Unveiling Political Rhetoric: Exploring Natural Language Processing Methods to Analyze Political Discourse

by

Tasnia Zaman
20301296
Sabrina Tajnim Sithi
21101044
Md. Sadi Ashraf
23141067

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
October 2024

Declaration

It is hereby declared that

1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Tasnia Zaman	Sabrina Tajnim Sithi
Tasnia Zaman	Sabrina Tajnim Sithi
20301296	21101044

MD SADI ASHRAF

Md. Sadi Ashraf

23141067

Approval

The thesis titled "Unveiling Political Rhetoric: Exploring Natural Language Processing Methods to Analyze Political Discourse" submitted by

1. Tasnia Zaman (20301296)
2. Sabrina Tajnim Sithi (21101044)
3. Md. Sadi Ashraf (23141067)

Of Summer, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on August 23, 2024.

Examining Committee:

Supervisor:
(Member)

Farig Yousuf Sadeque

Dr. Farig Yousuf Sadeque

Assistant Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Ethics Statement

This research uses natural language processing techniques to analyse political conversation while upholding the most rigorous moral principles. The research completely uses publicly accessible political writings and statements, guaranteeing the fact that no confidential or delicate data is retrieved or employed. Every single statement or allusions to certain political leaders are done in a way that respects the standards of objectivity as well as accountability. Furthermore, every piece of information utilised in the research is combined and categorised to avoid identifying specific viewpoints. The research appropriately adds to our comprehension of connecting speech, sentiment, and political communication while upholding the values associated with academic ethics, honesty, and impartial investigation regarding political speech.

Abstract

Politics has significant effects on how a country develops and how we live our daily lives, affecting things like public services, social norms, and economic policies, extending its impact to other countries. The United States, regarded as the most influential political entity, through its election results, not only influences national policies but also has far-reaching global effects across several fields. This research focuses on the evaluation of political textual data collected from Twitter and Reddit comments. This study's main objective is to enhance the detection accuracy of political comments. To achieve a significant improvement in the detection of political comments, we used advanced language models, specifically the Bidirectional Long Short-Term Memory (BiLSTM), Multilayer BiLSTM, Bidirectional Encoder Representations from Transformers (BERT), Robustly optimized BERT approach (RoBERTa), and A Lite BERT (ALBERT) models. These models were used to significantly increase efficiency and accuracy. By improving this detection ability, social media platforms will be able to effectively moderate political discourse and obtain deeper insights into public support for different political parties.

Keywords: Political discourse; Natural language processing (NLP); Political comments; Republican; Democrat; United States.

Acknowledgement

All praise to the Almighty Allah for whom our thesis have been successfully concluded with minimal disruptions. Lastly, we want to be thankful to our parents as without their assistance, it might not be feasible. Their prayers and support have brought us to the precipice of our graduation.

Table of Contents

List of Figures

Chapter 1

Introduction

Language is one of the most important utensils used by common citizens to express their political opinions towards political parties and also for politicians to advance their political goals, get support, win elections, and create change in society. Since textual data sources are a widespread and extremely comprehensive resource in political research, natural language processing (NLP) is becoming more prevalent in many social science branches, notably the study of political discourse analysis. According to the researchers, political discourse is a verbal representation of public behavior in the field of political-cultural backgrounds, which is a professional utilization of language that depends on the national and socio-historically shaped mindset of its speakers [13]. Social media platforms are becoming more important for political discussion. These mediums let the common public interact with current events, communicate with other people, and express their opinions. Any citizen can send a message to a politician on modern online social media platforms, such as Facebook, Twitter, Reddit, and YouTube, because all users are treated equally. We chose to focus on American political comments for our research due to the immense global influence of the United States in both the political and economic domains. The two massively influential political parties in the United States are the Republic and the Democratic Party.

Furthermore, the results of American elections and political choices have broad effects that impact international markets and policies. While numerous efforts have been made to identify political comments, our main focus is on the critical task of determining which political comment indicates which political party. We also want to find out which political party's supporters show more aggression based on the political comments, visualize their political sentiment towards each party, and show the changes in aggression, sentiment, and popularity of each political party's supporters in a certain period. Besides, we want to show the change in public sentiment and popularity during the 2024 election, first when Biden was the Democratic Party's nominee for president, and later when Kamala Harris became the nominee. We also want to compare each political party's sentiment, level of aggression, and popularity on Twitter and Reddit from January 2024 to now.

To achieve the goal, we have made use of a large dataset that contains a variety of textual information, including posts and comments that have been taken from social media sites, specifically Reddit and Twitter. Our team members have classified the dataset, and each category has been labeled with thoughtfulness. For our following NLP model creation and analysis, this categorization offers a strong basis. Next, by using advanced machine learning methods and advances in Natural Language Processing (NLP), we have concentrated on building strong NLP models. After complete training on the large dataset, these models can identify which side's political comments they agree with. Our focus is on the opinions and sentiments of people because it is interesting to research these as they can reveal their views on elections or other political events. Political parties may also be interested in knowing whether or not people are supporting their respective campaigns, the aggression of the people, and whether their public support is increasing or decreasing over time.

1.1 Research Statement

The primary objective of our thesis is to analyze public opinion and how people feel about various political parties, as well as politicians' opinions on their party's ideologies and those of their opponents. Our research aims to look at the political opinions of the public and the party ideologies of the political parties by extracting data from online social media.

1.2 Research Objective

This research aims to use a much more effective Natural Language Processing (NLP) technique to analyze political discourse among various political parties. To be more specific, this research seeks to categorize political texts into topics. We believe that this will facilitate the visualization and classification of ideas expressed within political party discussions.

Key Objectives:

1. Utilizing Natural Language Processing (NLP) techniques for the analysis of political texts.
2. Unveiling patterns and themes within political discourse.
3. Implementing sentiment analysis to gauge the general public's attitude towards political speeches.
4. Investigating the role of sentiment analysis in predicting electoral outcomes.
5. Understanding public opinion on specific political issues.

Chapter 2

Related Work

To explore NLP methods to analyze political discourse, we have studied some research articles. To select the articles, we looked into those that have been published recently and have a respectable amount of citations. In our collection of research articles, we used Google Scholar as a resource. At first, we looked for articles that related to our research topic, which is political discourse analysis. After gathering about thirty research articles, we selected about twenty of them that were more relevant to our study. The selected articles have been arranged in sequence according to their publication dates, which has allowed us to create a coherent and up-to-date knowledge base for our research.

In their paper, Conover et al. tried to develop useful characteristics for differentiating between people who lean left or right ideologically [1]. They explore manual annotations of training data that includes almost 1,000 individuals who are actively involved in U.S. political discussions to investigate multiple methods for differentiating between individuals with left- and right-wing political affiliations. They first concentrate on content-based characteristics, showing that a support vector machine can correctly identify a user's political identity in tweets 91% of the time using data provided by individuals. Primarily, they seeded their sample using two popular political hashtags, which are #tcot ("Top Conservatives on Twitter") and #p2 ("Progressives 2.0"), which lean left and right, respectively. They also build two networks from the collection of political tweets: one that relies on retweet edges and another based on mention edges. A weighted, undirected edge connects the nodes in the mentioned network that represent users A and B, if either. The total number of mentions among each of the users determines the weight of each edge. The structure of the retweet network is the same. Users were categorized into three groups by each annotator: "left", "right", or "ambiguous" depending on the content of the tweets they posted throughout the six-week duration of the study. Democrats and progressives were the groups most closely connected with a "left" political alignment, while Republicans, conservatives as libertarians, and Tea Party members were largely connected to a "right" political affinity. They apply linear support vector machines (SVMs) to distinguish people into the "left" and "right" groups for content-based categorization. They give the confusion matrix for every classifier and a rating of accuracy according to 10-fold cross-validation to measure performance for various feature sets. They train a model using support vector machines on a feature-user matrix that corresponds to the number of TFIDF-weighted terms (unigrams) included in each user's tweets to create a performance benchmark. Following these preprocessing stages, 13,080 features—each denoting a single term—are included in the final database. They provided details on several methods based on network and content analysis. They have, however, demonstrated that hashtag features with an abundant amount of information are nearly as good at capturing political affiliation and have the added advantage of generalizing without requiring the network to be reclustering to make room for new members. Finally, they identify the web pages that are particularly popular with Twitter users on the left and right of the political spectrum, respectively, as a demonstration of the concept for the usefulness of this prediction skill. This method provides new insights into the audience appeal of various media channels, which can be used to enhance the selection process for web-based advertising purchases.

The author of the next research paper, Patricia L. Dunmire analyzes PDA (political discourse analysis) in the larger framework of the political and linguistic shifts in the human and social sciences that took place in the last quarter of the 20th century [2]. In this study, the topic of PDA research, which arose amid linguistic and political shifts, is introduced. PDA and CDA share a critical perspective through which to examine the connection between politics and discourse, which makes them closely aligned approaches. Both methods examine how political dominance and power are expressed via discourse structures, looking at how speech shapes discipline, social structures, power, and economic exploitation. PDA addresses more general subjects like race and racism, broadening its focus beyond traditional political problems. The research's precise methodology and datasets aren't covered in length in the publication. Rather, it offers a broad summary of PDA, going over many theories of politics and discourse and evaluating research according to its theoretical underpinnings and sociopolitical concerns. The contribution of the work is in three parts: it gives a summary of PDA, clarifies its theoretical and practical notions, and highlights its function in analyzing discourse's political nature. It summarizes research on the composition and role of political discourse, the relationship between political conduct, discourse, and cognition, and the influence of speech and writing on political systems and procedures. The study also emphasizes PDA's role in the politicization of social life outside of conventional political spheres. The article doesn't offer precise findings in the conventional sense. Rather, it provides a summary of PDA, going over its theoretical basis, fundamentals, and salient features. It highlights the critical perspective through which PDA investigates social change and power relations in discourse, highlighting the importance of PDA in comprehending the nuances of political language. The absence of concrete conclusions or findings, a thorough explanation of PDA's limitations as a research method, and a critical analysis of potential biases in the evaluated studies or frameworks are only a few of the drawbacks of the study. The study emphasizes the use of a variety of conceptual frameworks, methodologies, and data in PDA research, even if it makes no clear suggestions for particular future projects. Future research on the relationship between PDA and CDA, the influence of political discourse on public opinion and behavior, the examination of the intersectionality of social categories in political discourse, and the analysis of political actors' discursive techniques are some possible avenues of inquiry.

Johnson et al. authored a paper focusing on Twitter data to forecast political positions [3]. They make use of Twitter content, frames, and temporal activity, whereas earlier efforts have mostly concentrated on online discussion forum data or societal structures. Using a binary predicate TWEETS (P1, ISSUE) to indicate whether a politician has spoken on a certain issue, the researchers provide a unique keyword-based heuristic to identify the topics politicians are tweeting about. This method is distinct from earlier research that frequently used social interactions or argument structures to predict position. Their research demonstrates the larger framework for studies in political discourse analysis on Twitter at the level of Twitter activity. Previous research has looked at several topics, including sentiment analysis, inferring social networks, forecasting ideologies, and analyzing Twitter network effects on political events. Their research differs as it focuses on forecasting politicians' positions through the use of a probabilistic modeling framework, namely Probabilistic Soft Logic (PSL). The process entails gathering information on the positions and recent tweets of 32 well-known American politicians on 16 distinct subjects. They use a three-model approach: Model 1 is initialized with knowledge about political party affiliation, followed by Model 2, which includes findings from local models for sentiment analysis and issues, and Model 3, which combines all of the previous models with higher-level Twitter activity. By predicting stance and agreement more accurately across a range of topics, the performance evaluation reveals that the PSL models perform better than a local baseline model. The work admits several shortcomings despite its contributions. It is noted that the issue detection process relies too heavily on a keyword-based heuristic, indicating the necessity for a more sophisticated method. Potential biases are introduced by manual annotation for politicians with unavailable replies on ISideWith.com, and a restriction is noted about the elimination of stance correlations among party members. Regarding future research, the report recommends examining how party member stance correlations may be included, improving the keyword-based heuristic, and determining whether the method can be used by less well-liked politicians. They also suggest contrasting their strategy with other approaches that are based on the argument structures of online discussion forums. Their study makes a substantial contribution to the area by presenting a novel approach that combines probabilistic modeling and text analysis to forecast politicians' positions on Twitter. The findings point to increased accuracy over baseline models, and the shortcomings that have been noted and the projects that have been proposed for the future offer insightful avenues for future study in the field of political discourse analysis.

According to research by Mohd et al., sentiment analysis is a tool for mining online social media platforms for user opinions expressed via text [18]. The aforementioned platforms house a large amount of written content that is influenced by different ideas. Thus, the observed growth is a result of the substantial rise in demand for sentiment analysis. There are several lexical and semantic elements included in this piece of art. These traits may be learned with the use of sentiment lexicons and semantic models. These features help deep learning keep data samples consistent in size, which gets rid of the necessity for zero padding. To evaluate the effects of lexico-semantic characteristics on classification performance, the authors utilize several lexicons and semantic models. Scientific research has shown that classifiers—which include machine learning and deep learning techniques—are more effective when these traits are incorporated. The study uses objective textual data from multiple online social networking platforms for sentiment analysis. The data is obtained from online resources. The authors assess the significance and impact of lexico-semantic parameters on the precision of classification using a variety of semantic frameworks and lexicons. Numerous classifiers have been taught utilizing the Lexico-Semantic features to do evaluations of experiments. There is proof from real-world use that adding Lexico-Semantic features enhances classifier performance. As a result, methods based on deep learning and machine learning perform better. Semantic models and lexicons can examine how lexico-semantic features impact categorization accuracy. To do experiment assessments, different classifiers undergo training utilizing additional criteria in addition to their lexico-semantic characteristics. The authors evaluate many of the most cutting-edge methods in the fields of machine learning and deep learning by using Lexico-Semantic features. We run a series of tests on these strategies to determine their effectiveness. Empirical studies show that classifiers perform better when provided with lexico-semantic information. Empirical research has shown that Lexico-Semantic characteristics, such as the use of machine learning and deep learning techniques, enhance classifier effectiveness for sentiment analysis. In deep learning contexts, an unchanged data collection size can be maintained by employing lexico-semantic features, which can be produced from sentiment lexicons and semantic models. As a result, zero padding, which is standard practice in the field, is rendered unnecessary.Zero padding, which is common practice in the field, is therefore superfluous. Incorporating lexico-Semantic features improves the system's data classification capabilities, according to the authors' comprehensive evaluation of many sophisticated methodologies. Research has shown that the addition of the additional Lexico-Semantic features improves the accuracy of sentiment analysis by making it easier to identify the polarity of sentiment in English text. Empirical research has shown this, thus we know this impact is beneficial.

In another paper, Paritosh D. Katre aims to use NLP techniques for text analytics and political speech transcript visualization [6]. The core intention of this paper was to get past the challenges associated with analyzing huge quantities of unorganized written information with computer systems methods. He provides a better alternative to conventional linguistic techniques in political discourse analysis by highlighting the successful and productive utilization of NLP. The research makes significant use of methods based on NLP in the interpretation of political written documents, and Python was chosen for programming for textual corpus processing. Important libraries such as Word Cloud, Numpy, Matplotlib, Pandas, and Counter are implemented for evaluation and display to create clusters of words and track the prevalence of words. After that, graphical representations like bar graphs, time series plots, and lexical dispersion plots are created using Python tools. His excellent example of how NLP-based text analytics can identify issues in an extensive set of spoken conversations is one of his greatest achievements. His research establishes that such practices facilitate sophisticated political speech examination and are a faster alternative to conventional syntactic procedures. Word clouds, lexical dispersion plots, and time series plots are a few examples of visualizations that help more comprehensively comprehend speech than transcripts. The dataset that was utilized in the study has not been addressed in the publication specifically. However, because it concentrates on political speech transcripts, the collection may include a wide variety of speeches from different political situations and even in other languages. The study focuses more on emphasizing the benefits of NLP approaches than on evaluating any potential downsides. Concerns over the precision and generalizability of the NLP techniques employed, as well as their suitability for usage in various linguistic or political situations, may be among the limitations. The paper does not specifically delineate forthcoming endeavors or suggest additional investigations. On the other hand, given the study's implications, future research could look into the application of NLP techniques in various linguistic and cultural contexts, examine how views on politics affect public sentiment, incorporate additional databases for a more thorough analysis, create sophisticated analytical models, and carry out long-term studies to monitor the development of political speech.

In the next paper, a study conducted by Beleveslis introduces a hybrid methodology to examine the sentiment of election-related tweets. Similar to the prior paper, this research provides significant contributions to not only the field of information technology but also political discussions [4]. This particular paper focuses on sentiment analysis of Greek tweets related to the recent European elections using a hybrid method. This method combines Greek lexicons and classification methods. The relevance of this method derives from the limited amount of study and resources devoted to sentiment analysis in the Greek language. This method entails using a probabilistic classification model to forecast sentiment. One of the key aspects of the data collection method involves the tweet data, which is mainly handled by removing hashtags, URLs, and emoticons using the "tweet-preprocessor" Python package. The authors discussed two precisely chosen Greek words that have undergone thorough review and evaluation to ensure their compatibility with the vocabulary included in the tweets, guaranteeing integration of the lexicons. This also improves the comprehensiveness of the sentiment analysis. The study experimented with a dataset consisting of 1,640 evaluated tweets linked to the Greek General Elections in January 2015. These tweets feature emotions distinguished as happy, negative, and neutral. The authors use several classification methodologies, such as Random Forest, Decision Tree, and XGBoost, to get substantial results. Besides, the authors include the adjustment of parameters, as this determines the most effective combination of parameters for the classification algorithms. In the paper, the authors successfully predict the emotion of the analyzed tweets accurately, which produces encouraging results. Mainly, the Random Forest approach emerged as the most effective of the classifiers used. This gives the highest levels of accuracy, precision, and F-score. Moreover, this paper shows the complex interplay of emotions, revealing certain pre-election events that had a discernible impact on public opinion. In addition, the study provides useful insights about the frequency of negative hashtags. This acts as a marker for categorizing tweets that convey negative feelings. Furthermore, the study also examined the temporal dimension of Twitter engagement. Also in their findings, the authors declare that a significant increase in activity was seen after May Day, which coincided with Easter festivities. During this time, there were intense debates on several subjects, including not only the disaster in Mati but also the speeches of major figures. This sentimental analysis provides critical perspectives on the political actions of both Europe and Greece. This uncovered a distinct and extensive resistance to austerity measures. These were achieved via the development and effective use of an innovative hybrid technique.

Next, a study by Stegmeier et al. is devoted to analysing Twitter discussion regarding two global elections using multiple methods of discourse analysis [7]. Here, the author aims to understand the importance of different topics within the debates.Concerning both scenarios, this defines the scope of global interaction on Twitter. On January 14 and March 6, 2015, the researchers streamed tweets with the hashtags Net Neutrality and Climate change in order to gather data from Twitter. The aforementioned hashtags were regarded by them as relevant search phrases for the international discussions in both of the political domains. . They collected a total of 884,729 tweets and retweets, with 380,890 tweets related to climate change and 503,839 tweets related to net neutrality. In order to allocate tweets to particular nations, researchers additionally conducted geolocation analysis using the Data Science Toolkit (DSTK). They also managed to ascertain the geo-location of 54 Climate change tweets and 56 internet neutrality tweets. The analysis focused on the ten most frequent countries in the corpus. We have observed that a multi-method approach that combines geolocation analysis, network analysis, and keyword analysis has been used to analyse Twitter communication. Twitter users' whereabouts were ascertained by geolocation analysis, which was also utilised to construct specialised sub corpora for linguistic and social research.This helped in understanding the transnational nature of the discourse. Network analysis was conducted to examine the relationships between discourse actors and identify the most visible and important actors in the debates. AntConc, a concordancer software, was employed to identify topic-specific vocabulary and analyse word frequency lists, allowing for the identification of significant keywords in the discourse.Combining these techniques yielded a multi-angle picture on how worldwide conversation is formed. As a result, this helped in understanding the interplay between political, social, thematic, and technical aspects of discourse formation. Here, the study shows insightful methods used in the analysis, which led to supplementary and complementary results. This provides a comprehensive understanding of the discourse on climate change and net neutrality on Twitter. Network analysis revealed that US US participants in government, business, and the press ruled the discourse on net neutrality, whereas the discourse on climate change remained more uniform.Keyword analysis provided insights into the most frequently discussed topics in each subcorpus.The researchers also emphasised how US American and Anglo-Saxon users and groups predominate in Twitter discussions on these two policy areas.Lastly, a more comprehensive understanding of the discourse on these global political issues on Twitter is needed.

According to another paper, Adinda Natassa Valentine Hutabarat investigates the political rhetoric that China has produced around the "One Belt, One Road (OBOR)" proposal [5]. Specifically concentrating on the impacts of discourse and the participation of translators in the process, this research was carried out based on written interviews that were done with Chinese professionals. To do this, a PDA is used. The purpose of this research is to investigate the level of correlation that exists between the educational backgrounds of academics and the translations that they produce, as well as to evaluate the level of comprehension that they have about the theoretical and practical components of the One Belt, One Road initiative. The study obtained its information from the official websites of the United Nations Mission in Melbourne and the Ministry of Foreign Affairs of the People's Republic of China in Melbourne. Both of these websites are located in Melbourne. By emphasizing alternative terms that carry the same meaning as those found in the original texts, the translators want to promote a good impression of China among the individuals who are supposed to receive the translation. The goal of this approach is to infuse the texts that are being targeted with new relevance while at the same time decreasing the influence of certain words. Throughout the whole of the translation process, five primary tactics are used. Omission, metaphor, foreignization, and domestication are some of the tactics that are used, along with modifications to the structure and substance of the text also being included. Through the use of the suggestive functions that are supplied by the components, it is possible to send messages that are in line with the discourse in an effective manner. The research task concentrates on the manner in which the stance of the Chinese government affects the evaluation of content as expressed by interpreters working within their operations. It concentrates on looking at the methods of interpretation employed in political debates surrounding the One Belt, One Road (OBOR) project. The study additionally illustrates how important translators are to reshaping the story and projecting an optimistic view of China. By pinpointing the essential components, the research obtained a thorough grasp of the way translation processes operate.

In a subsequent publication, Cabot and colleagues present combined approaches for political analysis that take into account argumentation and sentiment, as well as compare them with the context of multitasking learning [8]. According to the study, students significantly improved their ability to convey legislative issues, recognize officials' political associations, and discern political viewpoints in newspapers. The findings demonstrate the significance of sentiment and connection in discussions about politics since individuals demonstrated appreciable growth across the board. From a political perspective in the news media, there is a possibility of left- or right-leaning bias in political news. They categorize them as left, right, or center, noting the bias of each source. 2008 and 5761 articles make up 30% of the training data utilized for validation, together with training and test sets including 412 publicly posted Facebook posts by US politicians, for this task. There are 9792, 2458, and 2356 items in the test, training, and validation sets, correspondingly. Predicting a Republican or Democrat to fill a post held by an invisible politician is the task. Articles from the Media Frames Corpus4 cover five policy areas, including immigration, gun control, and Out of 23,580 articles, 15% of test and 15% of validation data are used to predict the framing dimension using article-level annotation, resulting in fifteen possible framing dimensions (political, economic, etc.). The VU Amsterdam dataset is used for metaphor recognition; it comprises 9,017 sentences with binary (literal or metaphorical) classifications for each word. They employed a dataset from SemEval to classify emotions. The tweets in the dataset were classified as neutral or belonging to one of eleven emotion classes. They show that the STL model outperforms Li and Goldwasser's (2019) text-based approach, with RoBERTa improving document encoding for political perspective prediction. The research demonstrates the notable gains in performance that MTL with metaphor detection made on all three tasks, greatly increasing the STL model in both MTL configurations. Despite these advancements, there is still an opportunity for progress, including exploring novel tasks like emotion and disinformation detection and utilizing more sophisticated MTL techniques.

