local Context Focus Mechanism for Aspect-Based Sentiment Classification," *Applied sciences*, vol. 9, no. 16, pp. 3389–3389, Aug. 2019, doi: https://doi.org/10.3390/app9163389.

[109] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *arxiv.org*, Nov. 2017, Available: https://arxiv.org/abs/1711.05101

[110] Y. Zhu *et al.*, "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books," *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, doi: https://doi.org/10.1109/iccv.2015.11.