Brian Sharber in his paper, tried to figure out the political aspects of various news sources in an attempt to develop a framework for researching the linguistic facets of polarisation in the news media [9].The research project examines the terminology employed through various political parties. It finds terms and blends of words which one category uses more frequently than another group. The following can point to a serious bias or slant regarding their dialogue. His thesis, which draws inspiration from earlier relevant work, uses NLP to identify polarisation in the political news sections of many political blogs and online news sources. Data from several political blogs and websites has to be gathered to start assessing polarisation among various political communities. Every top-page story, opinion or not, from the political news sections of all the media outlets was collected every day at different times, with the date of creation and content logged for each piece. The list's timeline covers articles released between January 1, 2020, and July 31, 2020. Following a performance evaluation of each classifier, the best 3 classifiers (Random Forest, SGD, and LinearSVC) are identified in his research work. Keeping this in mind, VotingClassifier, a combined learner, combines the best three classifiers to improve the accuracy of classification through a voting mechanism. Theoretically, this specially designed classifier would yield the best accuracy because it would aggregate the votes from all three classifiers, with a majority vote determining the result of classification for each dataset entry. The dataset was first randomly merged before being divided into two sets: a training set consisting of 80% of the data and a testing set of 20% to start training the models. This was done to ensure that every time, a fresh set of training and testing data would be used, protecting the dataset from outliers or weeks when there were no political news stories. Following the creation of the Pipeline object, the models were fitted to the training set using the proper labels in the pole column—0 denoting left-leaning articles and 1 denoting right-leaning articles. The research paper's outcome shows that the best-performing model, with an accuracy rate of 83.729%, suggests that the model has acquired some degree of success in learning the language used by both sides in their respective discourse. As anticipated, the best-performing model made use of the voting classifier. It suggested building a unique ensemble learner and using the three best-performing classifiers as a voting mechanism to classify documents. Thus, a framework for identifying and evaluating the language used by opposing political poles was presented in this research study. The small number of publications that were obtained is one of the research's limitations. The model would perform better at classifying documents if it had access to more training data. This report proposes future work that will use an expanded database of articles with a new schedule for training and testing new models.

The other paper by Kubin et al. is mainly a comprehensive systematic review as it investigates social media's influence on political polarization [11]. This paper rigorously analyzes 94 papers, which consist of 121 studies on political polarization. As a result, the authors demonstrate a valuable synthesis of existing research within the context of the Annals of the International Communication Association. This consistently demonstrates that media that promotes certain attitudes intensifies polarization. However, this study combines quantitative and qualitative methodologies to provide a deeper comprehension of the role of social media in political polarization. This study predominantly focuses on Twitter and American samples.

Moreover, the systematic review highlights a notable deficiency in current research. So, the authors stated the lack of studies that investigate the potential of social media in reducing polarization. Here, this paper highlights the necessity of measuring and defining polarization more precisely. The authors also show a clear difference between representative and convenience samples. Assisting the studies, the research assistant carefully recorded the kind of sample used in each case. However, not all findings from this paper are in agreement regarding the link between political polarization and traditional media. In addition, some studies used advanced analyses with multiple data types and samples. Moreover, some studies found no effect between traditional media and polarization. Others suggested that partisan media predicts affective polarization, but mainstream media does not. The analysis shows an increase in research on media and polarization over the past 10 years. However, there is a hyperfocus on analyzing Twitter and American samples. It indicates a need for more diverse and representative samples. This paper shows that the impact of mainstream media on polarization is inconsistent. It effectively detects important patterns, discrepancies, and potential subjects for further investigation. With this study, we have seen a thorough assessment of the present status of media and political polarization research.

In response to the gradually changing discussion of politics on Twitter, the research paper by Jayo et al. offers an innovative technique for improving immediate evaluation using computerized data methods [10]. The researchers employ a collection of labeled political declarations to create a written classification algorithm by accessing the Parallel Statements Initiative's classification structure, which is usually employed in analyzing text in political statements. Next, this sort of algorithm is used on information from Twitter, merging political analysis and NLP comprehension. The paper's writers investigate the influence of pertinent data, for instance, the sender's political beliefs or preceding tweets, to enhance the model's reliability. The objective of this research is to strengthen the study of political parties' conversations on internet sites, particularly Twitter, which is growing into a crucial means for open discussion among elected leaders and the wider population. A better categorization of election-related tweets is made possible by the approach suggested, which entails a separate picture of the candidates due to their ideological inclination. The writers employ phrase encapsulations and multilayer neural networks (CNNs) as the machine learning framework for recognizing texts using an information set that includes 5,000 identified tweets and 100 identified electoral statements. The strategy works well because it makes use of developed Word2Vec models, especially those that were predetermined using the search engine's dataset. The geometric mean, F-measure (macro), and accuracy rate are among the parameters utilized throughout the study to assess the algorithm's effectiveness. This examines how the presidential election in the US in 2016 was discussed on social media, mostly Twitter, offering perspectives on the electoral conversation that surrounded the vote. They found that merging ideological declarations as well as labeled tweet collections significantly improved the f-measure. Moreover, adding pertinent data improves the accuracy of the algorithm even more. The significant discrepancy across experiments resulting from just a handful of categorized tweets in comparison to larger instances is one of the constraints acknowledged by the article. To improve execution, the writers encourage evaluating neural network designs tailored to various labels and indicate the necessity of a larger structured sample for the analysis of multiple labels. The study draws attention to the linguistic distinctions between political statements and online platforms, highlighting the need for specialized methods when examining political speech on websites such as Twitter. In summary, this study makes recommendations for additional areas of study, such as delving further into the subcategories of electoral debate on online platforms while tackling possible prejudices and constraints associated with the use of set-up word insertion algorithms. Even though that study acknowledges its limits, it is still a significant achievement considering that it effectively employs several techniques to comprehend political discourse in the technologically advanced age of today.

On the other hand, Mishra et al. discuss a novel approach to data extraction using NLP [12]. This method is insightful, as the study focuses on not only sentiment analysis but also opinion mining. Speaking of the method used in this paper, the authors extract meaning from written words, i.e., study thoughts, feelings, judgments, and attitudes. The importance of sentiment analysis in various domains, for example, social networks and product opinions, has been discussed in this paper. It mentions the use of supervised machine learning and lexical-based techniques to measure sentiments captured in digital form. A variety of sources, including blogs, forums, social media (such as Facebook and Twitter), feedback on goods, and user-reviewed web pages, can provide the necessary data for analysing sentiment. Natural language processing is used in the procedure to interpret the data's conveyed emotions, regardless if they are adverse or affirmative. Evaluation of sentiment may be performed on particular phrases, the entire manuscript, or it may involve just a few chosen parts. However, assessing sentiment at the level of the file can be a little tricky.The use of language-based approaches and machine learning methods for sentiment detection is discussed in the study. Methods including Naive Bayes and SVM (Support Vector Machines) are mentioned. To describe it differently, the researchers talk about how to gather evidence by identifying both beneficial and detrimental speech. In addition, they discuss determining if evaluations are favourable or undesirable by utilising a machine learning model with training dataset.It discusses the use of a feature extractor to extract public opinions from acquired data and provide subjective and factual responses. Using specific training data and testing for accuracy, we can develop an N-gram model for categorization. Finally, in this paper, we learn about the use of supervised machine learning and lexical-based techniques to measure sentiments captured in digital form. However, the authors of the paper mention the challenges faced in sentiment analysis, which include the growing noise on the web due to abbreviations, slang, and emoticons.

The next article by Bestvater et al. examines the popular technique in political text analysis that approximates the author's position using sentiment scoring [14]. Its main contributions are that it questions the applicability of this methodology, distinguishes "stance detection" from sentiment analysis, and highlights the significance of precisely defining and quantifying stance for analyses of political phenomena that follow. Although the focus of the work is on open-ended survey responses and political conversations on social media, the dataset utilized in the study is not specifically addressed. Four popular sentiment analysis techniques are employed in this paper: two supervised classifiers (BERT and SVM classifier) and two dictionary-based techniques (VADER and Lexicoder Sentiment Dictionary). Using the R package Quanteda, the Lexicoder Sentiment Dictionary (LSD), intended for political texts, is deployed. The ground-truth sentiment and attitude labels for every text are used to train the supervised classifiers, BERT, and support vector machine classifiers. The mean score for F1 has been supplied, and the efficacy of the method is assessed using five times cross-validation techniques. According to the paper's outcomes, dictionary techniques and automated algorithms can evaluate public opinion adequately, yet both are not robust at detecting rooted textual locations. The work not only discusses the drawbacks and assessment biases associated with executing disposition through sentiment measurements but also provides tangible instances of how opinion and viewpoint in political text analysis are discordant. Additionally, the constraints of the study are that it relies too much on illustrations from online forum talks and ambiguous comments; it doesn't go into great detail about how different political scenarios or social circumstances may affect how opinion and position correspond; and it doesn't thoroughly explore other methods for effectively expressing a stand. The study proposes several potential areas of study, including looking into other strategies for stance identification, examining the effects of various political as well as cultural scenarios, creating enhanced methods of sentiment assessment to precisely capture true viewpoints, and assessing and contrasting the effectiveness of various sentiment methods to analyze data. The tangible consequences emphasize how important it is for scholars to understand the theoretical distinctions between sentiment and stance because utilizing sentiment as a stand-in for stance can lead to inaccurate findings, including severe inaccuracy when measuring.

Another work by Bryan Strawser uses discourse analysis techniques to examine the ideas about reality that are promoted by the US alt-right as well as Donald Trump [16]. The study makes a substantial contribution towards our acknowledgment of this new political discourse by emphasizing the crafting as well as presenting truth throughout a variety of news, including digital platforms, literature, and argumentation. The research mainly reveals the intricate ways that perspectives, relationships of authority, and communal grouping indicators like race, ethnic background, and belief system interact with assertions of fact through an analysis of discourse approach based on Foucault. This method additionally sheds greater clarity regarding the rhetorical techniques used by Trump, including the alt-right, yet at the same time emphasizes the larger repercussions on the rule of law and the general conversation. Being cognizant that there are threats such conversations pose to dominant descriptions, which consequence in the development of sharply divided and politicized social discussions, makes a crucial discovery. This study underlines the imperative of a thorough investigation concerning the argumentative tactics of significant political participants by highlighting the influence of political discourse on societal variation. By investigating the convoluted connection between speech patterns and social groupings, the investigation illuminates the complexity of today's political discourse. It reinforces its evaluation by combining actions with discussions about ethnicity, socioeconomic status, and sexual orientation. This reveals the subtle manner in which personal characteristics are deliberately used to promote election mobilization and separation. To help readers recognize the particular features of contemporary political discussion, this investigation points out inadequacies within the field along with advocates having novel techniques as well as architectures, highlighting the need for a worldwide viewpoint on discursive methodologies. The study highlights the important ramifications for democratic institutions of the relationship between these discursive activities and social identity concerns. The author gathers information from a variety of sources, including speeches, videos from public events, blogs, podcasts, social media platforms, internet forums, campaign rallies, and alternative news websites. While specific datasets or sources within each category are not addressed directly, the complete study guarantees a thorough comprehension of the public discourse related to Trump and the alt-right. Language and communication, the main techniques used in Foucault's discourse analyses, allow for a critical investigation of power dynamics, knowledge formation, and social behaviors. The research provides important insights into the discursive tactics used by Trump and the alt-right by examining speech, social media, publications, and public events. The effect of political context on interpretation, potential selection bias resulting from the large body of available literature, and time and word count limits are acknowledged limitations. The study acknowledges the possibility of selection and media biases in news reporting, especially when it comes to modern political personalities. The study suggests future research paths that focus on investigating the linkages between truth discourse and audiences' varying interpretations of it. Prospective directions for future research include comparative studies with international political players, analyses of the function of identity in political discourse, and a more thorough examination of the global ramifications of these discursive practices.

Another recent study by Németh gives us the application of NLP to political polarization [15]. This scoping review offers insights into not only dominant trends but also promising results. This scoping review aims to clarify how NLP research has conceptualized and measured political polarization. Besides, it characterizes the integration of different kind of research prototypes in this particular area. For the data collection, this paper does not explicitly mention the specific data sources used in the research. However, it primarily mentions that the studies included in the scoping review used various data sources, such as tweets. These tweets are mainly extracted from parliamentary speeches and other sources. First and foremost, this review paper highlights the importance of different aspects of data usability. For instance, when choosing data sources for studying political polarisation, accessibility, and fruitfulness are taken into consideration. In addition, this paper mentions that about 40 studies were used in the review of tweets because, in most countries, it is the major accessible source of politically relevant texts. Moving on, in this scoping review, the authors emphasize the fact that the data source has an impact on the result. As a result, the authors suggest that sources like parliamentary speeches might be more justified for certain research questions. Here, the authors of this paper acknowledge the significance of the nature of data by emphasizing not only how the data should be collected and filtered but also the context in which it was created, as this will ensure accurate and meaningful analysis. The searches for relevant studies were conducted using Google Scholar, and studies published between January 1, 2010, and June 29, 2021, were included. Moreover, the initial search terms used were "political polarisation" and "NLP", with additional synonyms added to both terms. Hard and soft exclusion criteria were defined to filter out irrelevant hits from the search results. The findings were synthesized in a narrative report, and technical details were provided in the Supplementary Material. The author of this review paper also observes various text analytic methods used in the studies, including Wordfish, Wordscores, Wordshoal, topic modeling, word embedding, supervised machine learning, and sentiment analysis. Qualitative methods and simple quantitative approaches were excluded from the review. In addition, the author of this review paper discussed the operationalization of polarization, which includes the measurement of political positions using either scaling or classification approaches. Here, the authors identified 154 relevant studies on the use of NLP in research on political polarization. The authors found biases towards the US context, Twitter data, and the use of machine learning approaches in the studies. This review covered different layers of the political public sphere, but very few involved more than one layer. Lastly, the political position of the authors must be neutral.

In their most recent piece of work, Torregrosa et al. showed a mixed approach to analyzing aggressive political discourse on Twitter. It sheds light on the complexities of online communication [17]. This methodology allows for a faster and more reliable analysis. Also, this technique bypasses the need for manual coding, which automatically eliminates any errors in data entry. This research has connections with the 2008 financial crisis and the escalating political tensions in Europe. However, the data presented in this study include tweets gathered during the time covering the disbandment of Madrid's regional parliament on March 10th to the election day on May 26th, 2021. This paper suggests a Nobel technique that includes sentiment and emotional analysis. Thus, analyzing political narratives and framing, detecting misinformation, and evaluating COVID-related narratives on social media. This paper primarily studies tweets that are spread by candidates during the CAM election on the Twitter platform. This facilitates the assessment of their communication habits and emotional displays. The data consists of textual information derived from the tweets. This study includes qualitative content analysis and quantitative metrics such as sentiment analysis, n-gram analysis, and hashtag analysis. The Valence Aware Dictionary and Sentiment Reasoner (VADER) are used here for sentiment analysis; this is intended for analyzing sentiment in social media material. It is particularly trained and intended to handle social media information. Moreover, N-gram analysis is used to detect emergent patterns in the tweets, which offer more intricate information than a simple word count. In addition, hashtag analysis is used as a complementary approach to the n-gram methodology because it acts as a semantic aggregator of meanings. Content analysis includes a review of tweets selected by using NLP metrics given through quantitative analysis. Then, transformers are used for semantic analysis of certain terms like 'freedom' and 'freedoms'. In this study, using this hybrid method, researchers found that the political campaign was shown to be less acrimonious than previously assumed. Interestingly, no relationship has been observed between the tone of the speech and its level of variation. Leftist groups used more aggressive rhetoric in contrast with their moderate competitors. Twitter largely serves as a tool for analyzing public emotion rather than as a forum for explicitly expressing political messages. Plus, researchers have discovered the rise of new political entities that emphasize the instability of the existing political system. Finally, this research provided substantial details about the patterns of political communication in the Internet domain during this crucial campaign period.

Chapter 3

Dataset Description

3.1 Analysis of Dataset

The dataset consists of comments sourced from two popular social media sites: Reddit and Twitter. These platforms act as platforms for user-generated content, presenting a wide variety of opinions, sentiments, and conversations from people worldwide. Our main objective is to identify the political comments of the two popular political parties, which are the Republican Party and the Democratic Party, and also find the sentiment, aggression, and popularity of both political parties. We also aim to highlight shifts in public sentiment and popularity throughout the 2024 election, first when Biden was the Democratic Party's presidential candidate and then when Kamala Harris took over. Furthermore, from September 20 to October 4, 2024, we have graphically demonstrated the daily shifts in popularity, aggression, and sentiment for each party. The textual entries in this dataset have been manually categorized and annotated by our team members to distinguish and classify the comments into three unique categories: Republican, Democrat, or Non-Identical. Every comment has been carefully examined to figure out whether it corresponds with these political affiliations or whether it is distinct from other comments.

We collected at least 10.9k comments in total from Twitter and Reddit for making predictions in our NLP models. The total dataset contains 10926 rows and 11 columns. It is divided into three segments: training, validation, and test sets. The training set comprises 9,287 rows and 6 columns, while the validation and test sets consist of 1,638 rows and 6 columns individually. Additionally, for our 15-day graphical visualization, which is from September 20 to October 4, 2024, we have collected around 6k comments from Twitter. Moreover, we have gathered around 1700 data points after Kamala Harris became the nominee of the Democratic Party. Lastly, we collected around 6k data points of 2024 political comments from Reddit for a graphical comparison of each party's sentiment, aggression levels, and popularity on both Twitter and Reddit from January 2024 to now. To determine the sentiment of each political party supporter, we have made a column named sentiment in every dataset using Valence Aware Dictionary and sEntiment Reasoner (VADER) which is a sentiment analysis tool specifically designed for social media contexts. Based on the comment's sentiment, this tool has assigned positive, negative, or neutral in the sentiment column. Then, we used the Detoxify model to analyze toxicity in a comment, which generated toxicity scores for each comment in categories such as toxicity, severe_toxicity, obscene, threat, insult, and identity_attack and we appended the toxicity scores as new columns in the dataset.

3.2 Category Construction

The political comments collected from Twitter and Reddit are categorized into three distinct labels for analytical purposes. Comments expressing support for the Republican Party are given label 1, while those supporting the Democratic Party are given label 2. Comments that do not support either party or are neutral are labeled as 0. This categorization helps a deeper comprehension of public opinion patterns about these two major political parties by organizing and evaluating the political sentiment expressed throughout these social media platforms. In a table, the definitions and categories are highlighted. In Table [3.1], the table is displayed.

Text	**Republic**	**Democrat**	**Non-identical**
This just keeps getting worse. Here's a like for you for sounding the alarm and a one-fingered salute to #BidenHarris for selling out our country and using our tax dollars to import #Democrat voters.	1	0	0
What former President Trump did... is inexcusable but, BUT WHAT?! Well you can add yourself to the list of no longer republican, now a tRumplican. You can give yourself over to the #trumpCult, sell your soul, sell out your country. BUT I WON'T! f*** your endorsement	0	1	0
Some Tucker texts let raw truth out Which violates Fox rules The private comments left no doubt Fox knows its fans are fools 4/5 Then Tucker crudely criticized His colleagues and his boss So Tucker's ass was tossed aside A red line had been crossed 5/5	0	0	1

Table 3.1: Table of political statements

3.3 Annotation

The entire dataset was manually annotated by our three dedicated team members. Before the start of the annotation, the members studied the features of the US political parties and the parties' political ideologies so that, while annotating the comments, they could identify the categories. Weekly meetings were held by the team to discuss difficult cases and enhance their understanding and annotation approach. This devoted work and collaboration ensured that our dataset was of a superior standard. To annotate the entire dataset, three steps were taken. Firstly, for each batch of comments, pairs of annotators were assigned to work on. For simplicity, let's refer to the members as A, P, and S. A and P marked the first 500 comments individually, followed by A and S for the following 500, and finally P and S for the final 500. A and P annotated the next 500 once more, and so on. The pairs, therefore, covered the entire dataset. Secondly, to verify their annotations, the members met regularly. The third member, who was not in the pair for that batch, served as a neutral party if there was any dispute amongst the participants in that particular batch. For example, S would choose the most appropriate value if A and P could not agree on a specific comment. The third person would thus provide a new take on the problem and settle the disagreement amongst the other members. After our dataset's training on multiple models, we chose RoBERTa, the model with the highest accuracy, to auto-label a new CSV dataset. However, since the RoBERTa model did not give 100% accuracy, we conducted a manual check of the assigned labels to ensure accuracy for further analysis.

3.4 Statistics

There are approximately 10.9k comments in our dataset overall for running in our models, all written in English. For our 15-day graphical visualization around 6k comments taken from Twitter, around 1700 data points after Kamala Harris became the nominee of the Democratic Party have been collected from Twitter, and we collected around 6k data points of 2024 political comments from Reddit for a graphical comparison between Reddit and Twitter. These comments cover a broad range of views on politics and are systematically arranged to indicate their neutrality or support for either the Democratic or Republican parties. By examining this dataset, the different lengths of comments are displayed in a text length distribution graph, which provides information about the typical length of political speech on social media. A word cloud is another tool used to emphasize common themes and subjects in the dataset by highlighting the most frequently used words for all the categories. Data distribution percentages provide a clear picture of the dataset's political leanings by displaying the percentage of comments within each category.

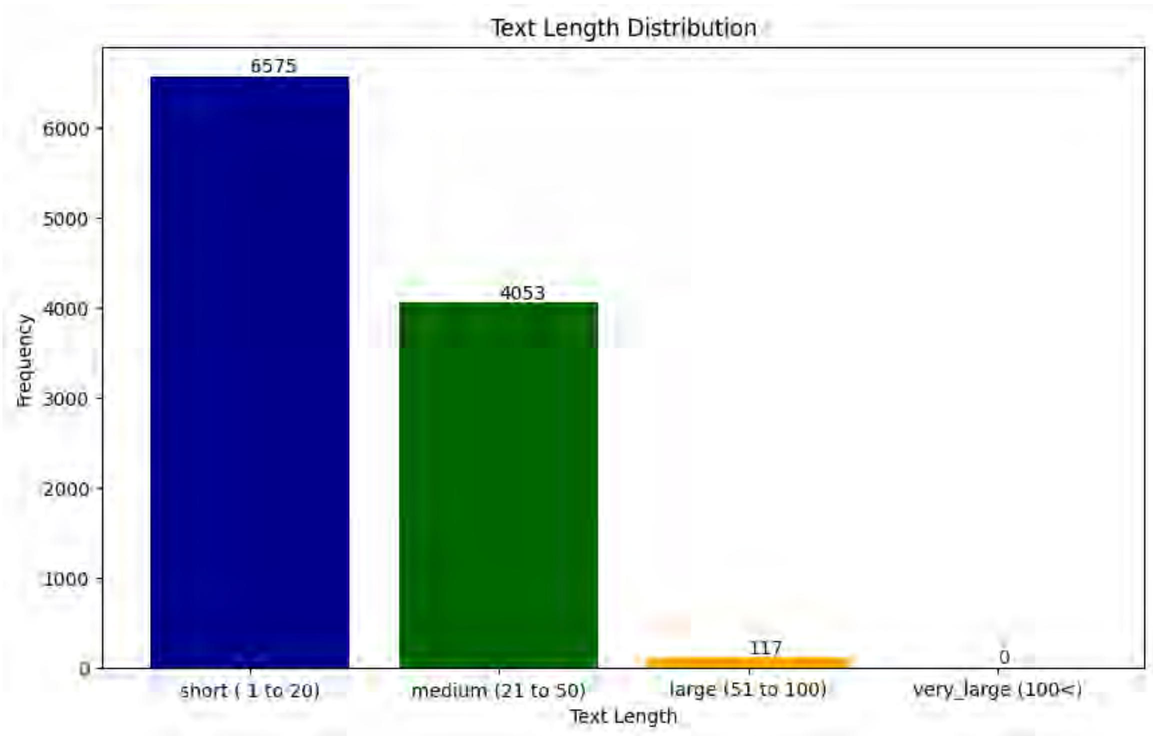


Figure 3.4.1: Text Length Distribution

The dataset, as seen in Fig. [3.4.1], has approximately 6500 short texts that contain 1-20 words. Moreover, 4053 entries fall within the 21-50 word range. It is worth noting that very few texts extend beyond 100 words. This distribution makes it clear that the dataset places a strong focus on concise sentences.

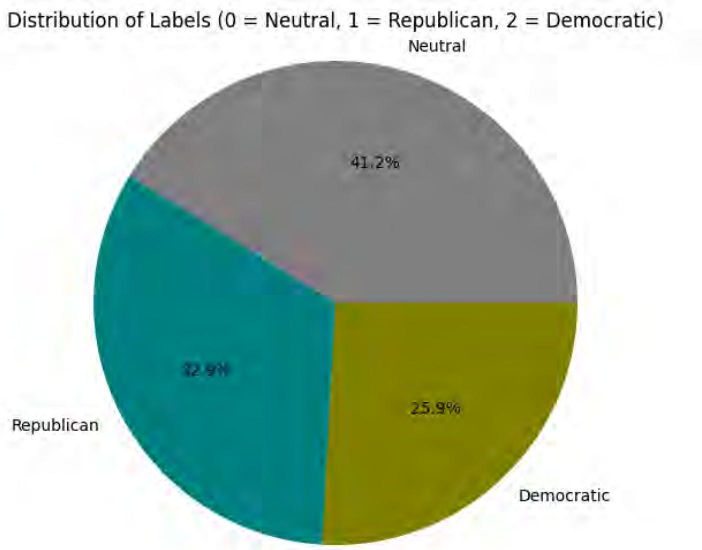


Figure 3.4.2: Distribution of Classes

Fig. [3.4.2] shows that 41.1% of the text and comments in the dataset are classified as non-identical (label 0), making up the majority of the data. The remaining 33% are identified as Republican (label 1) and 25.9% are indicating Democrat (label 2) comments

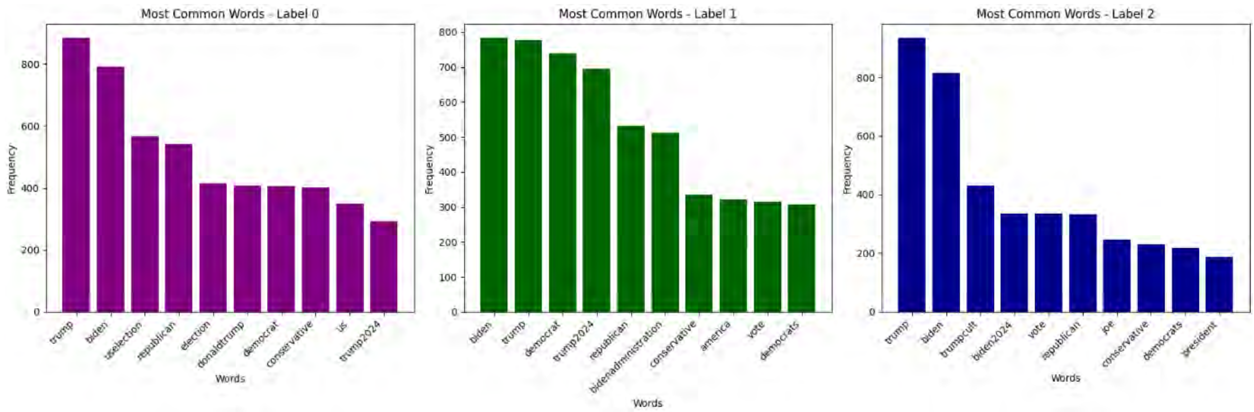
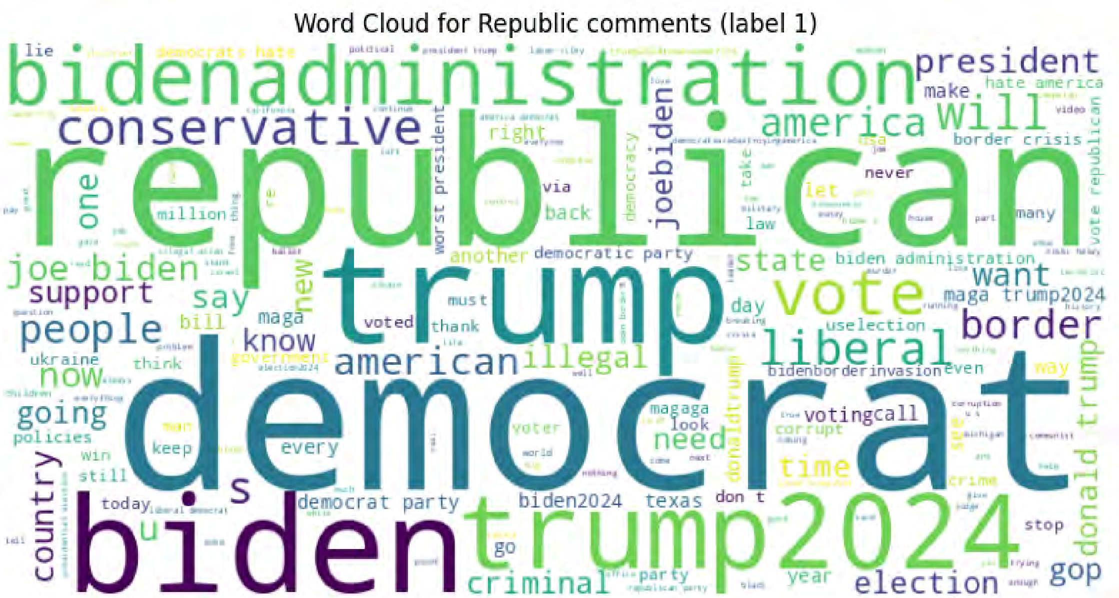


Figure 3.4.3: Common Words for Each Category

According to the findings from Fig. [3.4.3] displays popular words for every category of communal words. To make this analysis more clear, frequent but unimportant stop words like 'the', 'and', 'from' were removed from all the categories. This method provides a greater understanding of the details of political speech patterns by ensuring a more concentrated and relevant examination of the 21 unique linguistic qualities inherent to each category. The most common words with high frequency for the labels 0, 1, and 2 are 'biden' and 'trump'. In the US political context, these words act as unique identifiers. For example, "President Biden has thrown down the gauntlet, challenging Trump to cease playing politics and join him in getting the border deal approved. Your move, Trump", "Donald Trump traitor, hiding it. Never Trump #TrumpIsATraitorAndCriminal #Trump #America #USelection #NeverAgain". Additionally, words like 'democrat', 'republican', 'conservative', 'trump2024', and 'biden2024' are common words among the three labels, showcasing their significance across different political discussions.



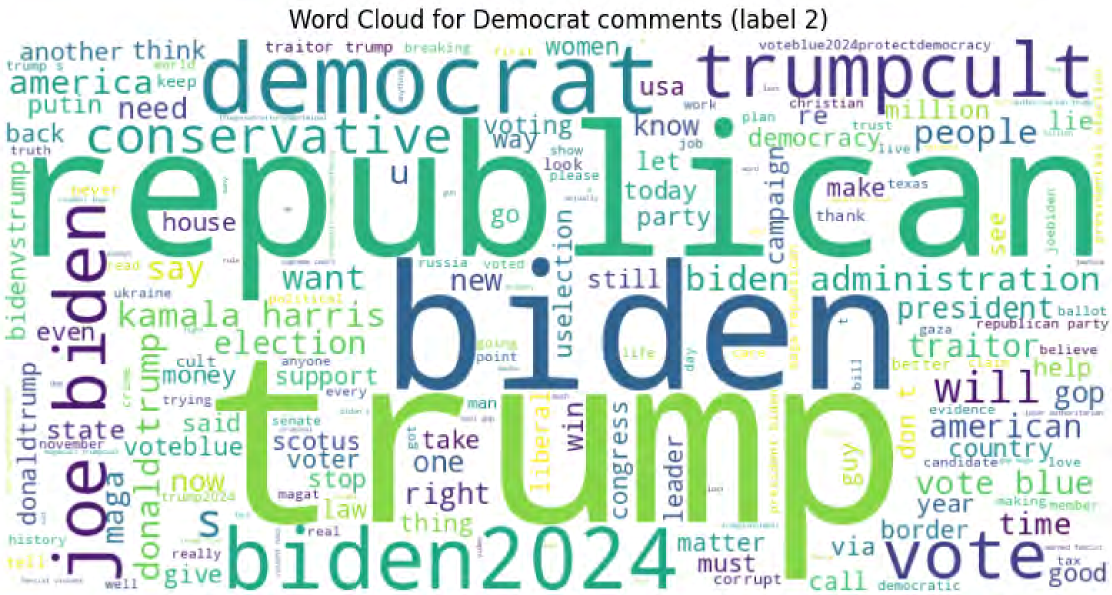


Figure 3.4.4: Word cloud

Wordcloud Fig.[3.4.4] shows unique terms that are important to every category while removing frequent but common irrelevant words like stop words. It effectively draws focus on distinctive and category-specific linguistic aspects. For instance, the words 'republican', 'democrat', 'trump', and 'biden' have been used frequently and numerous times in our dataset

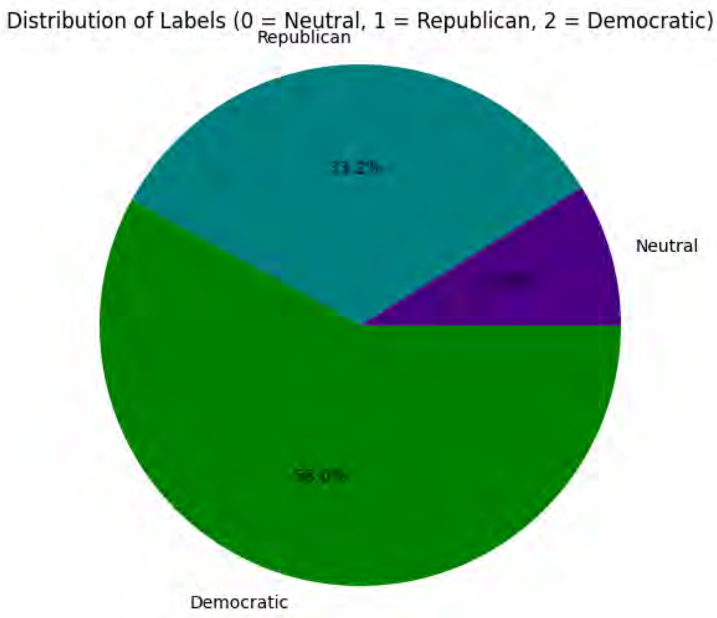


Figure 3.4.5: 15-day data Distribution of Classes

Fig. [3.4.5] shows the 15-day data from September 20 to October 4, 2024, that 8.9% of the text and comments in the dataset are classified as neutral (label 0), making up the majority of the data. The remaining 58% are identified as Democratic (label 2), and 33.2% are indicating Republican (label 1) comments.

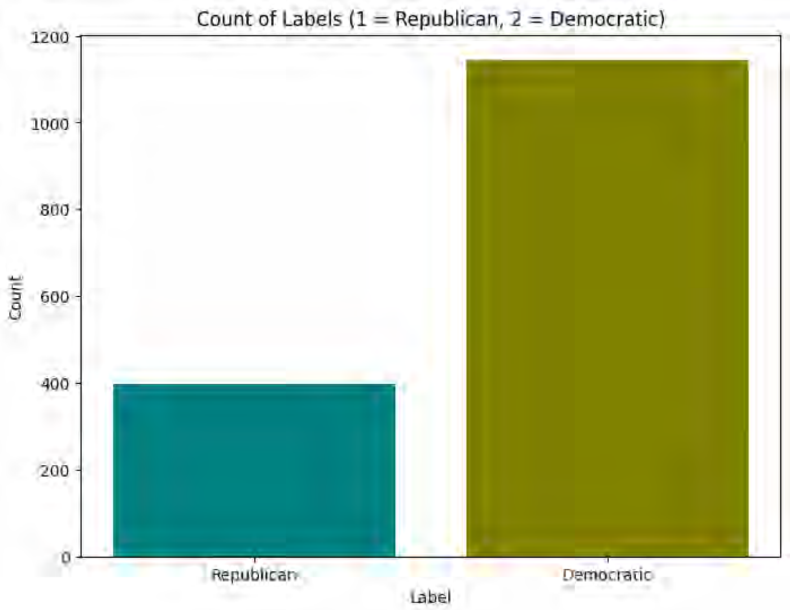
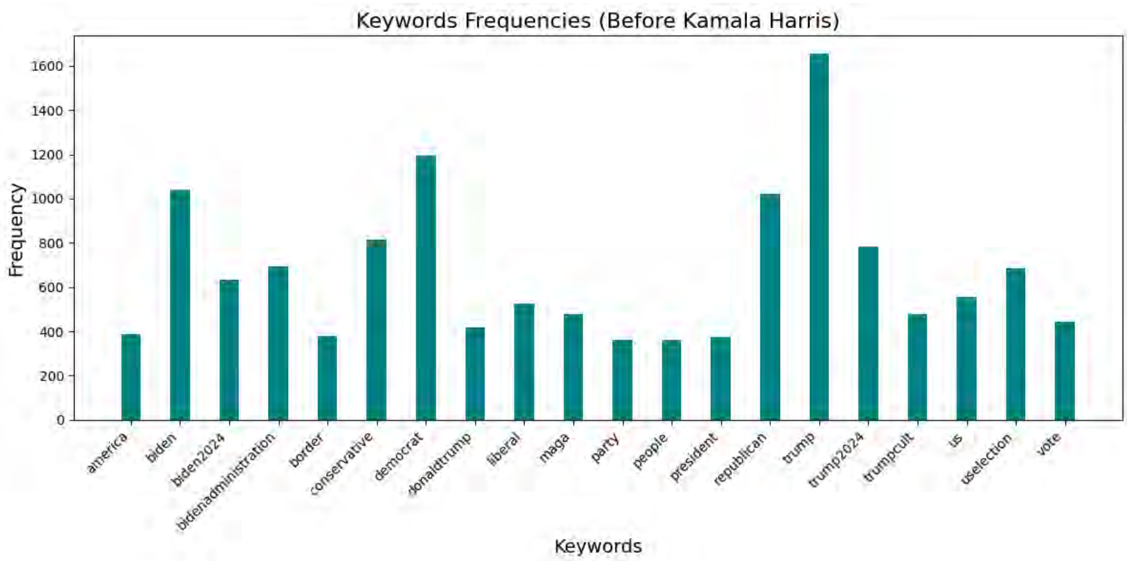


Figure 3.4.6: Data Distribution of classes after Kamala Harris came

Fig. [3.4.6] shows the data after Kamala Harris became the Democratic Party's presidential candidate, where more than 1050 comments in the dataset are classified as Democratic (label 2) and around 400 comments indicate Republican (label 1) comments.



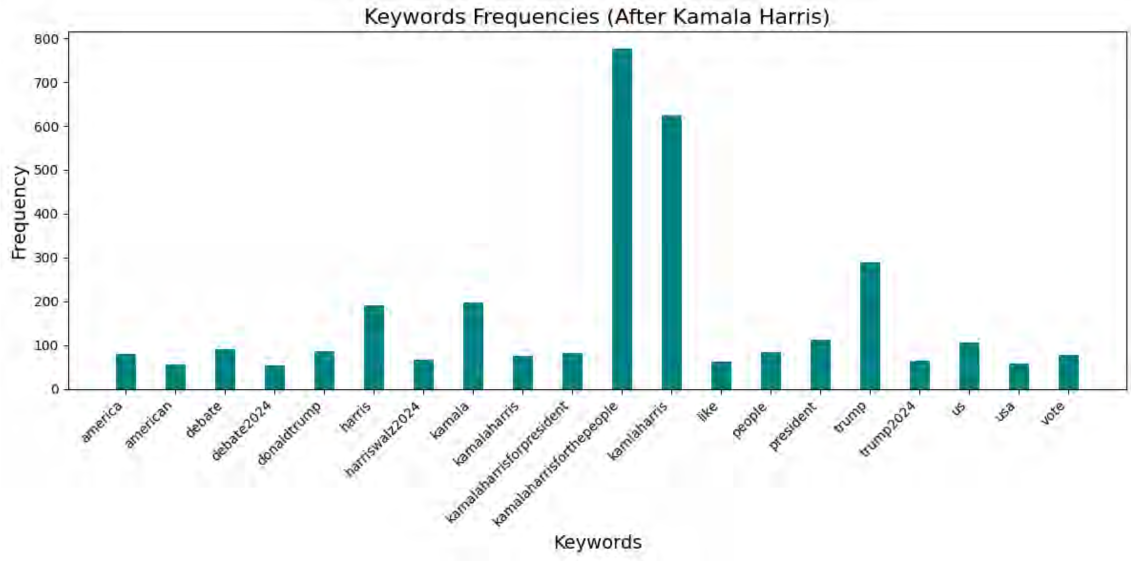


Figure 3.4.7: Most Frequent Words Before and After Kamala Harris

According to the findings from Fig. [3.4.7] , the most frequent word in the dataset before Kamla Harris elected as the Democratic Party's presidential candidate is 'trump' with a frequency of more than 1600 and 'democrat' with a frequency of 1200. As an example,'republican' and 'biden' show the approximately equal distribution of frequencies between 800 and 1000. Also, words like 'trump2024', 'conservative', 'vote', 'uselection', 'bidenadministration', and 'biden2024' display frequency distributions that range from about 400 to 600.
Then, after Kamla Harris was elected as the Democratic Party's presidential candidate the most frequent word in the dataset is 'Kamalaharrisforthepeople' with a frequency of more than 750 and 'kamalaharris' with a frequency of 600.Also, words like 'trump','kamala','harris' display frequency distributions that range from about 200 to 300. Stop words have been excluded for the most frequent word analysis of the dataset.

Fig. [3.4.8] shows from January 2024 to October 2024 Reddit political comments, where 7.5% of the text and comments in the dataset are classified as neutral (label 0), making up a minority of the data. The remaining 45.8% are identified as Democratic (label 2), and 46.8% are indicating Republican (label 1) comments.

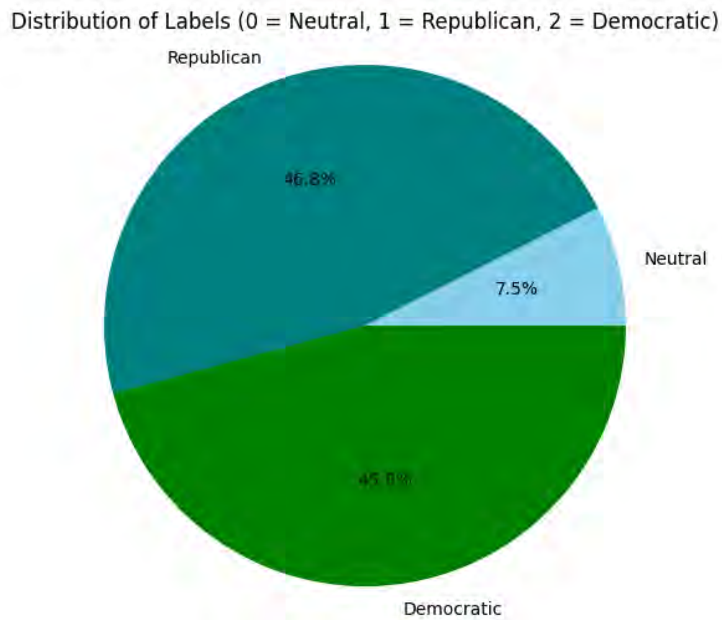


Figure 3.4.8: 2024 Reddit data Distribution of Classes

According to the findings from Fig. [3.4.9], the most frequent word in the dataset in 2024 on Reddit is ‘trump’ with a frequency of more than 1000 and ‘republic’ with a frequency of 900. As an example,‘biden’,‘democrat’ and ‘harris’ show the approximately equal distribution of frequencies between 600 and 800. The cause of most frequent word “trump” can be that the number of Republican Reddit users are higher than the Democratic supporters. Also, words like ‘us’,’trumps’,”campaign”, and ‘debate’ display frequency distributions that range from about 200 to 250.

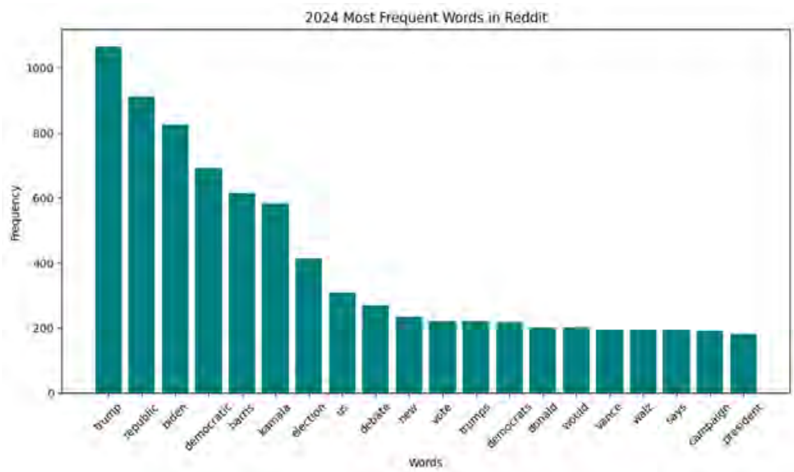


Figure 3.4.9: Most Frequent Words in Reddit

Fig. [3.4.10] shows from January 2024 to October 2024 Twitter political comments, where 21.4% of the text and comments in the dataset are classified as neutral (label 0), making up a minority of the data. The remaining 39.6% are identified as

Democratic (label 2), and 39.0% are indicating Republican (label 1) comments. The percentage of Republican and Democratic comments on Twitter is nearly the same because Twitter is a very well-known social media political discussion platform, and the numbers of Republican and Democratic supporters on Twitter are nearly the same.

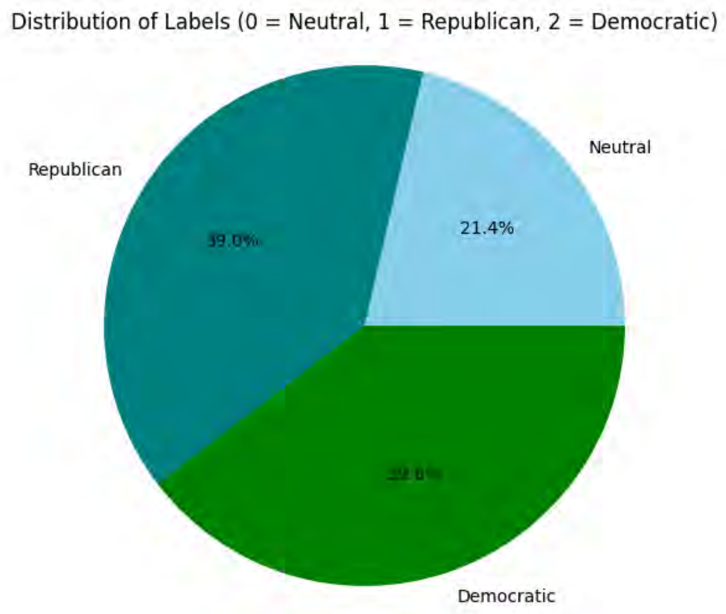


Figure 3.4.10: 2024 Twitter data Distribution of Classes

According to the findings from Fig. [3.4.11], the most frequent words in the dataset in 2024 on Twitter are 'biden' and 'trump' with a frequency of more than 2000. The cause of equal frequency frequent word 'biden' and 'trump' can be that the number of Republican and Democratic Twitter users is equal. As an example,"Kamala,","republic and 'democrat' show the approximately equal distribution of frequencies between 1800 and 2000. Also, words like 'us','trumps',"campaign", and 'president' display frequency distributions that range from about 500 to 1000.

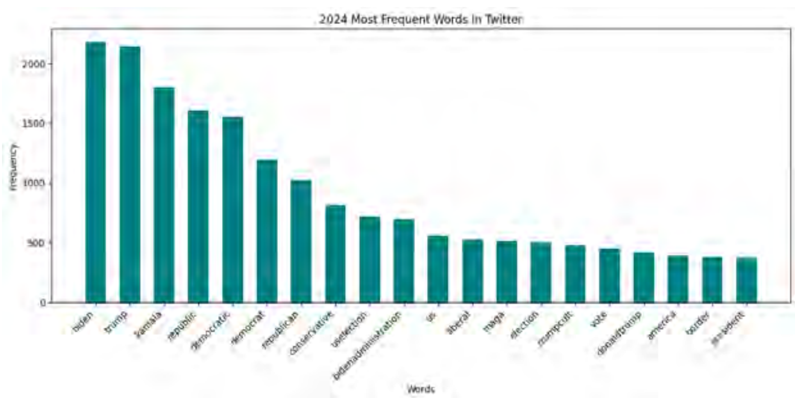


Figure 3.4.11: Most Frequent Words in Twitter

Chapter 4

Data Preprocessing

4.1 Data Cleaning

The custom-labeled dataset with nearly 10.9k comments is loaded using the Pandas program. To prevent errors in the chosen NLP model, we cleaned our dataset using the following steps-

Removing Duplicates:
First of all, we delete duplicate entries from the dataset. This stage improves the model's accuracy and helps ensure that our analysis is not skewed by eliminating unnecessary data items.

Handling Missing Values:
We then deal with missing values, which eliminates any rows that have any NaN values in them. By doing this, we reduce the possibility of biased analysis and make sure that the data we use for our later analyses is accurate.

Removing unwanted sections:
We remove characters from comments that don't contribute to their meaning to improve accuracy. Thus, we use regular expressions to get rid of URLs, html tags, hashtags, mentions, and numbers. We use the regex patterns in the re-modules to remove the emojis and replace them with empty text.

4.2 Data Preprocessing

To ensure efficient training of our selected model, we preprocess data to improve its quality, and then we visualize and analyze the results using the steps listed below-

Removing Stop Words: Eliminating Stop Words: To access English, we use the nltk library. After that, we take it out of the text data column.

Removing empty columns: Eliminating mentions, urls, stop words, and emojis may result in empty rows in the dataset. Thus, we remove any empty rows to eliminate errors in training.

Text Normalization and Cleaning: Regular expressions are used to clean and normalize text data. This entails expanding contractions to their full forms ("won't" to "would not"), eliminating URLs, keeping only alphanumeric characters and certain punctuation, condensing multiple instances of repeated punctuation to one, eliminating specific unwanted substrings, changing the text to lowercase, and replacing multiple spaces with a single space.

Part-of-Speech Tagging: Using the spacy library, we tag parts of speech in the cleaned text to determine the sentences' grammatical structure.

Named Entity Recognition: We recognize proper nouns, dates, and other important aspects in the text by extracting named entities using spacy.

Noun Chunk Extraction: To identify the subject and object of sentences, we utilize spacy methods for obtaining noun chunks from the text.

4.3 Visualization

Our aim is to visually analyze our custom-annotated dataset for deeper insights and understanding. To demonstrate the label distribution and implications, we plot graphs.

Since we want to handle multi-class classification for data that has three distinct categories for our defined labels in the dataset description separately. Our objective is to specifically identify the views present in the comments, which are labeled as follows: label-0 denotes non-identical comments or neutral, label-1 indicates Republican comments, and label-2 denotes Democratic comments. To gain deeper insights into our dataset and find linguistic patterns, we use word clouds to visually demonstrate the English terms that appear often in our political dataset. We use the word cloud technique to examine word frequency in three different categories. We create visual representations that emphasize the most frequently occurring words connected to each category by utilizing the word cloud library. This technique allows us to quickly identify and explore significant language components within our dataset, making it easier to understand and analyze the core concepts and attributes associated with each group. We also plot graphs to visually demonstrate the changes in sentiment, aggression, and popularity of each political party's supporters.

Chapter 5

Models, Experiment and Results

5.1 Models Introduction

With the goal to anticipate the probable result of the 2024 U.S. election for president by examining the consensus on networking sites such as Twitter and Reddit, our research addresses political speech analysis utilizing natural language processing (NLP) techniques. Bidirectional Long Short-Term Memory (BiLSTM) and Multi-layer BiLSTM models are good at managing sequential input and comprehending its proper context; we also used BERT, RoBERTa, and Albert models because they make use of transformer layouts with attention mechanisms, which improves their ability to capture complex contextual relationships throughout entire text sequences more effectively. These models are useful for forecasting election results according to sentiments expressed on the internet because they allow us to observe the subtleties in discourse about politics.

5.1.1 BiLSTM Model

By processing data in both forward and backward orders, Bidirectional Long Short-Term Memory (BiLSTM) Fig.[5.1.1.1] ,which is an instance of RNN (recurrent neural network) design, performs conventional Long Short-Term Memory (LSTM) networks. Because of their multi-tasking capabilities, BiLSTMs are able to extract background information through an entire information sequence. This feature is especially useful for applications like natural language processing (NLP), during which the significance of a term can be inferred through both its preceding and following terms. In NLP usage, the input layer of a BiLSTM model receives a sequence of data in the shape of phrase integration. Such embedded information, which captures lexical commonalities among terms, is represented by vectors that reflect every term in an infinite vector space. Keywords in textual information can be transformed into compact vectors using an integrated layer.

The LSTM cells are the central component of a BiLSTM. A collection of gates (forget gate, output gate, and input gate) controls how data transfers within every LSTM cell. The gate that receives input regulates the amount of newly acquired data via the stream of inputs that gets transferred to the currently active cell state. The Forget Gate establishes how much of the data preceding the prior state of the cell ought to be ignored. Determines the amount of cell state that ought to be output to the following layer via the gate at the output.With the help of these gates, LSTMs can preserve dependence over time and solve the problem of declining gradients that conventional RNNs frequently encounter. A pair of LSTM layers handle the supplied contents in a BiLSTM. The content is processed by one set in a forward motion (from beginning to end) as well as by another set in a reverse way (from the end to the start). As an outcome, interconnections inside the process can be represented by the system using current and potential perspectives. Concatenation of the forward and reversed LSTM layers' responses results in an accurate portrayal of the process scenario during every single phase. The forward and backward LSTM layers' signals are combined (typically by stringing). By capturing data from every angle, this amalgamated result improves the characteristic description. Following the initial LSTM layers, layers for dropouts may be inserted to avoid excessive fitting.

During every training modification, dropout arbitrarily turns a portion of the input values into zero, assisting in reorganizing the framework. The densest layer (or layers) connects the intended outcome to the extensive depictions of features that are acquired through the BiLSTM layers. This stage usually concludes by presenting a softmax or sigmoid activation function, commonly used in sentiment evaluation, which provides probabilities for multiple mood classifications such as positive, negative, and neutral. Finally, results are produced by the output layer. It may provide an individual probability score for binary sentiment analysis, reflecting the possibility that the overall mood is favorable or adverse. It may provide an estimated dispersion throughout many sentiment classifications, enabling multi-class sentiment estimation.

BiLSTMs are far better at grasping the significance of each word in statements where the situation is important because they are capable of comprehending context in all directions. BiLSTMs employ LSTM cells to handle dependence over time efficiently. This helps to preserve pertinent data across extended durations and reduce problems such as a gradient that disappears. Since the model has a deeper comprehension of what is happening, processing events in each direction usually results in enhanced accuracy in activities like analyzing emotions. In addition to the analysis of sentiment, BiLSTMs can be used for a wide range of synthesis operations, such as recognizing speech and automated translation. BiLSTM models are especially useful for analyzing sentiment. We may identify modest sentiments from the expressions encompassing us, according to the significance that they are intended to provide.Discussions along with additional contextually dependent events play a crucial role when assessing sentiment, while BiLSTMs are superior at comprehending their effects.The ability of the model to generalize from the training data to previously unknown samples is enhanced by the improved depictions of features derived by bilateral analysis.

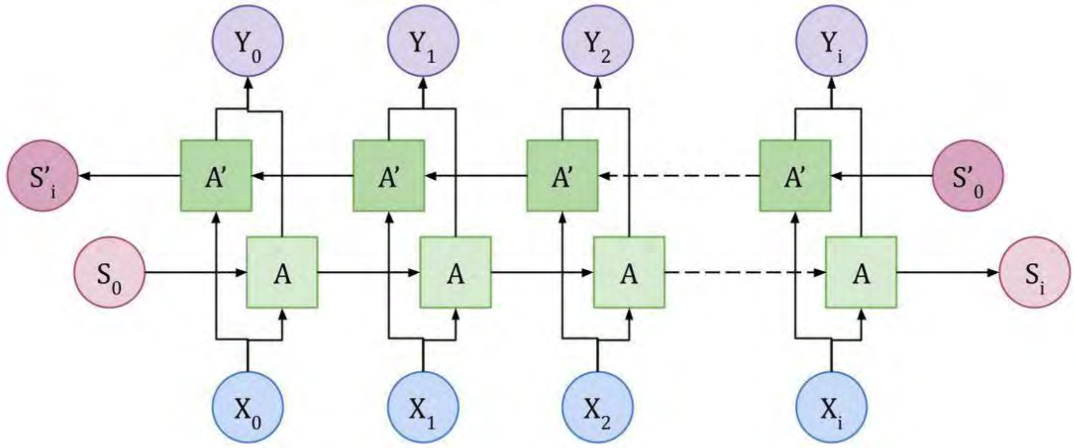


Figure 5.1.1.1: Bidirectional-LSTM

5.1.2 Multilayer BiLSTM Model

A Multilayer Bidirectional Long Short-Term Memory (Multilayer BiLSTM) model Fig.[5.1.2.1] comprises several separate units of bidirectional LSTM (BiLSTM) layers . The BiLSTM unit analyzes the sequences of input in both directions simultaneously , while the stacked layers permit the system to acquire ordered characteristics using datasets. The ability of the model to recognize and evaluate intricate trends in series is improved by its extensive architecture, which enables it to extract ever more conceptual and intricate characteristics from the information that is provided. This method works especially well for activities like analyzing sentiment, in which the precision of estimations can be greatly affected by catching minute details in speech. In tasks involving NLP, the input layer generally gets the order of the data as word embeddings. Words are represented as highly packed vectors in an infinite vector domain through these embedded data. Text may be represented as dense vectors using an embedding structure. For creating vibrant, complemented word vectors, pre-trained embeddings such as Word2Vec, GloVe, or BERT embeddings can be used. Numerous layers on top of BiLSTM modules make up the Multilayer BiLSTM design's central component.

Every BiLSTM layer captures dependence on both past and future contexts by processing the given input order both forward and backward. During every single move, the first BiLSTM layer generates both forward and backward concealed states after receiving the starting input order. The final result of the initial BiLSTM layer is formed by concatenating several concealed states. The combined results of the preceding layer provide the input for each BiLSTM layer that comes after it. Each layer may acquire further intricate characteristics as well as more advanced concepts using the input because of its organizational layout. The input, forget, and output gates found throughout every LSTM cell of the BiLSTM layers control the data movement, allowing the model to preserve dependence over time and handle the issue of disappearing gradients. To lessen excess fitting, layers of dropouts can be inserted among BiLSTM layers. To assist in regularizing the framework, dropout operates by arbitrarily changing a portion of the input quantities to zero at every conditioning phase. The results generated by the forward and reverse orientations are combined at every BiLSTM layer. Complete background data is captured in the two perspectives by this composite depiction.One or more thick layers are placed after the last BiLSTM layer. Combined BiLSTM results are fed through those completely linked layers, which convert the results to the appropriate structure for the ultimate forecast. This usually requires a couple of robust layers in the analysis of sentiment, which result in a softmax or a function called the that provides estimates for various emotions. Utilizing the results obtained from the full layer, the output layer produces the ultimate estimations. The result for multimodal classification of sentiment could be an individual sentiment-indicating possibility ratio. The distribution of probabilities over several sentiment classes might indicate the result of multi-class sentiment categorization. By layering numerous BiLSTM layers, the neural network can acquire broader and more complicated attributes, leading to better pattern recognition.

An additional layer enables the framework to more accurately reflect subtle backdrops and far-reaching connections, which is important for applications including sentiment analysis, whereby the implications of individual phrases may vary depending on their distance from one another. Multilayer BiLSTMs are structured so that they can operate faster upon a variety of analyzing initiatives, especially sentiment analysis, because the algorithm can take advantage of rich contextual data. Dropout segments are inserted across BiLSTM layers to minimize excessive fitting, which strengthens the framework and increases its ability to generalize to new evidence.

Multi-layer BiLSTM models provide several benefits regarding sentiment analysis. Because of their sophisticated design, these algorithms can recognize sentiment patterns that more straightforward approaches could ignore. The model's comprehension of a variety of phrases and circumstances is improved by its capacity to handle episodes within forward as well as backward orientations across several levels. This improves the precision of sentiment prediction. Multilayer BiLSTMs may efficiently deal with long text sequences by identifying relevant data dispersed throughout a significant amount of text.

Our work combines the benefits of multi-layer BiLSTM models to precisely evaluate the collective sentiment expressed on social media platforms like Reddit and Twitter. The study's findings will help forecast the most likely outcomes of the 2024 presidential election in the United States since they provide valuable data on general public opinion patterns.

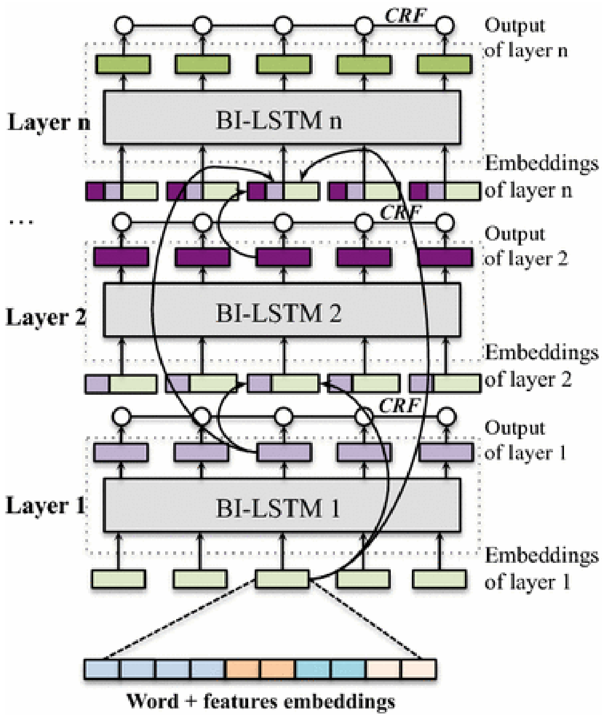


Figure 5.1.2.1: Multilayer BiLSTM Model

5.1.3 BERT

In this paper, we use the Bidirectional Encoder Representations from Transformers (BERT) model to anticipate voter inclinations in the upcoming US presidential election to analyse public sentiment. Google created BERT, a cutting-edge language model that has transformed Natural Language Processing (NLP) tasks by excelling previous approaches in understanding contextual keyword embedded data. Because BERT is bidirectional, it can analyse a word's context in both its left and right contexts, which improves its performance in sentiment analysis tasks. BERT performs better at recognizing sophisticated language,a crucial skill for political discourse analysis. BERT's contextual knowledge facilitates in the decoding of sentiments, conflicting emotions, and implied views that are very frequent in public posts on Reddit and Twitter. Because the model has been trained on language earlier, it can be used to a wide range of problems, including political speech analysis. The transformer design, which is dependent on a concentration process, is the cornerstone of BERT (Bidirectional Encoder Representations from Transformers).It has simultaneous interpretation of the entire data series.The BERT-Base and BERT-Large models are composed of 12 and 24 stages, respectively, of transformers.

There are two essential parts to every transformer section. Multi-Head Self-Attention technique enables BERT to focus on various phrase segments while recognizing word associations at various points. Every "head" approaches the statement from a distinct angle.Feed-Forward Neural Networks carry out non-linear transformations, each token's representation passes across an entirely linked feed-back network following self-attention. BERT processes text in both directions, in contrast to systems such as GPT that only analyse words in one direction (from side to side or from centre to left). This is important for operations like detecting emotions and question answering since it enables comprehending meaning from both viewpoints of a word in an expression.Three types of embedded data are used by BERT for tokens. Embeddings of tokens express each word or token separately.Segment Embeddings declare if a token comes from text A or text B (helpful for addressing questions, for example).Positional Embeddings, since the transformer does not automatically comprehend word order, make sure to record the placement of every word in the text.

The primary benefit of BERT is derived from its pre-training and fine-tuning procedures. At first, BERT is pre-trained using two unsupervised tasks on a big dataset of text, such as Wikipedia and BooksCorpus.Language Modeling using Masks (MLM) trains the model to predict words that have been randomly masked, representing a fraction of the input tokens. This compels the model to pick up on words' bidirectional contexts.Predicting the Next Sentence (NSP) of the model to comprehend the connection between two texts is crucial for tasks like conversation or natural language interpretation.Following pre-training, BERT is trained on a particular task, such as question replying, sentiment analysis, or labelled data identification. Because BERT has developed general language representations through pre-training, fine-tuning is quicker and more effective.The task-specific data is fed via BERT during fine-tuning, and its settings are adjusted to suit the current job. Because of its excellent language understanding abilities, BERT is frequently utilised for a range of Natural Language Processing (NLP) tasks. BERT is perfect for evaluating social media postings, assessments, or political conversation since it can categorise content based on sentiment.Many QA infrastructure, like Google Search, rely on BERT to select the most pertinent response from a page when a user inputs an inquiry.It helps in recognizing and categorising textual objects (such as individuals, locations, or events).Text categorization is the process of grouping texts into distinct groups for purposes like identifying spam, topics, or phoney information. Also to enhance the accuracy of interpretations, translation systems can employ BERT as a language model.For activities like summarising, in which knowing how sentences relate to one another is crucial.

BERT also has certain drawbacks.BERT needs a lot of processing strength, particularly while it's being pre-trained and fine-tuned. High-end GPUs and lots of storage are needed for fine-tuning on an enormous data set or using BERT in real life operations. BERT may operate slowly when making predictions on newly acquired data during reasoning, especially for large-scale or real-time jobs, because of its expansive structure and bidirectional approach. Limited Length Handling: The largest amount of input tokens that BERT can process is 512. Any content that is longer than that must be divided or shortened, which might result in the loss of essential data, particularly when working on lengthy papers. Although BERT is adept at recognizing the wider context of a phrase, it lacks substantial factual knowledge and reasoning skills. Without any logical justification or understanding, it just makes predictions according to sequences it has already observed. BERT attempts to anticipate phrases by masking them arbitrarily during the pre-training stage. This does not, however, necessarily correspond to jobs in the actual world where entire phrases are frequently used. As a result, its training goal could not always be consistent with some real-world NLP tasks.BERT is susceptible to adversarial scenarios, in which minute modifications to the input (such as synonyms or misspellings) drastically affect the forecasts made by the model.

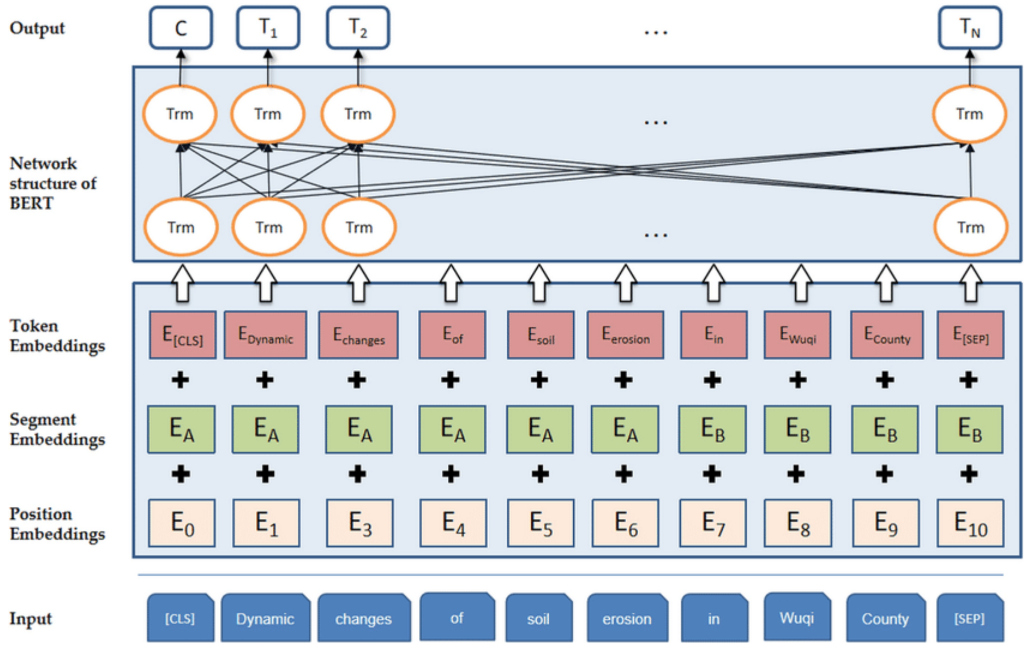


Figure 5.1.3.1: BERT

5.1.4 RoBERTa

Robustly Optimised BERT Pre Training Method, or RoBERTa, is a potent BERT model variation that aims to greatly enhance BERT's initial training and fine-tuning techniques. RoBERTa, which was created in 2019 by Facebook AI (now Meta AI), aims to improve different facets of BERT's method of training, producing better outcomes in a variety of natural language processing (NLP) tasks, including text categorization, question answering, and sentiment analysis. Although RoBERTa's structure is similar to that of BERT, it adds some significant advancements that distinguish it when it comes to precision as well as efficiency.Through pre training process optimization, RoBERTa improves on BERT. It makes use of bigger stages, lengthier patterns, and greater training data without BERT's Next Sentence Prediction (NSP) task. Further reliable outcomes in subsequent tasks follow from this. RoBERTa was incorporated into our study to enhance prediction accuracy by capturing more accurately linguistic patterns from Reddit and Twitter data. Using a significantly wider training dataset is one of the main enhancements that RoBERTa brings. While RoBERTa trained on 160GB of data, BERT was pretrained on a 16GB dataset mostly consisting of text from BookCorpus and Wikipedia. 'WebText' (material taken from Reddit), 'OpenWebText' (an free version of GPT-2's WebText), 'CommonCrawl News'(a dataset of 63 million news items), and 'Stories Dataset' (a database containing narrative-based text) are some of the data streams used in RoBERTa.RoBERTa's capacity for extrapolation to different NLP tasks is enhanced by this wide and massive dataset, which allows it to gather data from a wider range of language systems, scenarios, and patterns.

RoBERTa introduces an adaptive disguising method to enhance BERT's 'Masked Language Modeling' (MLM). When using BERT, the masked tokens that the model predicts are chosen only once at the beginning of the training phase and don't change. RoBERTa uses dynamic masking, which means that whenever the model comes across the same phrase, an alternate token is masked. Because RoBERTa successfully trains on numerous copies of the exact same text, with distinct concepts masked every time, it is able to acquire stronger depictions of the language. Two goals of BERT's pretraining were next sentence prediction (NSP), which teaches the system to forecast whether one sentence will come after a different one, and masked language simulation. Facebook AI's studies, meanwhile, revealed that eliminating the NSP task enhanced the accuracy of the model. Without an interruption of sentence-level predictions, RoBERTa removes NSP and concentrates solely on the MLM target, enabling it to focus on word-level as well as knowledge of the context. It has increased batch size and training time.In comparison to BERT, RoBERTa makes use of significantly bigger mini-batch lengths and greater training times. RoBERTa, for instance, utilises batches with capacities of up to 8,192, whereas BERT was trained with sample sizes of up to 256. This makes it possible to simulate the gradient more effectively, which promotes greater convergence and improved applicability over NLP tasks. Furthermore, RoBERTa had been trained for more than 500,000 steps (as opposed to BERT's 1 million steps), which improved its capacity to absorb knowledge gathered from the collected data. Tokenization is done by RoBERTa using "byte-level Byte-Pair Encoding (BPE)", which helps the model do a better job of handling unusual phrases and subwords. RoBERTa's byte-level technique enables it to more effectively organise phrases that occur rarely or are complicated, which improves its efficacy on languages and situations with varied terminologies, whereas BERT uses WordPiece tokenization. RoBERTa achieved the highest standards on multiple NLP standards, demonstrating its excellent efficiency.

GLUE (General Language Understanding Evaluation), A series of exercises designed to assess a model's capacity to comprehend natural language. In this comparison, RoBERTa significantly surpassed BERT.RoBERTa broke consistency records in SQuAD (Stanford Question Answering Dataset) task by performing exceptionally well at responding to queries based on the provided text.RoBERTa did exceedingly well on RACE (Reading Comprehension from Examinations) , a difficult benchmark, which tests comprehension of texts from high school and middle school English tests. RoBERTa is ideal for a variety of NLP work due to its exceptional language modelling features.RoBERTa is helpful in scenarios such as news categorization, detecting spam, and article monitoring since it can classify articles or text extracts into specified classes.RoBERTa can ascertain the mood conveyed in social networking content feedback, and other types of written information by examining big text corpuses.RoBERTa may retrieve and understand the required material from huge texts or databases of knowledge to deliver appropriate responses in tasks such as providing client service depending on preexisting facts.RoBERTa is helpful in Named Entity Recognition (NER), when conducting data extraction work since it can reliably recognize objects in written content, such as dates, names, and locations. Due to its improvements, RoBERTa is an effective tool for natural language processing (NLP) jobs, particularly when complex language comprehension is needed. Through the use of bigger datasets, dynamic masking, and more thorough training, RoBERTa builds on the framework established by BERT. Because of these advances, RoBERTa is now widely used in both business and academia to handle challenging linguistic jobs more accurately and effectively.

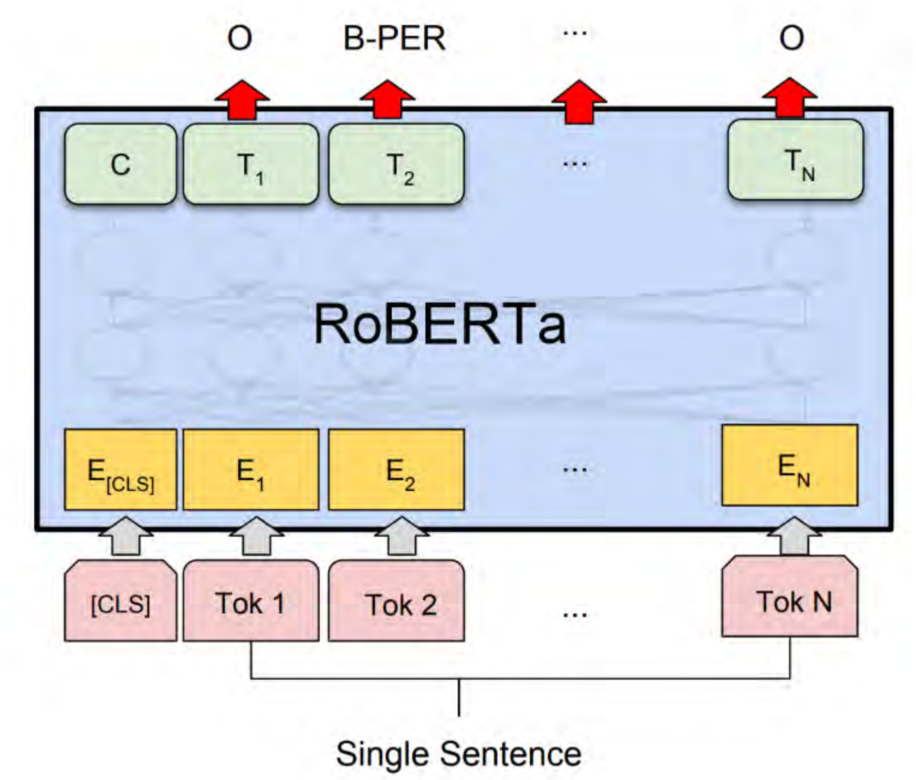


Figure 5.1.4.1: RoBERTa

5.1.5 ALBERT

ALBERT, also known as A Lite BERT, is a BERT model variation that is intended to minimise the model's size without sacrificing its great effectiveness for tasks involving NLP. ALBERT, created by the researchers at Google, aims to overcome BERT's storage and processing limitations by implementing specific enhancements without compromising precision. With fewer variables than BERT, these enhancements enable ALBERT a better fit for a range of uses in natural language processing.We used ALBERT, a model created to lessen BERT's space and temporal complexity without sacrificing efficiency, to solve its operational shortcomings. By adjusting the embedded matrix and sharing variables between layers, ALBERT does this. Through this approach, we were able to deal with bigger datasets effectively while maintaining forecast accuracy.The method that ALBERT organises its embedding layers represents one of the main modifications.

There are a lot of variables in BERT because of the close relationship between the dimension of the list of words embedding and the layers that are hidden sizes. By projecting phrases into lower-dimensional regions and then projecting them onto higher-dimensional hidden layers, ALBERT splits both of these processes. This decomposition drastically lowers the amount of space needed and the total amount of features without compromising the efficiency of the algorithm. When more transformer layers are added, BERT employs separate sets of characteristics for each one, which significantly expands the model's complexity. ALBERT solves this through using one set of attributes shared by all layers, as opposed with every layer maintaining an entirely distinct set of attributes. Because of this collaboration, the model's ability to develop and adapt throughout layers is maintained, despite a large decrease in the overall number of features. ALBERT presents 'sentence-order prediction (SOP)' as an initial training target, in contrast to BERT, which utilises the next sentence prediction (NSP) task. Sentence consistency and rational sequencing are critical for numerous NLP duties, including comprehending texts and long-form analysis of texts, and SOP aims to strengthen the framework's knowledge of these concepts. SOP anticipates if two subsequent phrases are in the right sequence or may have switched, as opposed to NSP's prediction of when a single phrase precedes another. Sentence-level comprehension improves as a result of this minor adjustment.

On NLP standards, ALBERT gets high scores even though it has fewer parameters. For example, because of the recently added improvements, the ALBERT-large model has considerably less parameters than BERT-large but operates comparably throughout various workloads. For situations where storage and processing speed are critical, such as distributing simulations on handheld gadgets or in real-time environments, this makes ALBERT more attractive. Because ALBERT uses factorised embedding and parameter collaboration, it can expand to huge structures effectively without drastically raising the number of parameters. Because of this, scholars are able to train bigger models with larger amounts of information while maintaining quite cheap expenses for processing.

ALBERT has produced state-of-the-art scores on numerous NLP standards, despite being lighter than BERT. A number of aspects of natural language understanding are measured by the "GLUE" (General Language Understanding Evaluation) activities, such as linguistic implication, emotion categorization, and paraphrase recognition. ALBERT did great on these assignments.ALBERT outperforms BERT on the "SquaAD 2.0" (Stanford Question Answering Dataset) competency level when it comes to solving problems pertaining to a particular written content.The difficult benchmark known as "RACE" (Reading Comprehension from Examinations) evaluates an algorithm's ability to reply to questions from language exams. ALBERT performed admirably in this test. It includes a lot of parameters, with values that range from 110 million (BERT-base) to 340 million (BERT-large) and cuts the total number of characteristics substantially, with models that vary from 235 million (ALBERT-large) to 12 million (ALBERT-base).Bigger parameter numbers result from the embedding length being identical as the layer that is concealed size.ALBERT utilise factorised embedding parameterization, which reduces the number of variables by separating the embedding size from the hidden layer's length,doesn't have parameters, allowing every single layer to have a particular set of parameters, carries out cross-layer combining parameters, that utilises the same parameters throughout several layers to cut down on the total number of factors.

It also utilises two pre-training targets, namely Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). Sentence Order Prediction (SOP) is used as a substitute of NSP and concentrates on predicting sentence-level consistency by checking if the two phrases belong to the correct sequence. It is particularly memory-efficient due to combining parameters and less integrating sizes, thus rendering it appropriate for circumstances with a limited amount of RAM. It requires more time for training than it should due to its greater amount of parameters. It trains rapidly since it has less parameters and utilises memory effectively.

Because to its effectiveness,ALBERT may be customised for activities like sentiment estimation, identifying spam, or information classification. It is ideal for a range of natural language processing uses, particularly in circumstances in which computing facilities are scarce.When it comes to providing precise answers according to situational comprehension, ALBERT shines in chatbots and artificially intelligent assistants.With its ability to analyse vast amounts data text, ALBERT is far better when used in manufacturing settings, producing brief overviews.Considering the additional benefit of lower computing costs, ALBERT may be utilised to locate dates, names, and other important elements in an item. Due to advancements like sharing variables and factorised embeddings, ALBERT is a highly effective substitute for BERT that produces results that are equivalent while using a significantly smaller number of parameters. It addresses the storage and computational problems related to massive models while maintaining the transformative architectural capacity to manage challenging linguistic comprehension applications. ALBERT becomes a more adaptable instrument for use in NLP by adding the sentence-order forecasting challenge, which strengthens its ability to comprehend sentence-level consistency. Because of its effectiveness, it is especially appropriate for situations where scalable strategies are needed, such as distributing models in contexts with limited resources.

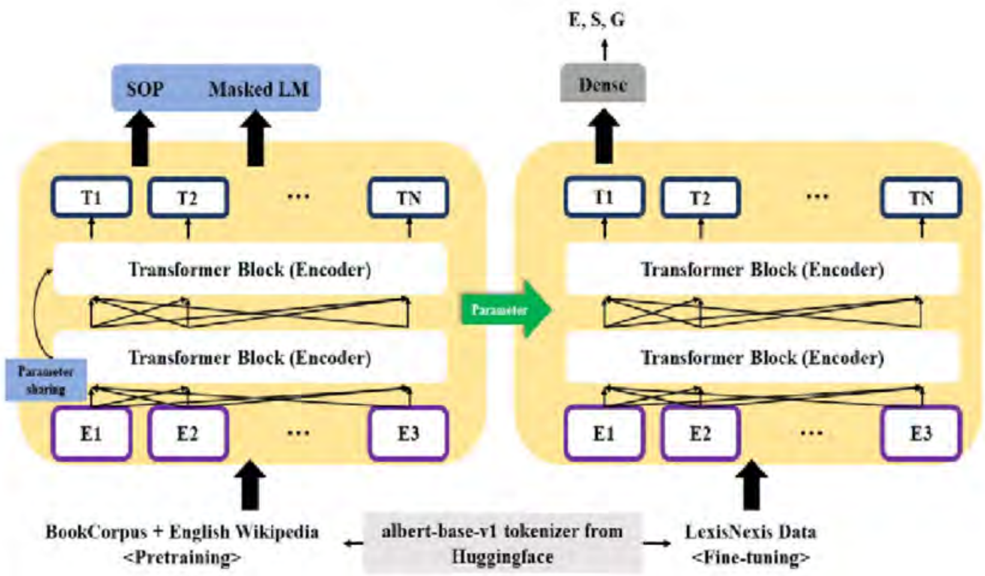


Figure 5.1.5.1: ALBERT

5.2 Multi-Class Classification

Data Split: We partition our dataset using a cross-validation process to avoid any form of overfitting. We separated the dataset into train and test sets, with the train containing 85% and the test containing 15% of the original dataset. We evaluate BERT, RoBERTa, and ALBERT models separately on short and long text, using a threshold of 20 words to distinguish between them. Additionally, we train BiLSTM, multilayer BiLSTM, BERT, RoBERTa, and ALBERT models without separating the texts by length to compare their overall performance.

Tokenize: We tokenize both the training and testing sets, utilizing various techniques to properly arrange our data for input into our BiLSTM, Multilayer BiLSTM, BERT, RoBERTa, and ALBERT models. Using the RegexpTokenizer, we break text up into words to give us flexibility when working with various kinds of data. To ensure that our models work with text, we also convert text into a string of integers using TensorFlow's Tokenizer. These token sequences are padded to provide constant lengths across data samples. We organize our tokenized data into a PyTorch dataset to enable seamless interaction with PyTorch's data loader for efficient model training and validation.

Training and Evaluating: On our dataset, we train the BiLSTM and Multilayer BiLSTM models for labels 0, 1, and 2. On the other hand, we train BERT, RoBERTa, and ALBERT models for labels 1 and 2. Custom training variables are set, and we adjust them in response to the performance of the model on the test dataset. We use accuracy, precision, recall, F1 score, and loss to evaluate our models. We present the results on accuracy and loss graphs across epochs and a confusion matrix to illustrate it for both models.

5.2.1 BiLSTM

Architecture

We start our experimentation with a Bidirectional Long Short-Term Memory (BiLSTM) model. BiLSTM is an effective approach for text classification tasks like detecting political bias in social media comments because it captures dependencies in both forward and backward directions, which makes it suitable for sequence modeling uses.

The comments are first tokenized and transformed into dense vector embeddings. These embeddings are passed into the BiLSTM layer, which processes the text in both directions, allowing the model to capture context from both past and future words. To avoid overfitting, the output from the BiLSTM layer—which mixes forward and backward hidden states—is then run through a dropout layer. Finally, a fully connected linear layer processes the output and calculates the logits for each class (Democratic, Republican, or Neutral). By reducing the loss between predicted and true labels during training, the algorithm is better able to identify political bias in comments on social media.

Performance Evaluation

The figure's loss function Fig.[5.1.1.1] shows a clear pattern of the BiLSTM model: the validation loss gradually increases as the number of epochs increases, whereas the training loss consistently decreases over the training process. It is evident from this significant difference between training and validation loss that overfitting is occurring in the model. Besides, after a few epochs, the validation accuracy reaches a plateau and starts to decline significantly, but the training accuracy continues to increase. This divergence between training and validation accuracy suggests that the model is overfitting, where it learns the training data well but fails to generalize to new, unseen data

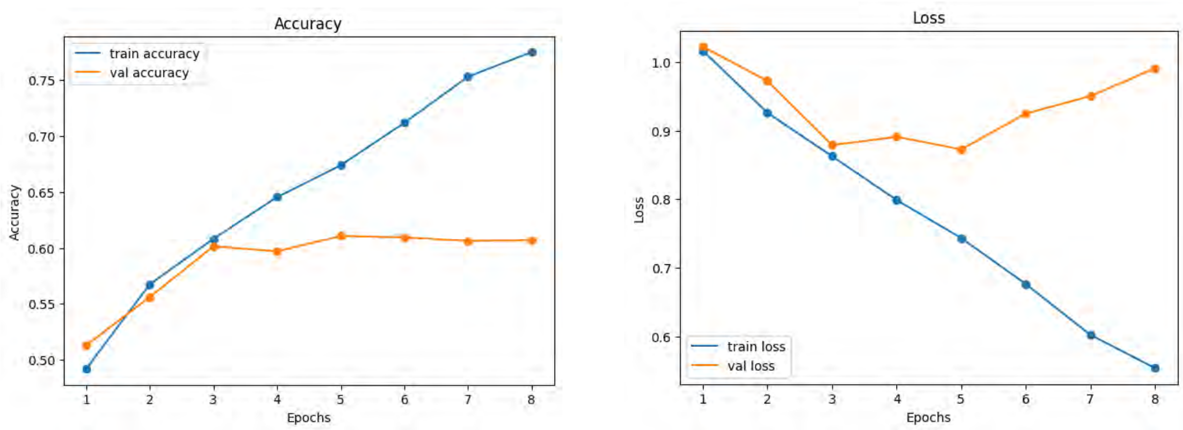


Figure 5.2.1.1: Loss and accuracy graph for BiLSTM model

By evaluating our BiLSTM model, we observe that the model provides a notable accuracy of 61.06%, indicating that it can accurately categorize comments. Even still, the loss score of 0.8726 shows that there is still room for development in terms of reducing errors made during training. Furthermore, the model's overall performance is demonstrated by the F1 score of 0.612, which shows a respectable balance between recall and precision. Precision and recall scores of 0.612 suggest a consistent ability to handle false positives and false negatives. A thorough examination of the confusion matrix Fig.[5.2.1.1] would offer insightful information about the model's advantages and potential areas for improvement, allowing for a greater understanding of its behavior.

Significant points are shown by looking at the confusion matrix in Fig.[5.2.1.2]. Class 0 examples are identified by the model with excellent accuracy, as seen by the topleft area of the matrix, which has 493 correct predictions. On the other hand, a number of off-diagonal elements—mostly in the middle and bottom rows indicate misclassifications. The model indicates a significant amount of misunderstanding for class 1, with 193 cases incorrectly assigned to class 0 and 86 to class 2. In the same way, 116 instances of class 2 are incorrectly classified as class 0 and 59 as class 1. This indicates regions where the model had difficulty accurately categorizing cases for these specific categories.

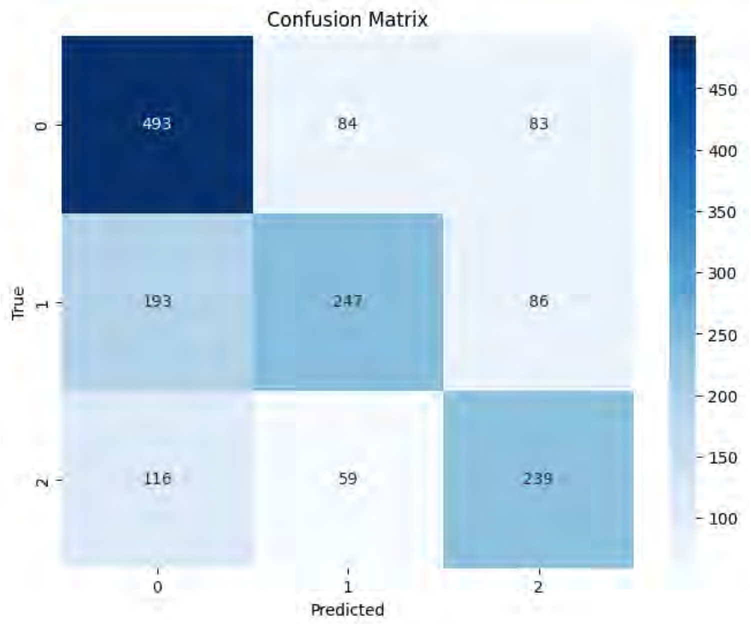


Figure 5.2.1.2: Confusion Matrix BiLSTM model

5.2.2 Multilayer BiLSTM

Architecture

To improve the categorization of social media comments into Republican (1), Democratic (2), and Neutral (0) categories, we applied a multilayer Bidirectional Long Short-Term Memory (BiLSTM) model. By layering numerous BiLSTM layers, this multilayer configuration enables the model to capture more complex patterns and dependencies, leading to a higher understanding of the sequence.
The input text is tokenized, and then embeddings are created and run through several BiLSTM layers. Richer contextual information is captured at every stage as each layer processes the text both forward and backward. The output of the last BiLSTM layer is first regularized by a dropout layer before being processed by a fully connected layer that calculates logits for every class.Deep contextual knowledge is used in this design to increase classification accuracy.

Performance Evaluation

The figure's loss function is Fig.[5.2.2.1] the graph shows a clear pattern of the Multilayered BiLSTM model: the training loss gradually decreases as the model learns and improves its performance on the training data. As the training loss gradually decreases, the model is picking up new skills and becoming more proficient with the training set. On the other hand, the validation loss starts increasing and stabilizes after first decreasing and reaching a low around the third epoch. This divergence points to the possibility of overfitting, in which the model works well on training data but finds it difficult to apply its predictions to unseen validation data. Besides, the training accuracy gradually rises, reaching approximately 75% by the fifth epoch. However, after rising initially and peaking during the third epoch, the validation accuracy starts to decline after fluctuating slightly. It means that the model begins to overfit after the third epoch since it loses generalization on the validation set but performs better on the training data.

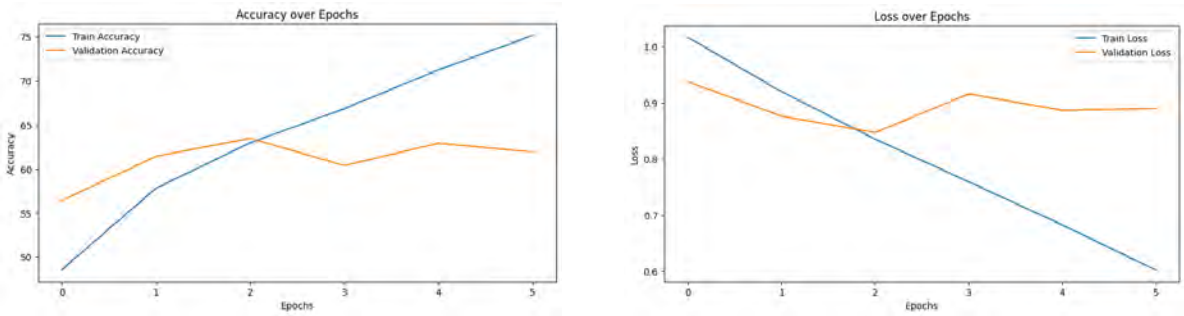


Figure 5.2.2.1: Loss and accuracy graph for Multilayer BiLSTM model

By evaluating, our Multilayer BiLSTM model, we find that it can reliably classify comments with an accuracy of 63.48%. However, a loss score of 0.8470 indicates that training errors can be decreased. The F1 score and recall score of 0.61 indicates consistent handling of false positives and negatives, while the model's precision of 0.62 shows a fair balance between recall and accuracy. Observing the confusion matrix Fig.[5.2.2.2] highlights important details. For class 0, the model shows good accuracy, with 449 accurate predictions. There are also significant misclassifications, though: 142 cases in class 1 are incorrectly categorized as class 0 and 56 as class 2, 32 whereas 93 cases in class 2 are incorrectly classified as class 0 and 109 as class 1. This indicates areas where the model had difficulty accurately categorizing cases for these specific categories.

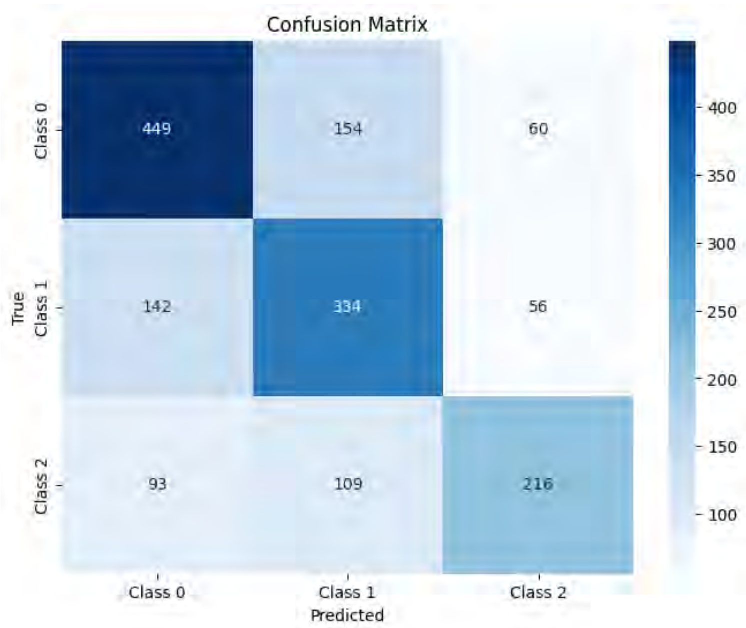


Figure 5.2.2.2: Confusion Matrix Multilayer BiLSTM model

5.2.3 BERT

Architecture

To categorize social media comments into Republican (1) and Democratic (2) categories while excluding neutral (0) comments, we have used a Bidirectional Encoder Representations from Transformers (BERT) model. Input text was transformed into tokens using BERT's pre-trained tokenizer, and these tokens were then mapped to dense embeddings.
The embeddings were passed through BERT's multi-layered bidirectional transformers, capturing both forward and backward context in the text. To classify, we have used the [CLS] token's pooled output, which captures the context of the entire sequence.The logits for the Republican and Democratic labels are computed by passing the pooled [CLS] token output through a fully connected layer. The BERT model successfully identified between the two political categories according to the text's context by reducing the loss between predicted and true labels.

Performance Evaluation

The BERT showed noticeable results with 0.85684211 accuracy, 0.85499135 precision, 0.85423787 recall, and an f1 score of 85.459882. The figure's loss function is Fig.[5.2.3.1] the graph shows the training loss gradually drops as the number of epochs increases, suggesting that the model is improving at reducing errors on the training set. The validation loss, on the other hand, shows an initial decrease. This growing gap between the training and validation loss highlights that the model is overfitting. Besides, the training accuracy in this BERT model accuracy graph increases significantly from about 75% in the first epoch to about 98% by the fifth epoch, suggesting that the model is doing very well of learning the training data. But in the early epochs, the validation accuracy increases a little, peaks at about 86% in the second epoch, and then starts to significantly decrease. As the model's performance on the validation data plateaus or even decreases after the second epoch, yet its training accuracy keeps getting better, this indicates that the model is beginning to overfit.

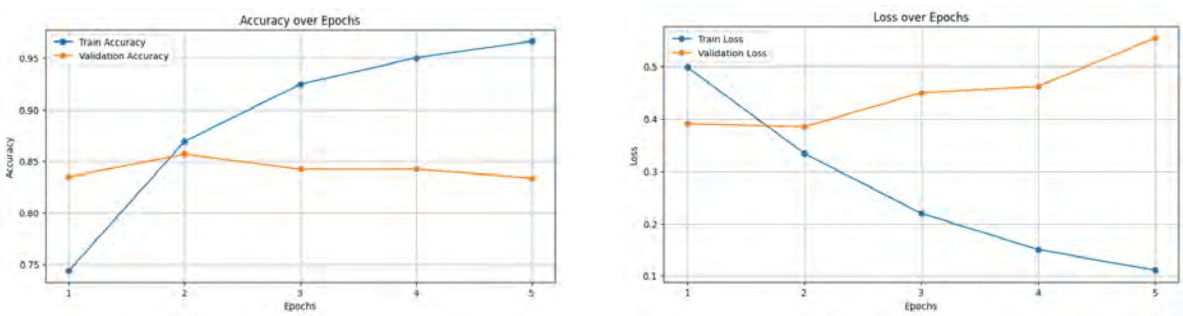


Figure 5.2.3.1: Loss and accuracy graph for BERT model

Significant points are shown by looking at the confusion matrix in Fig.[5.2.3.2].The confusion matrix provides a clear view of the model's classification performance. As shown by the top-left cell, where 466 occurrences are correctly identified, Class 1 is predicted with great accuracy. However, there are some misclassifications, with

66 instances of Class 1 being incorrectly labeled as Class 2. In the same way, in the second row, 348 Class 2 cases are accurately identified, but 70 Class 2 instances are incorrectly anticipated to be Class 1. These off-diagonal elements indicate areas where the model struggles, particularly in differentiating between Class 1 and Class 2, suggesting opportunities for improving the model's performance in these regions.

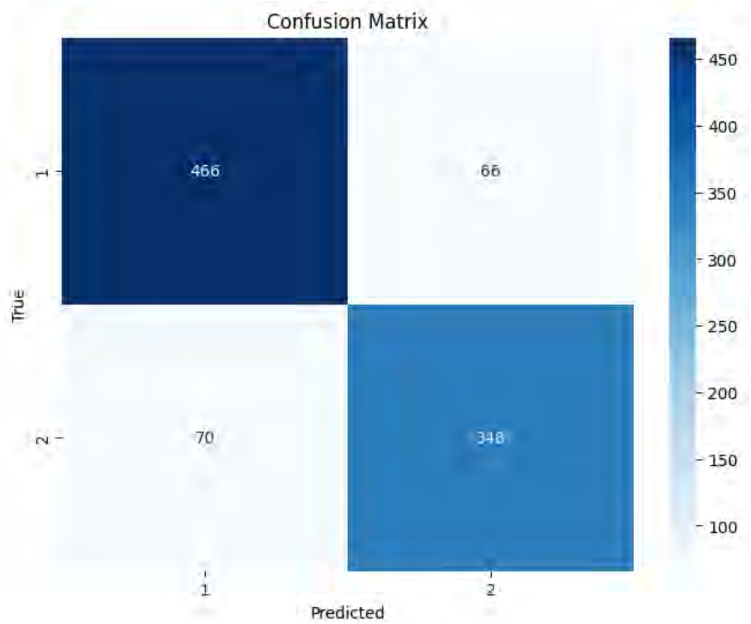


Figure 5.2.3.2: Confusion Matrix BERT model

5.2.4 RoBERTa

Architecture

We have employed the Robustly Optimized BERT Pretraining Approach (RoBERTa) model for classifying social media comments into Republican (1) and Democratic (2), ignoring Neutral (0) comments.Similar to BERT, RoBERTa identifies complex contextual relationship in text, with optimized training methods for better performance

The input text was tokenized using RoBERTa's pre-trained tokenizer, which produced tokens that were then transformed into dense embeddings. The multi-layered bidirectional transformer architecture of RoBERTa, capturing context from both directions within the sequence, thereafter receives these embeddings.Classification was done using the pooled output that linked to the [CLS] token, which represented the context of the entire sequence. After passing this output through a dropout layer to avoid overfitting, the logits for the Democratic and Republican labels were calculated using a fully connected layer. By minimizing the loss between predicted and true labels, and using dropout to improve generalization, RoBERTa effectively classifies political bias in social media comments.

Performance Evaluation

Where accuracy is 0.86631579 and f1 score is 0.85182709 with precision 0.85125257 and recall 0.85611757, the figure's loss function Fig.[5.2.4.1] the graph shows that as the number of epochs increases, the training loss gradually decreases, suggesting that the model is improving at reducing errors on the training set. Since the model is learning the training data effectively but not generalizing to new data, this discrepancy between training and validation loss indicates that the model is overfitting. Besides, the model appears to be learning the training data effectively, as evidenced by the accuracy graph, which shows a rapid increase in training accuracy from about 60% in the first epoch to 90% by the fifth. The model does well on training data, but it has difficulty maintaining its accuracy on validation data. This is demonstrated by the validation accuracy, which peaks early in the second epoch at about 86% and then begins to decline or plateau. This further emphasizes the overfitting issue in the model.

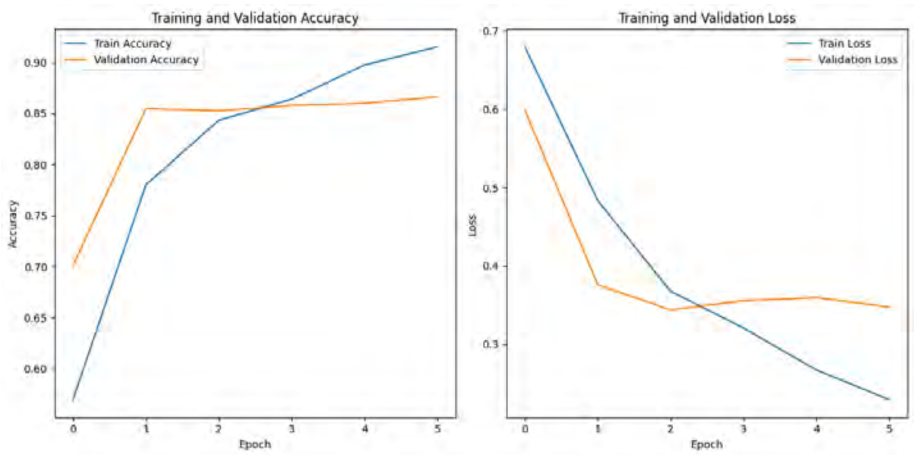


Figure 5.2.4.1: Loss and accuracy graph for RoBERTa model

Significant points are shown by looking at the confusion matrix in Fig.[5.2.4.2]. The confusion matrix provides a clear view of the model's classification performance. As shown by the top-left cell, where 440 occurrences are correctly identified, Class 1 is predicted with great accuracy. However, there are some misclassifications, with 92 instances of Class 1 being incorrectly labeled as Class 2. Similarly, in the second row, 370 Class 2 cases are accurately identified, but 48 Class 2 instances are incorrectly predicted to be Class 1. These off-diagonal elements indicate areas where the model struggles, particularly in differentiating between Class 1 and Class 2. This suggests there are opportunities for improving the model's performance in these regions, especially in reducing the misclassifications between the two classes.

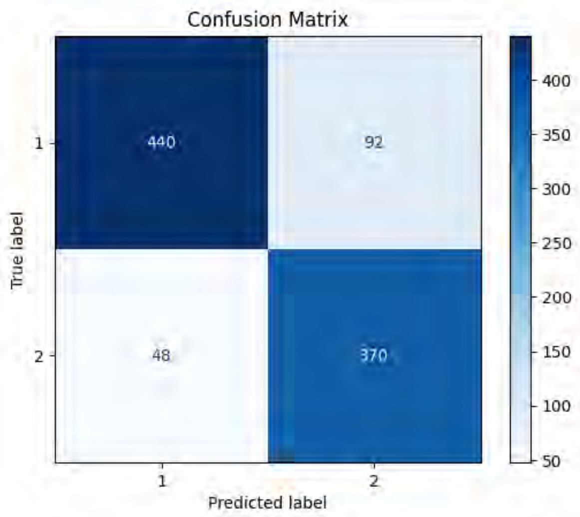


Figure 5.2.4.2: Confusion Matrix RoBERTa model

5.2.5 ALBERT

Architecture

We have classified social media comments into two categories using the A Lite BERT (ALBERT) model: Republican (1) and Democratic (2). Neutral (0) comments have been ignored. ALBERT is a simplified version of BERT that has the capacity to record complex contextual relationships while reducing the size of the model.
ALBERT's pre-trained tokenizer was used to first tokenize the input text, turning it into tokens that were then mapped to dense embeddings. The multi-layered bidirectional transformer architecture of ALBERT was applied to these embeddings in order to capture the sequence's forward and backward context. Classification was performed using the pooled output corresponding to the [CLS] token, which represents the context of the entire sequence. A dropout layer was subsequently applied to this [CLS] output in order to reduce overfitting. To calculate logits for the Republican and Democratic labels, the output was fed into a fully connected layer following dropout. In order for ALBERT to correctly identify political bias in social media comments, the model was trained by minimizing the loss between predicted logits and true labels. Dropout also helped to increase generalization.

Performance Evaluation

Where accuracy is 0.83789474 and f1 score is 0.81601171 with precision 0.81655809 and recall 0.81553315, the training and validation performance during five epochs is depicted by the loss function in Fig.[5.2.5.1]. The training loss gradually drops as the number of epochs rises, suggesting that the model is gradually reducing errors on the training set. However, the validation loss shows that although the model is doing a good job of fitting the training data, its capacity to generalize to unseen validation data is declining after an initial increase. Looking at the accuracy graph, the model shows a steep increase in training accuracy, from around 65% in the first epoch to over 95% by the fifth epoch. This suggests that the model is learning the training data effectively. However, the validation accuracy peaks early, reaching around 82% by the second epoch and then showing little to no improvement, even slightly declining. This further confirms that the model is overfitting, excelling on the training set but struggling to maintain its performance on the validation set.

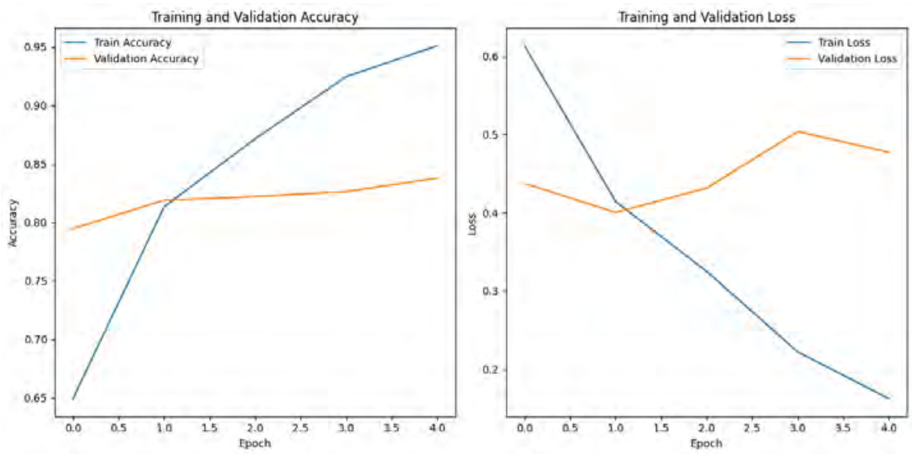


Figure 5.2.5.1: Loss and accuracy graph for ALBERT model

The confusion matrix in Fig.[5.2.5.2] provides a detailed analysis of the model's performance. The top-left cell shows that 449 instances of Class 1 are correctly predicted, indicating a strong ability to identify this class. However, 83 instances of Class 1 are incorrectly labeled as Class 2, representing some misclassification. Similarly, in the second row, 329 instances of Class 2 are correctly predicted, but 89 instances are incorrectly classified as Class 1. This indicates that the model struggles to distinguish between the two classes in certain cases. The off-diagonal elements highlight areas where the model is misclassifying, suggesting that the model could benefit from further refinement to improve its ability to differentiate between Class 1 and Class 2 more effectively.

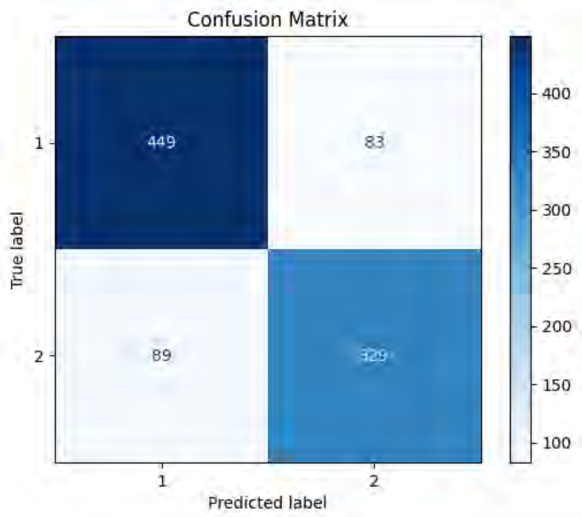


Figure 5.2.5.2: Confusion Matrix ALBERT model

5.2.6 Sentiment and Aggression Analysis

From the Fig. [5.2.6.1] we can see that the number of aggressive posts of the Republican Party is larger than that of the Democratic Party. We used the 10.9k data collected from Reddit and Twitter to show the comparison of aggressive posts between the Republican and Democratic parties. It demonstrates that the Republican party had a higher number of aggressive posts, with 447 posts compared to 426 for the Democratic party. Aggressiveness was defined by thresholds on toxicity-related features, such as threats, insults, and identity attacks.

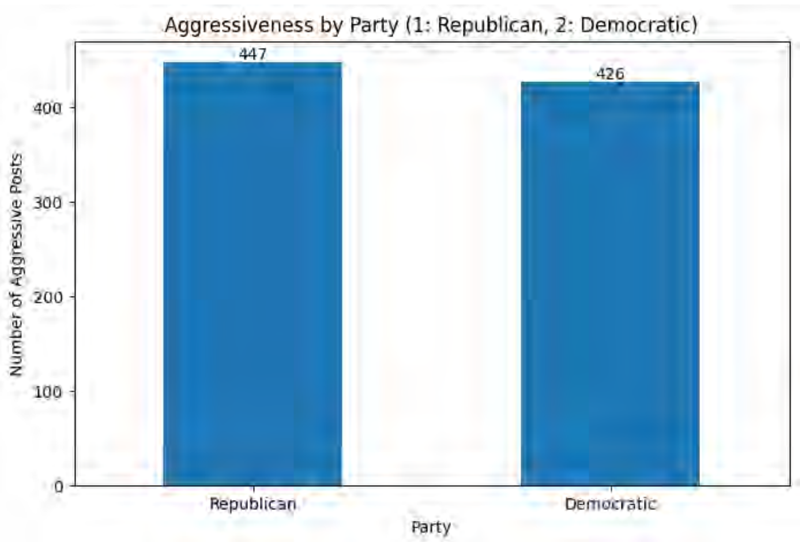


Figure 5.2.6.1: Comparison of aggressiveness of the Republic and the Democratic party

The Fig. [5.2.6.2] graph highlights the percentage of toxic posts by the Republic and the Democratic Party. Although both parties have similar toxicity percentages, Republicans displayed a little higher toxicity rate at 12.60%, compared to Democrats at 11.39%. This analysis reflects the prevalence of toxic content across political affiliations, based on the defined thresholds for toxicity.

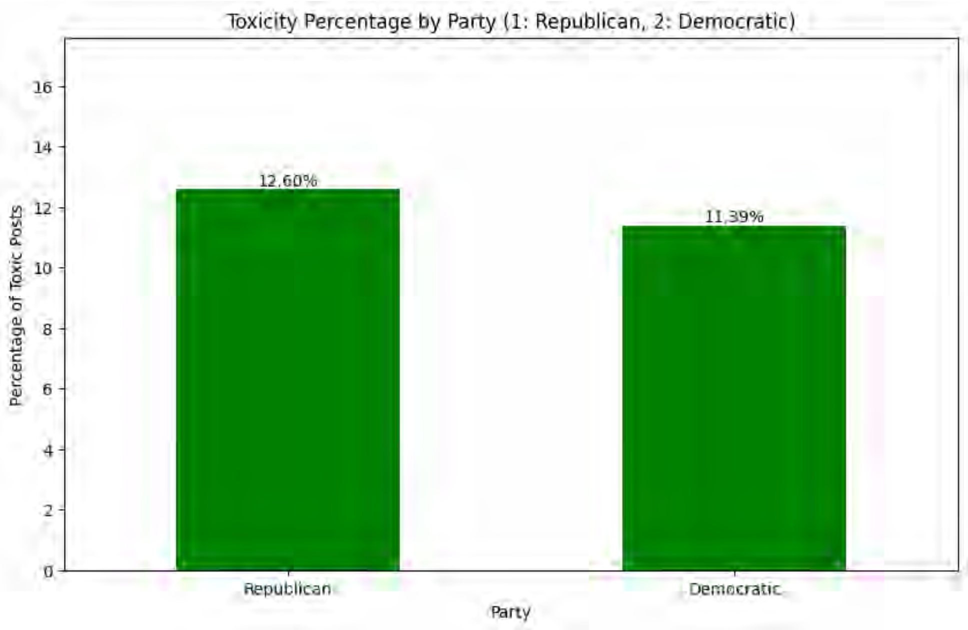


Figure 5.2.6.2: Comparison of toxicity of the Republic and the Democratic party

The third Fig. [5.2.6.3] illustrates the sentiment distribution by political party, comparing the Republican and Democratic parties. The posts are divided into three sentiment groups, which are negative, neutral, and positive, by the bar chart. With more than 3,500 posts, the Republican Party had more overall. The majority of them were negative, then the neutral, and the smallest were positive. On the other hand, the Democratic Party has fewer posts than the Republic Party . In the Democratic Party majority comments are negative, a significant number are neutral, and the smallest group is positive. As we can see, there are more posts from Republicans across all viewpoint groups, but there are also a lot of negative comments from both parties. This graphic provides light on the differences in sentiment among political parties.

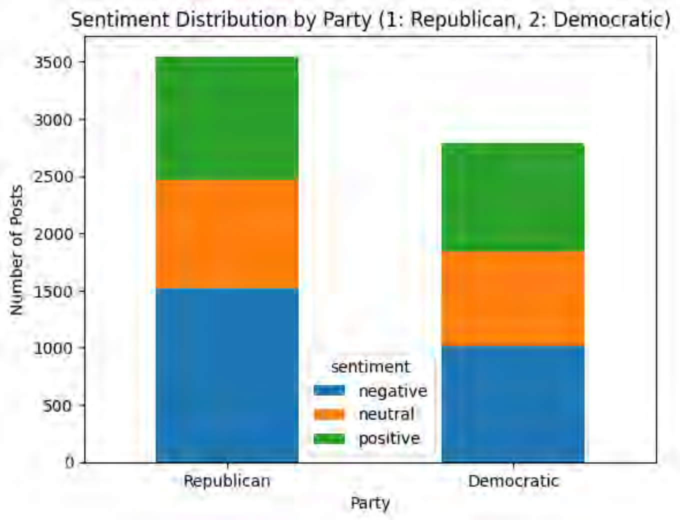


Figure 5.2.6.3: Sentiment Distribution of the Republic and the Democratic Party

5.2.7 15-day Data Analysis

The graph Fig. [5.2.7.1] illustrates the trend of aggressive posts related to Democrats on Twitter, collected over the period from September 20, 2024, to October 4, 2024. Initially, on September 20, the number of aggressive comments was high, but this number dropped drastically on September 21. Major political news on September 21, like Kamala Harris challenging Donald Trump to a second US presidential debate, may have contributed to the notable decrease in aggressive posts from Democratic Party supporters on September 21.
Then we can see there was a sharp increase in aggression on September 23, marking the highest point during this timeframe. This significant increase in aggressive comments occurs in connection with significant political occasions, like Donald Trump's rallies on September 23. After this peak, the number of aggressive posts fell significantly from September 25 to September 30. There might be fewer significant confrontational events during this time, and the political debate may have been less aggressive. With another peak on October 1 and a subsequent decrease in the final days of the period, the trend remained fluctuating. The news on 1st October: Planning for Emergencies A memo from the Council on Foreign Relations had issued a warning about the possibility of election-related violence, focusing on the nation's growing divisiveness, inflammatory rhetoric, and former President Trump's legal issues, which can be a reason for the increase of aggressive posts on 1st October, and it reached peak on 2nd October. Overall, the graph shows significant fluctuations in the frequency of aggressive posts during these 15 days.

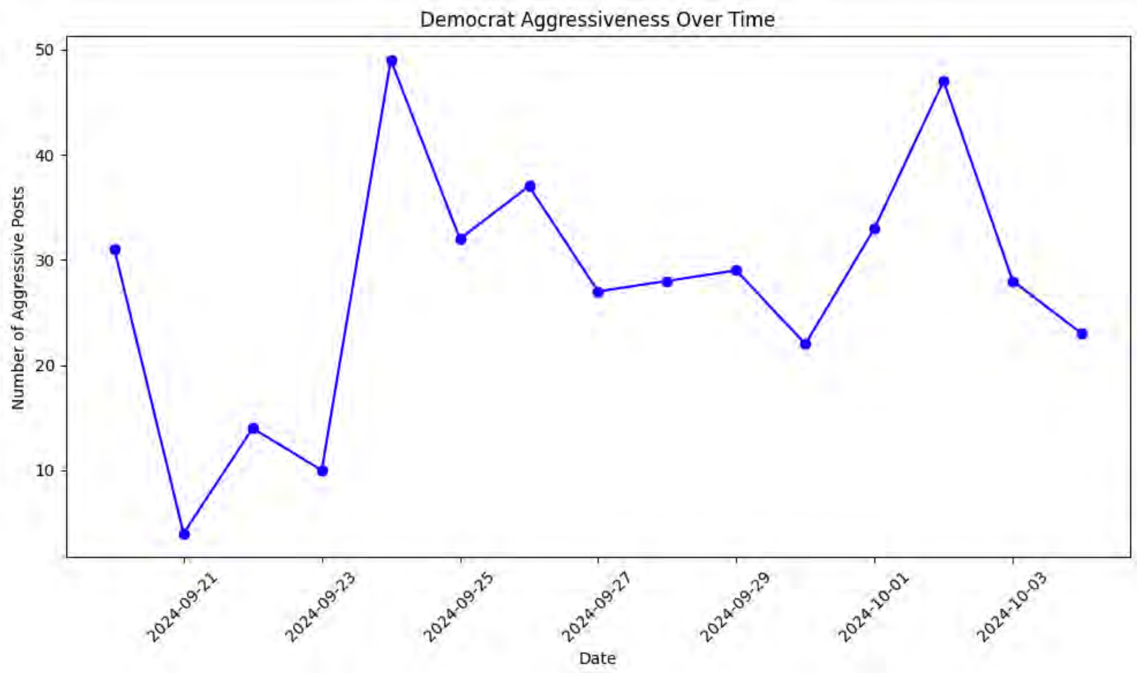


Figure 5.2.7.1: Aggressiveness of The Democratic Party over the 15 days

The initial part of the graph Fig. [5.2.7.2] shows a high level of aggression on September 21. Major political news on September 21, like Kamala Harris challenging Donald Trump to a second US presidential debate, may have contributed to the notable increase in aggressive posts from Republic Party supporters on September 21. As Kamala Harrris challenged the Republic Party President, the republican might have been triggered and showed high aggressiveness.
However, after this peak, there is a significant and steady decline in aggressive posts over the next few days. By September 25, the number of aggressive posts drops. This sharp decline in aggressive posts coincides with key political events, such as rallies held by Donald Trump on September 23 and September 25. Trump's rally on September 25 in Mint Hill, North Carolina, could have contributed to a reduction in aggressive posts, as supporters likely shifted their focus to the rally rather than engaging online. This is further supported by the decrease in posts following September 25.
For the rest of the period, from September 26 to October 1, the level of aggression remains relatively low and fluctuating. The cause of the aggression decasing can be that many politicians were talking on behalf of the Republic and supporting this party. For example. On September 27, the post where RFK Jr. tells Michigan supporters to vote for Donald Trump indicates support for the Republican Party. However, the quantity of aggressive posts significantly decreased at the end of the period, starting on October 1. On October 2, the aggressive post increased, and the cause can be that Kamala Harris said that Trump is unstable and unfit for the presidency, and the Republican supporters might have been triggered, so the number of aggressive comments of the Republican party increased.

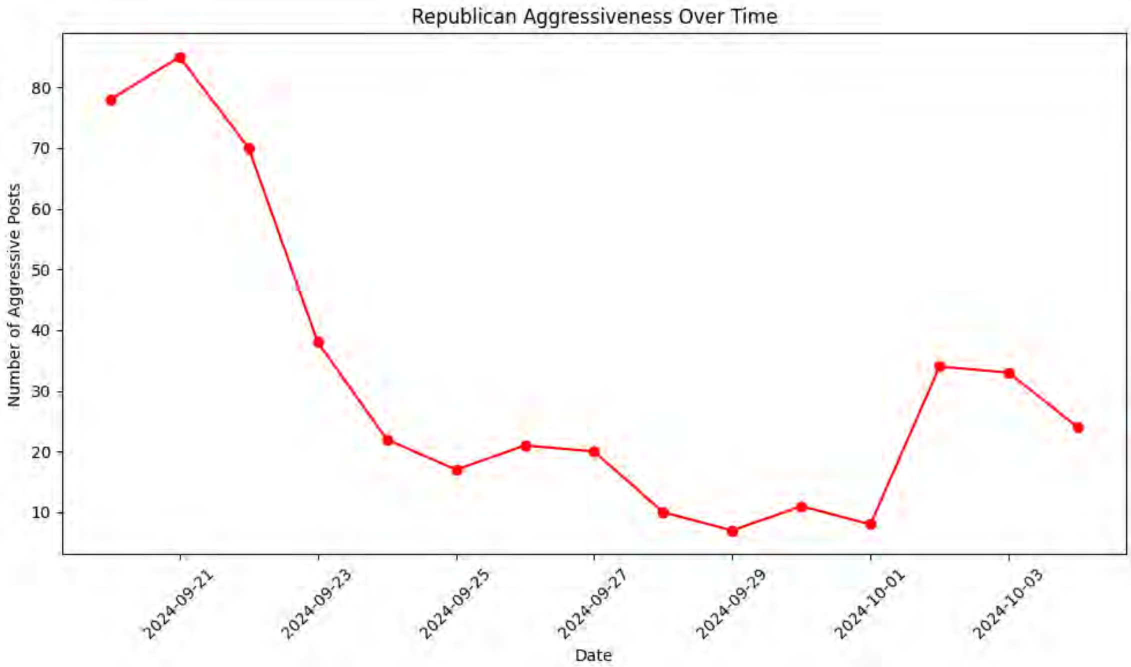


Figure 5.2.7.2: Aggressiveness of The Republic Party over the 15 days

The first graph, Fig. [5.2.7.3] shows the sentiment distribution of comments related to Democrats on Twitter from September 20 to October 4, 2024. On October 1, 2024, there is the most noticeable peak, with over 300 comments overall, mostly positive, then neutral, and a smaller percentage of negative comments. On September 30, 2024, there is another spike in comments, dominated by positive sentiments, though neutral and negative comments are also notable. September 24, 2024, sees a more evenly distributed increase in comment volume across all sentiment categories, although a smaller but still noteworthy increase.
Additionally, the lowest number of comments, with little activity across all sentiments, is seen on September 21, 2024. Similarly, there is a significant decline on October 4, 2024, with less than 150 comments. Though neutral and negative feelings are still there, positive comments tend to predominate throughout the course of the 15 days. The spikes in sentiment likely reflect increased engagement due to significant political events during this period.

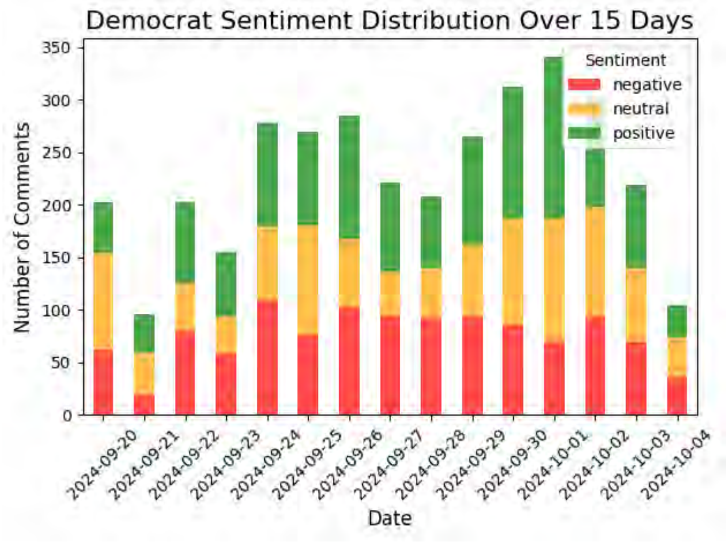


Figure 5.2.7.3: Sentiment Distribution of The Democratic Party over the 15 days

The second graph, Fig. [5.2.7.4] reflects the sentiment distribution for Republicans The most notable peak occurs on September 22, 2024, with a majority of them being positive, followed by neutral and negative sentiments. The number of comments sharply decreases after September 23 and is much lower between September 26 and October 1. On October 3, 2024, however, sentiment rises once more, with a rise in positive comments increasing the distribution. Although neutral and negative comments are constantly present, positive sentiment continues to be the most prevalent sentiment throughout the timeframe. These variations most likely represent reactions to significant political debates or occurrences during this period.

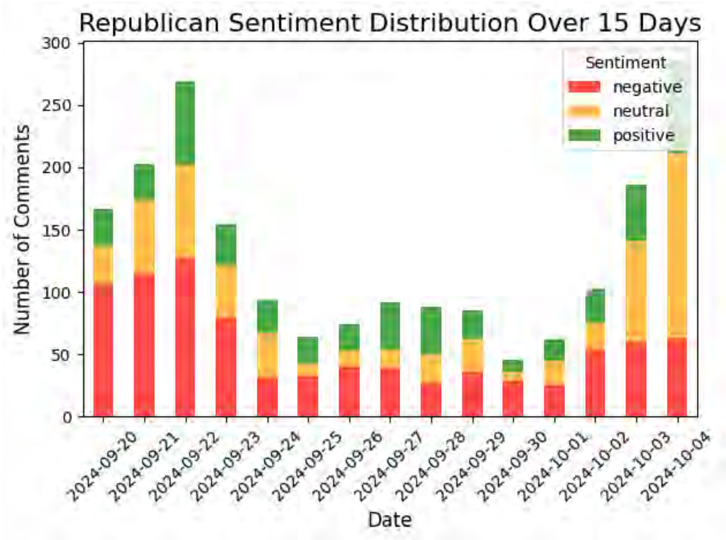


Figure 5.2.7.4: Sentiment Distribution of The Republic Party over the 15 days

In Fig. [5.2.7.5] from September 20, 2024, to October 4, 2024, the graph shows changes in the Democratic and Republican parties' sentiment scores. The sentiment of the Democratic Party stays mostly constant, around the neutral line (0). On

October 1, 2024, the sentiment score reaches roughly 0.2, marking the Democrats' highest point. Favorable political news or popular campaign strategies, like Vice President Kamala Harris challenging Trump in a second presidential debate or other significant occurrences this election season, may be responsible for this encouraging change.
In contrast, sentiment about the Republican Party generally moves more negatively, with notable declines. When the sentiment falls below -0.4 on September 30, 2024, it reaches its lowest point. Negative news or disputes involving Republican leaders or their policy stances may be to blame for this steep drop. For example, ongoing legal challenges faced by former President Donald Trump, combined with divisive rhetoric, may have contributed to increased negative sentiment. However, the Republican sentiment somewhat rose between September 28 and October 3, most likely as a result of favorable political developments or favorable media coverage.

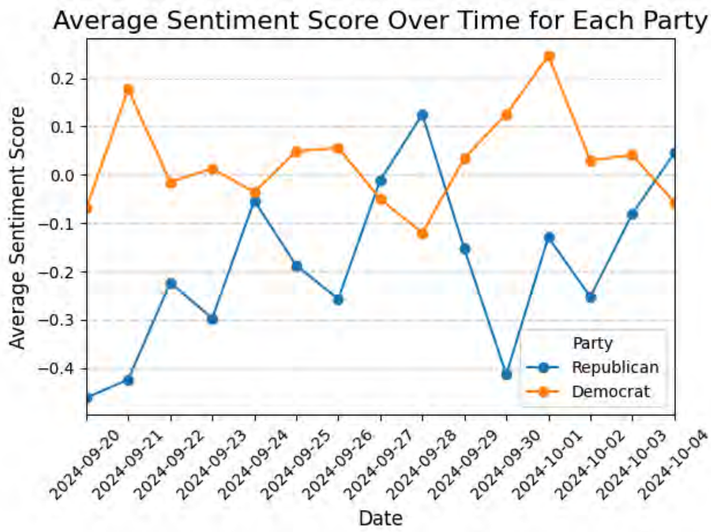


Figure 5.2.7.5: Average sentiment score of The Republic and The Democratic Party over the 15 days

In the graph Fig. [5.2.7.6] we can see that after September 21, the Democratic Party's positive sentiment (blue line) increases gradually until peaking on October 1 with about 140 positive tweets. Favorable political events like Kamala Harris challenging Donald Trump to a second presidential debate may have contributed to this rise in positive sentiment by boosting supporters and encouraging more positive discourse. The number of positive tweets sharply declines after this high.
On the other hand, the positive sentiment (red line) for the Republican Party stays comparatively low for the majority of the time, with a minor increase on October 3, perhaps as a result of rallies or favorable press coverage. The Republican Party's negative sentiment (red dashed line), on the other hand, is consistently present and peaked on September 21 and October 3, possibly linked to ongoing controversies or negative media attention, including the legal challenges faced by Donald Trump, which may have dampened the party's image. Because of the impact of significant political events on public opinion, the Democratic Party has maintained higher levels of positive sentiment overall over this time than the Republican Party.

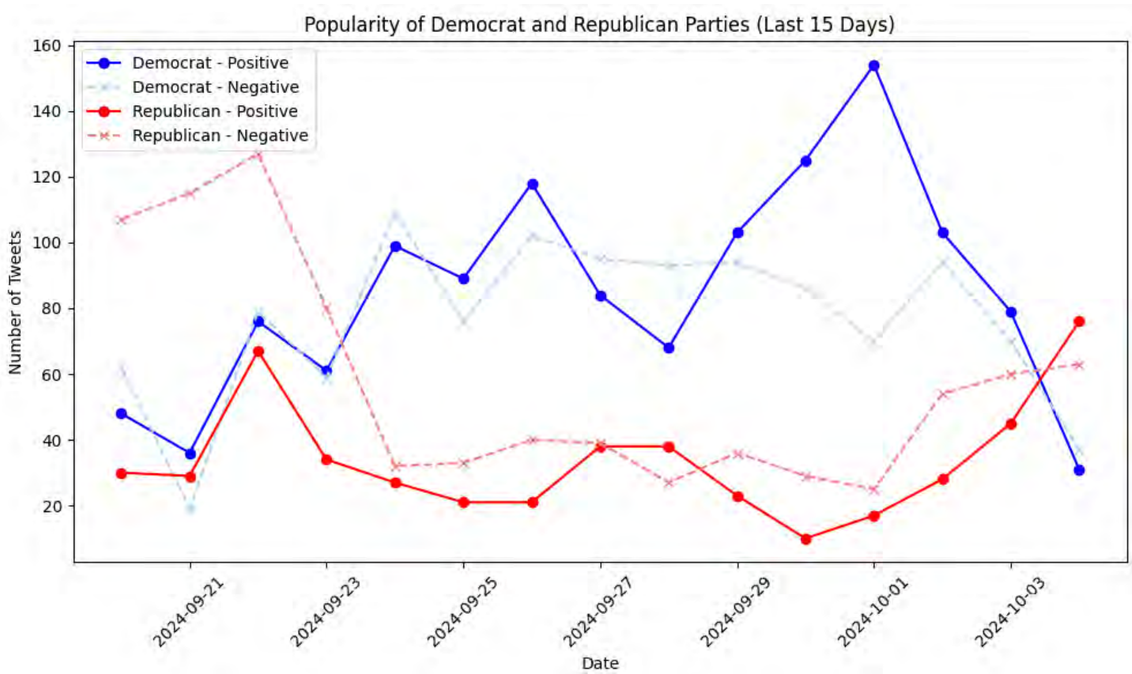


Figure 5.2.7.6: Change of popularity of The Republic and The Democratic Party over the 15 days

The graph Fig. [5.2.7.7] shows the number of tweets, both positive and negative that have been sent by each party during a period of 15 days. positive sentiment is continuously higher for the Democratic Party, particularly on September 26 and October 1, when there are noticeably more positive sentiment tweets than negative ones. Conversely, attitude toward the Republican Party varies significantly, with more negative tweets than positive ones on days like September 22 and October 3. Republican sentiment exhibits greater fluctuation between positive and negative comments, although Democratic opinion is often still more positive.

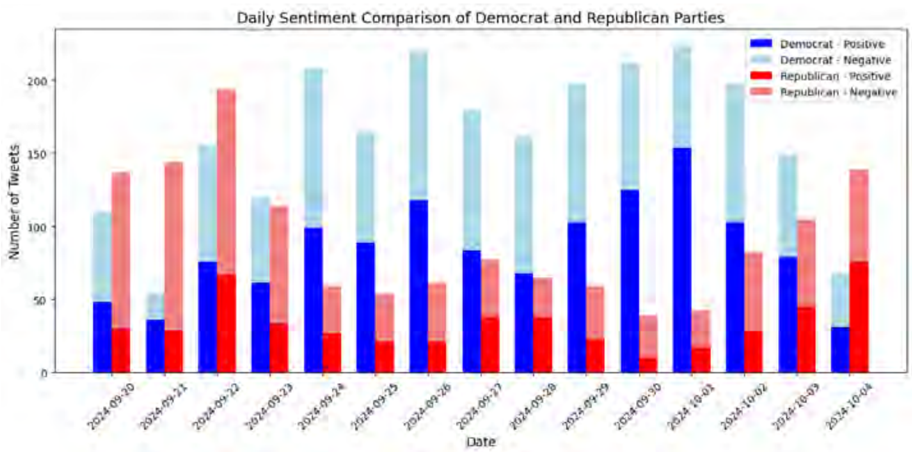


Figure 5.2.7.7: Change of sentiment of The Republic and The Democratic Party over the 15 days

5.2.8 Public Sentiment Comparison: Joe Biden vs. Kamala Harris

The graph Fig. [5.2.8.1] illustrates the difference in the level of aggression of comments made by Democratic supporters when Joe Biden was the party's nominee for the 2024 U.S. presidential election and after Kamala Harris replaced him on July 20, 2024.

About 250 aggressive comments were made by Democratic supporters when Joe Biden was the candidate, accounting for 14.9% of all comments. There could be a number of reasons for this increased aggression. For starters, Biden's campaign as a continuation of his presidency, probably came under criticism for the policies of his previous government, which included issues like foreign policy disputes and economic worries. In addition, Trump haters within the Democratic Party are partially responsible for the increase in aggressive posts from Democratic supporters. These people were fiercely anti-Donald Trump, anticipated his return, and had a vested interest in seeing him defeated.

On the other hand, the percentage of aggressive Democratic comments dropped significantly to 3.1% when Kamala Harris was elected. There are a number of reasons for this dramatic drop. First, by bringing a new face to the Democratic campaign, Harris' candidacy may have revitalized the support base and raised the tone of hopeful optimism. Her nomination might have been viewed as a fresh chance for reform, especially by supporters who were hoping for a candidate with a different leadership style or a more progressive agenda. Additionally, Harris' historic role as a woman of color on the ticket might have encouraged more positive discourse among Democratic supporters, as it symbolized progress in representation and diversity, leading to less negativity in the comments.

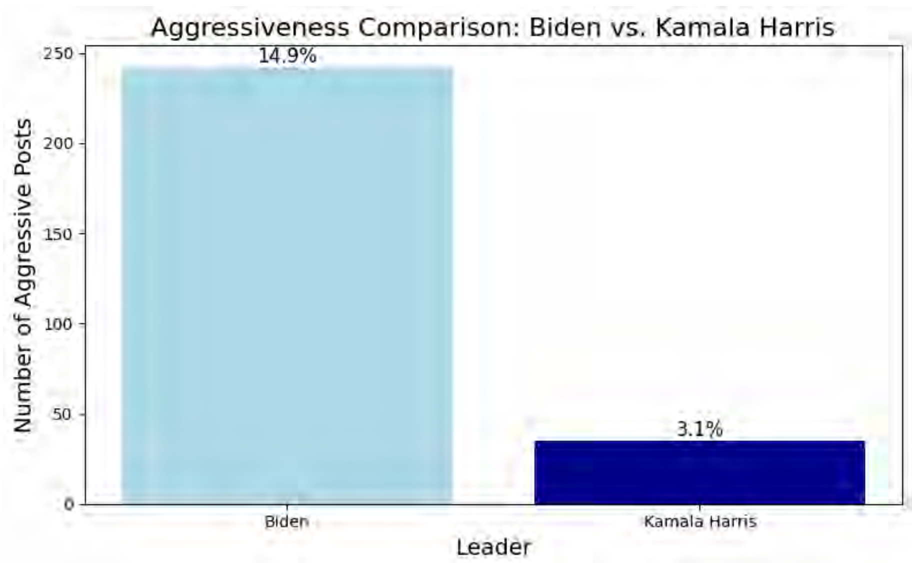


Figure 5.2.8.1: Aggressiveness Comparison of the democratic suppoterts when Biden was in Position and After Kamala Harris Came into Position

The graph Fig. [5.2.8.2] shows the percentage of aggressive posts from Republican and Democratic supporters before and after Kamala Harris became the Democratic presidential candidate in 2024, replacing Joe Biden.
Before Kamala Harris took over, 13.79% of Republican supporters' comments were aggressive. Harris's nomination caused this aggression to drop a little to 11.59%. The relatively small decline in aggressive posts indicates that the transition from Biden to Harris did not significantly change the overall sentiment of Republican supporters toward the Democratic nominee. This indicates that rather than focusing on the particular party leader, their opponents were more concerned with general Democratic principles. Harris's arrival may not have much changed their perception of the Democratic threat, resulting in a degree of aggressiveness that persisted but was somewhat lowered.
The shift in aggressiveness among Democratic supporters, however, was much more significant. 14.93% of their remarks throughout Biden's campaign were aggressive, indicating inner resentment that was probably fueled by worries about Biden's capacity to defeat Donald Trump, his leadership, or his policy choices. The progressive wing of the party, which frequently believed that Biden's moderate approach fell short on important topics like healthcare, economic inequality, and climate change, may have contributed to this increased aggression.
However, the percentage of aggressive comments made by Democratic supporters dropped dramatically to 3.06% once Kamala Harris was elected as the candidate. This points to a notable change in tone, perhaps brought on by a number of circumstances. It's possible that Harris's candidacy, as a newcomer and historic leader, renewed the party's passion and optimism, reducing internal friction and rallying supporters around her campaign.The graph demonstrates a significant decrease in Democratic aggression after Harris's nomination, while Republican aggression decreased barely. although Harris's arrival might have had a minor impact on Republican discourse,it significantly reduced internal tension and aggressive behavior among Democratic supporters, likely reflecting their increased unity and optimism under her leadership.

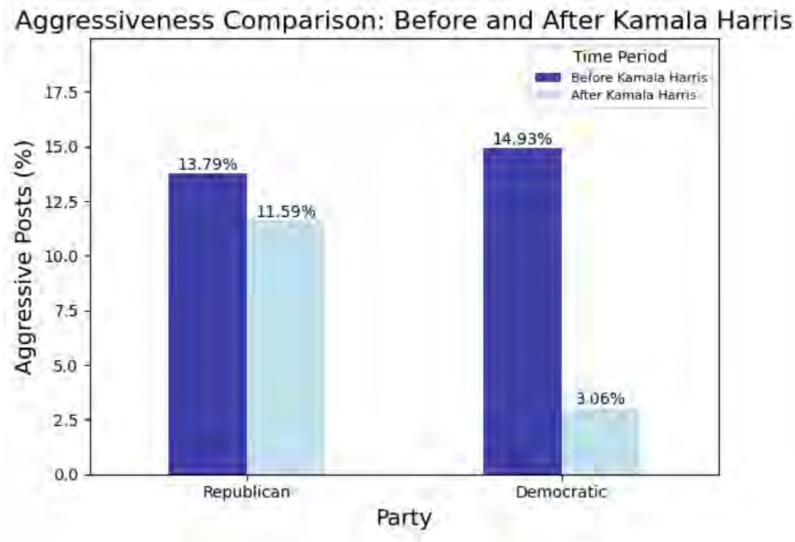


Figure 5.2.8.2: Aggressiveness Comparison of both of the political parties when Joe Biden was in Position and After Kamala Harris Came into Position

The graph Fig. [5.2.8.3] presents a breakdown of positive, neutral, and negative sentiment toward the Democratic and Republican parties during Joe Biden's campaign and following Kamala Harris's handover as the Democratic nominee. Under Biden's time, 37.14% of Democratic comments were positive, compared to 32.23% under Harris'. Positive sentiment among Republican supporters also increased throughout Biden's campaign (33.76%), but it dropped to 26.45% once Harris was elected. Following Harris' nomination, Democratic Party neutral sentiment increased sharply, from 21.41% under Biden to 53.45% under Harris, suggesting a shift in Democratic supporters' responses toward more moderate ones. The neutral attitude among Republicans was comparatively stable, standing at 23.78% under Biden and 30.73% following Harris's takeover.
After Harris became the candidate, Democrats' negative opinion fell precipitously, from 41.46% to 14.32%. Republicans' general negative opinion of the opposition party, on the other hand, remained relatively constant at about 42 before and after Harris' nomination. This implies that Kamala Harris's nomination led to a more neutral and positive tone among Democratic supporters, while Republicans remained consistently negative in their sentiment towards the Democratic leadership.

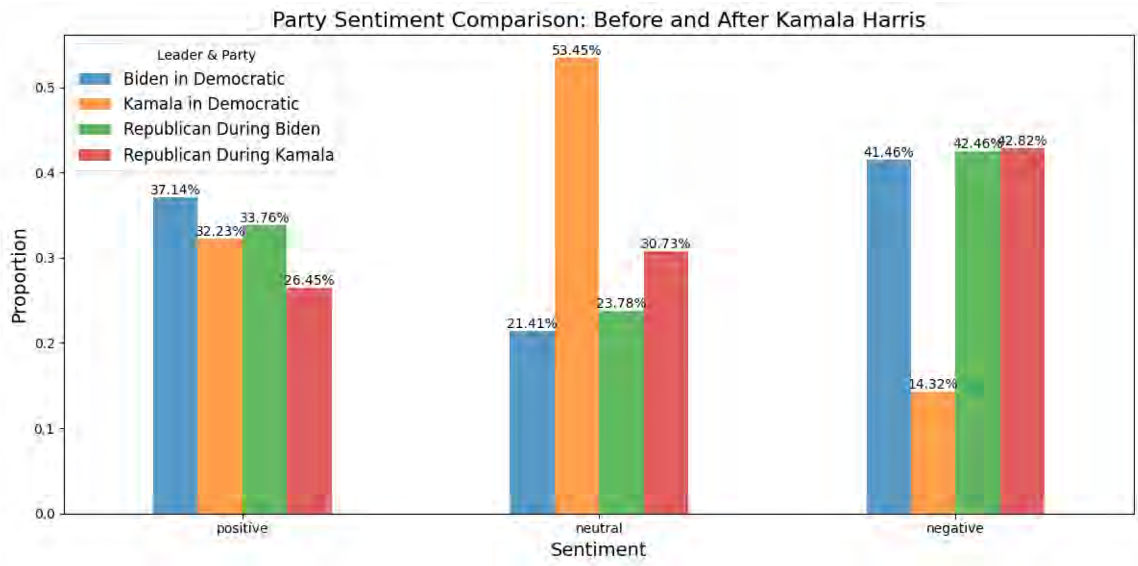


Figure 5.2.8.3: Sentiment Comparison of both of the political parties when Joe Biden was in Position and After Kamala Harris Came into Position

5.2.9 2024 Twitter Data vs Reddit Data

The graph Fig. [5.2.9.1] shows the distribution of negative, neutral, and positive sentiment among Reddit comments classified as Republican and Democratic is depicted in the bar chart. The data indicates that more comments linked with Republicans are viewed negatively, as evidenced by the somewhat larger number of negative sentiments in Republican-labeled comments than in Democratic ones. Both groups display comparable percentages for neutral sentiment, with Republican comments slightly outnumbering Democratic ones. Compared to their Republican sentiment, the Democratic-labeled comments exhibit a somewhat higher volume of positive emotion, suggesting that more comments attributed to Democrats are perceived positively. Overall, the graph shows that the two groups' sentiment distributions are balanced, with just little variation in the percentages of negative, neutral, and positive sentiment.

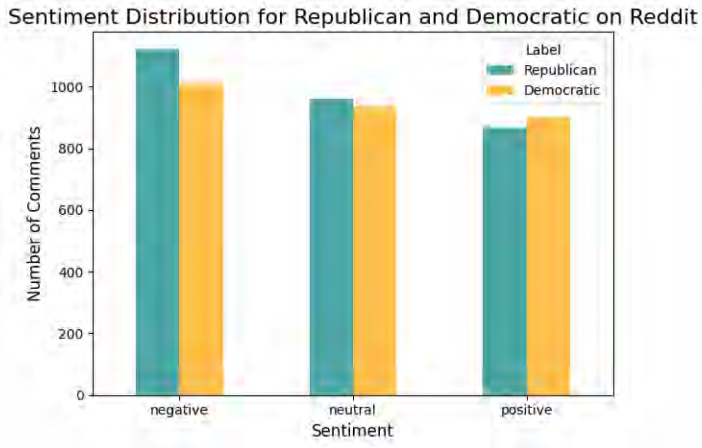


Figure 5.2.9.1: 2024 Sentiment Distribution for Republican and Democratic on Reddit

The graph Fig. [5.2.9.2] shows the sentiment distribution of Twitter comments classified as Republican and Democratic is depicted in the bar chart. Republican-labeled comments have a larger count than Democratic-labeled comments for negative sentiment, suggesting that more Republican comments are viewed negatively. Republicans have a slightly higher number than Democrats in the neutral sentiment survey, which is close between the two groups. Democratic-labeled comments outnumber Republican-labeled comments in terms of positive sentiment, indicating that Democratic views are more often viewed positively. Overall, the graph indicates that people are more negative about Republicans and more favorable about Democrats, while the two groups' levels of neutrality are comparable.

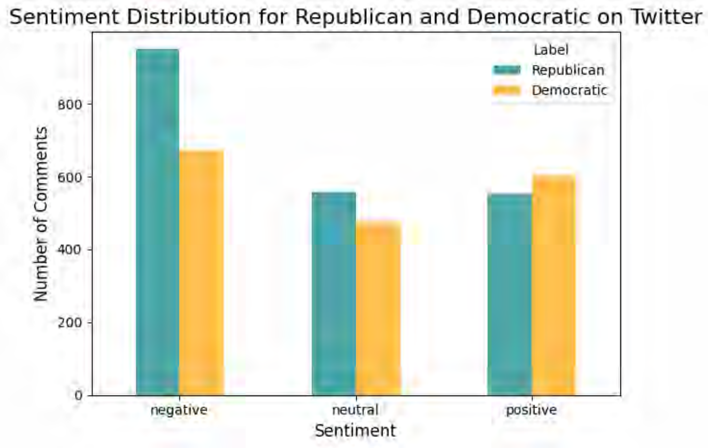


Figure 5.2.9.2: 2024 Sentiment Distribution for Republican and Democratic on Twitter

The graph Fig. [5.2.9.3] shows Republicans' and Democrats' monthly aggression on Reddit in 2024, as depicted in the bar chart. Aggression between the two parties increased in July after Kamala Harris was announced as the Democratic nominee and Joe Biden resigned on July 20. This change in leadership probably fueled heated political debates. Republican hostility peaked in September, which also happened to be the month of Donald Trump's legal issues and the September 23 protests in his favor. Posts that were aggressive peaked on September 25 and then started to decline. In October, as the election neared, Democrats showed more aggression, likely driven by final campaign efforts, while Republicans' aggression remained high due to ongoing controversies.

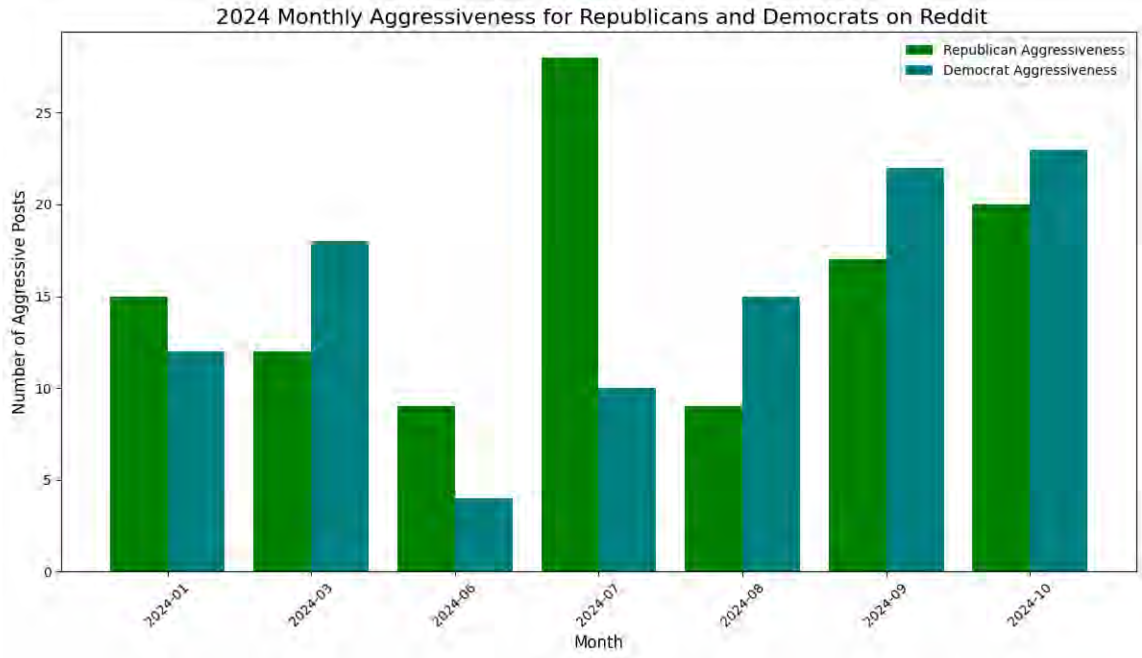


Figure 5.2.9.3: 2024 Monthly Aggressiveness for both Political parties on Reddit

The graph in Fig. [5.2.9.4] shows how aggressive Republicans and Democrats were on Twitter each month in 2024. Republicans are more aggressive in January, most likely as a result of early election talks. Democrats' aggressiveness increases significantly by March, maybe as a result of the primary season and political arguments. While Democrats continue to be less aggressive, Republican aggressiveness significantly increases in July, most likely as a result of the Republican National Convention. Both parties reach their most aggressive peak in September, which also happens to be the month of Donald Trump's scandals and vigorous campaigning. Democratic aggression marginally outpaces Republican aggression in October, probably as a result of Kamala Harris's claim that Trump is unqualified to be president, which sparked more heated online debates.

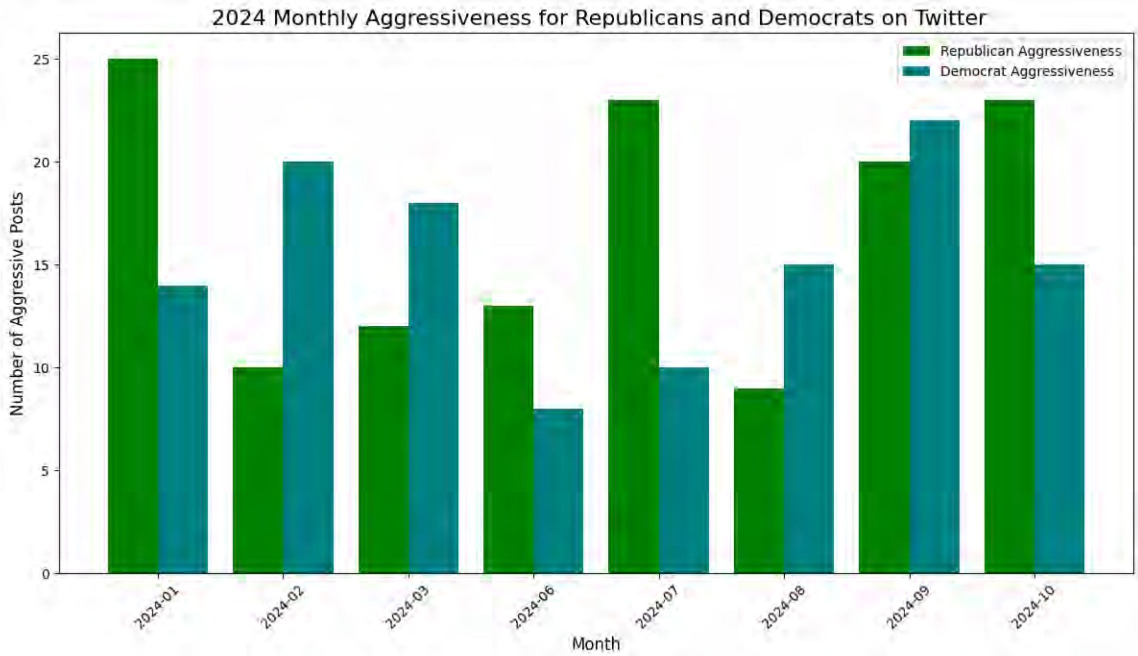


Figure 5.2.9.4: 2024 Monthly Aggressiveness for both Political parties on Twitter

Chapter 6

Result Comparative Study

6.1 Multi-Class classification

Model names	**Epoch**	**Accuracy (%)**	**F1 Score (%)**	**Precision (%)**	**Recall (%)**
BiLSTM model	30	61.066336	59.840948	61.127695	59.794871
Multilayer BiLSTM model	20	63.48419	61.14469	62.216638	60.726357
BERT	15	85.684211	85.459882	85.499135	85.423787
RoBERTa	15	86.631579	85.182709	85.125257	85.611757
ALBERT	15	83.789474	81.601171	81.655809	81.553315
BERT - Short text	20	83.333333	82.656168	84.68599	82.160774
BERT - Long text	20	84.552845	83.444211	85.402619	82.758414
RoBERTa - Short text	20	83.676975	83.171084	83.603801	82.945533
RoBERTa - Long text	20	84.281842	84.023198	84.666147	84.139147
ALBERT - Short text	20	87.285223	86.659125	86.576125	86.78452
ALBERT - Long text	20	85.365853	81.873078	82.194936	81.656068

Table 6.1: Evaluation Metrics for Multi-label Classifications of different models

From the above table 6.1, we can see that the RoBERTa model achieves the highest accuracy, making it the most overall correct model in prediction across all instances by providing a notable accuracy of 86.63157%. And a loss score of 0.3472 shows that the model is performing efficiently, minimizing errors. Furthermore, the F1 score of 85.18% shows good balance of the model with precision and recall scores of 85.13% and 85.61%.

With the highest F1 score (85.46%), which achieves a balance between precision and recall, BERT is clearly maintaining a strong balance between the two. With an F1 score of 85.18%, RoBERTa also shows a very strong performance with an F1 score of 85.18%, while ALBERT performs slightly lower with 81.60%.

Precision, which measures how many of the predicted positive cases were actually correct, is also led by BERT (85.49%), followed closely by RoBERTa (85.12%). These models' high precision scores demonstrate their exceptional dependability in producing accurate predictions.

With a recall of 85.61%, RoBERTa leads in detecting true positives, demonstrating the model's high capacity to identify true cases, and BERT comes in second with 85.42%.

BERT's accuracy for short text is 83.33%, greater than its accuracy for long text (84.55%). This suggests that BERT performs slightly better with short texts than with lengthy texts.

Similarly, RoBERTa performs 83.67% for short text and 84.28% for long text. The strongest performer for short text classification is ALBERT, which achieves an impressive 87.28% accuracy on short texts compared to 85.36% on long texts.

Seeing the overall result, RoBERTa and BERT consistently surpass the other models in the majority of metrics, especially when it comes to efficiently handling both short and long texts. For this task, transformer-based models such as BERT and RoBERTa are far more effective than the BiLSTM and Multilayer BiLSTM models, which lag behind significantly.

Chapter 7

Limitations

Limitations in Data Diversity

We may have overlooked some political data even though we gathered our dataset from social media sites like Reddit and Twitter, which gave it some degree of robustness. Our findings may have been affected since our dataset may not have included all possible social media comment variations. Social media networks like Facebook, Instagram, and YouTube have different user profiles and interaction patterns, which may show different aspects of the public opinion. Therefore, it is evident that more improvement is required.

The Importance of Regular Data Updates

Like all social media platforms, languages are always changing. Unfortunately, our dataset is not dynamic. Our dataset may become invalid and misleading in the near future if new data becomes available. Comments on politics are changing daily. There may be new abbreviations added. These variables will not be present in our dataset, and it may become outdated in the future.

Challenges in Capturing Different Languages

We have worked with English language comments, so our model does not understand political comments in any other languages, though those comments are correlated with our research findings.

Limitation in Multi-Political Party Analysis

Third-party candidates such as the Green and Libertarian parties were not included in our analysis, which was limited to the Democratic and Republican parties. Even though they are rarely successful, third-party candidates have the power to affect election results by dividing votes or detracting from front-runners. For instance, George W. Bush's victory against Al Gore in 2000 is frequently attributed to Ralph Nader's Green Party campaign. Third-party candidates may be included in future studies to gain a deeper understanding of their influence on election outcomes and voter behavior.

Limitation in Data Collection Period

This study's data period, which only spans January through October 2024, limits its applicability by leaving out important information from prior election cycles. It is more difficult to identify long-term patterns or comprehend how voter sentiment changes over time in the absence of historical data. Furthermore, it is more difficult to assess how incumbents, voter turnout, and campaign tactics will affect future voting trends when historical election outcomes are ignored. Such information would provide a more thorough understanding of electoral behavior.

Absence of Previous Election Data

Political sentiment is frequently shaped by voter histories, long-standing party affiliations, and reactions to significant events like political scandals or economic crises. Information from the U.S. presidential elections in 2016 and 2020 might have shed light on how voter attitudes and behavior have changed over time. Historical patterns that could have improved the understanding of electoral dynamics are ignored in this study, such as shifting voter loyalties and the long-term impacts of political polarization. These elements would have provided a more thorough picture of the results of subsequent elections.

Contextual and Geographical Limitations

Our study only looked at forecasting the US presidential election in 2024, but this limited emphasis has drawbacks. Election trends vary greatly between nations, particularly in parliamentary democracies or multi-party systems. Cultural and organizational factors influence social media engagement, public discourse, and political dynamics. As a result, our results might not be generalizable to other countries. Future research could broaden the focus by examining elections in nations with various political structures, exposing distinctive patterns, and providing more comprehensive understandings of electoral behavior worldwide.

Chapter 8

Future Work

There is a lot of room to improve the ability to identify political comments in the domains of deep learning and natural language processing. So we aim to build up on our existing work in the future. Such areas are:
Dataset Diversity: For further studies, broadening the scope of information sources is essential. Although we used data from Twitter and Reddit, a more complete picture of public opinion might be obtained by including traditional media as well as websites like Facebook, Instagram, and YouTube. A larger range of political debate, including those from older or less technologically proficient individuals, would be captured by this more extensive data collection.
Expanding Sentiment Analysis to Multi-Party Systems: Our sentiment analysis methods could be used in future studies to observe public sentiment toward various parties and investigate the ways in which smaller parties affect election outcomes. A deeper understanding of election tactics and voter behavior in multi-party settings may be gained by examining coalition-building in systems with proportional or mixed-member voting.
Incorporating Historical and Regional Data for Enhanced Accuracy: By accounting for changes in views on politics, voting trends, and tactics, historical election data could enhance prediction models and produce more accurate projections. Because U.S. states have different political environments, future research may also benefit from examining discussion at the regional level. For instance, Texas and Florida are typically conservative, while California and New York lean Democratic. Swing states like Pennsylvania and Michigan sometimes determine presidential outcomes.
Global Analysis of Political Support for U.S. Parties: In our future work, We plan on expanding the scope of our study in the future by collecting global comments to identify the countries that support particular U.S. political parties. . Additionally, we aim to adapt our methodology to analyze political discourse across various nations, making our approach applicable globally. This will not only provide insights into international perspectives on U.S. elections but also enable the study of political trends and voter sentiment in diverse political systems.

Finally, to improve performance, further testing and fine-tuning will be done using different deep learning models. In addition to identifying overt instances of political remarks, we also want to increase the model's capacity to recognize political remarks and more accurately assess political sentiment and popularity.

Chapter 9

Conclusion

In conclusion, political discourse analysis plays an important part in our social lives in the present day. Our research has explored the challenges associated with evaluating and understanding the political comments on English-language textual content on social media, allowing us to identify the comments supporting each political party. Our team members carefully classified a large dataset of political speeches into specific categories. To accurately categorize and analyze the subject categories of the texts, we used advanced language models, especially Multilayer BiLSTM and BiLSTM, BERT, RoBERTa, and ALBERT models. These models improved our ability to understand and analyze political discourse effectively. In our thesis, the sentiment, aggression, and popularity of each political party's supporters throughout a certain time period—including shifts in sentiment after a change in political candidates have all been graphically depicted. Additionally, we compared the differences in political sentiment among party supporters on two major social media platforms, Twitter and Reddit. So, monitoring supporter sentiment and tracking the popularity of political parties are important, as these factors significantly influence election campaigns. We believe that our unique dataset will help future researchers in this interesting field of research. We explored at several kinds of models and provided the findings from the top-performing models. There is always room for improvement, though. Our near future goal is to enhance the models' performance. We had a number of limitations. We aim to overcome these restrictions in the future. We hope to continue researching and improving our findings. In the future, we could be able to attain outcomes in the 90th percentile. We will continue to examine existing literature and research in this field to uncover any subtleties or insights we may have overlooked. Overall, we believe our study will offer valuable insight on the many strategies and techniques used for political discourse analysis efficiently for future researchers.

Bibliography

[1] M. D. Conover, B. Goncalves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of twitter users," in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 2011, pp. 192–199. DOI: 10.1109/PASSAT/SocialCom.2011.34.

[2] P. L. Dunmire, "Political discourse analysis: Exploring the language of politics and the politics of language," *Language and Linguistics Compass*, vol. 6, no. 11, pp. 735–751, 2012. DOI: https://doi.org/10.1002/lnc3.365. eprint: https://compass.onlinelibrary.wiley.com/doi/pdf/10.1002/lnc3.365. [Online]. Available: https://compass.onlinelibrary.wiley.com/doi/abs/10.1002/lnc3.365.

[3] K. Johnson and D. Goldwasser, "Identifying stance by analyzing political discourse on Twitter," in *Proceedings of the First Workshop on NLP and Computational Social Science*, D. Bamman, A. S. Doğruöz, J. Eisenstein, *et al.*, Eds., Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 66–75. DOI: 10.18653/v1/W16-5609. [Online]. Available: https://aclanthology.org/W16-5609.

[4] D. Beleveslis, C. Tjortjis, D. Psaradelis, and D. Nikoglou, "A hybrid method for sentiment analysis of election related tweets," in *2019 4th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, 2019, pp. 1–6. DOI: 10.1109/SEEDA-CECNSM.2019.8908289.

[5] A. N. V. Hutabarat, "Political discourse analysis in translation of speech texts on "one belt one road (obor)"," *Proceedings of the Fifth Prasasti International Seminar on Linguistics (PRASASTI 2019)*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:203078532.

[6] P. D. Katre*, *Nlp based text analytics and visualization of political speeches*, Sep. 2019. DOI: 10.35940/ijrte.c6503.098319. [Online]. Available: http://dx.doi.org/10.35940/ijrte.C6503.098319.

[7] J. Stegmeier, W. J. Schünemann, M. Müller, M. Becker, S. Steiger, and S. Stier, "Multi-method discourse analysis of twitter communication: A comparison of two global political issues," in *Quantifying Approaches to Discourse for Social Scientists*, R. Scholz, Ed. Cham: Springer International Publishing, 2019, pp. 285–314, ISBN: 978-3-319-97370-8. DOI: 10.1007/978-3-319-97370-8_10. [Online]. Available: https://doi.org/10.1007/978-3-319-97370-8_10.

[8] P.-L. Huguet Cabot, V. Dankers, D. Abadi, A. Fischer, and E. Shutova, "The Pragmatics behind Politics: Modelling Metaphor, Framing and Emotion in Political Discourse," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 4479–4488. DOI: 10.18653/v1/2020.findings-emnlp.402. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.402.

[9] B. Sharber, "Analyzing political polarization in news media with natural language processing," Ph.D. dissertation, University Honors College Middle Tennessee State University, 2020, pp. 24–28.

[10] A. Bilbao-Jayo and A. Almeida, "Improving political discourse analysis on twitter with context analysis," *IEEE Access*, vol. 9, pp. 104 846–104 863, 2021. DOI: 10.1109/ACCESS.2021.3099093.

[11] E. Kubin and C. von Sikorski, "The role of (social) media in political polarization: A systematic review," *Annals of the International Communication Association*, vol. 45, no. 3, pp. 188–206, 2021. DOI: 10.1080/23808985.2021.1976070. eprint: https://doi.org/10.1080/23808985.2021.1976070. [Online]. Available: https://doi.org/10.1080/23808985.2021.1976070.

[12] S. Mishra, S. Choubey, A. Choubey, N. Yogeesh, J. Durga Prasad Rao, and P. William, "Data extraction approach using natural language processing for sentiment analysis," in *2022 International Conference on Automation, Computing and Renewable Systems (ICACRS)*, 2022, pp. 970–972. DOI: 10.1109/ICACRS55517.2022.10029216.

[13] P. N. Valerevna and K. F. Rakhmatovna, "The concept of "political discourse"," *European Scholar Journal*, vol. 3, no. 3, pp. 88–91, 2022.

[14] 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. 235–256, 2023. DOI: 10.1017/pan.2022.10.

[15] R. Németh, "A scoping review on the use of natural language processing in research on political polarization: Trends and research prospects," *Journal of Computational Social Science*, vol. 6, no. 1, pp. 289–313, Apr. 2023, Received: April 1, 2023, ISSN: 2432-2725. DOI: 10.1007/s42001-022-00196-2. [Online]. Available: https://doi.org/10.1007/s42001-022-00196-2.

[16] B. Strawser, "A discourse analysis of donald trump's and the american alt-right's conception of truth and its implications for public discourse," Jun. 2023, Available at SSRN: https://ssrn.com/abstract=4473952 or http://dx.doi.org/10.2139/ssrn.4473952.

[17] J. Torregrosa, S. D'Antonio-Maceiras, and e. a. Villar-Rodríguez Gregorio, "A mixed approach for aggressive political discourse analysis on twitter," *Cognitive Computation*, vol. 15, pp. 440–465, Mar. 2023. DOI: 10.1007/s12559-022-10048-w. [Online]. Available: https://doi.org/10.1007/s12559-022-10048-w.

[18] M. Mohd, S. Javeed, Nowsheena, M. A. Wani, and H. A. Khanday, "Sentiment analysis using lexico-semantic features," *Journal of Information Science*, vol. 0, no. 0, p. 01 655 515 221 124 016, 0. DOI: 10.1177/01655515221124016. eprint: https://doi.org/10.1177/01655515221124016. [Online]. Available: https://doi.org/10.1177/01655515221124016.