Social Media Technologies 2023
Research Focus & Paper Summaries

Patrick Balent
Clemens Hofmann
Rohit Kaushik
Marcel Lohfeyer
Paula Nauta
Saeed Saadati Pour
Monika Raffalt
Tobias Stöckl
Florian Werkl
Thomas Zenkl
Nina R. Zettl

December 4, 2023

Keywords— data-mining, social media, sentiment analysis, political, reddit, twitter

1 Research Focus

In our research topic, we data-mine two social media platforms for political posts and perform sentiment analysis towards two political figures. Our main motivation for this task is to assess modern methods for sentiment analysis and identify political bias comparing both platforms. We chose to compare Twitter and Reddit regarding sentiments towards Joe Biden and Donald Trump.

Datasets:

1. Reddit: Using the Reddit API, we scrape all comments on posts from selected subreddits during the time of the 2020 US election (approx. jan. 2020 - april 2021).
2. Twitter: We use a Twitter Dataset, grouped into Donald Trump and Joe Biden hashtags, also during the time of the 2020 US election.

We aim to collect our Reddit data by selecting certain[4] subreddits containing known political discussions and scraping all the comments from them. We will then filter these comments based on certain keywords like "Trump" or "Biden" and perform sentiment analysis over these topics. Identifying political comments well is key towards having clean data. More complex approaches could use classifier networks[4] or machine learning, but we are not sure how well those will perform.

Research Question: How did the sentiment of Joe Biden and Donald J. Trump shift during and after the 2020 US election period?

Additionally, we also want to prove or disprove the following Hypotheses:

- Is Donald Trump more favored on Twitter?
- Is Joe Biden more favored on Reddit?
- Is Donald Trump more featured on Twitter?
- Is Joe Biden more featured on Reddit?
- Was sentiment towards Donald Trump negatively impacted on both platforms after the Jan. 6^{th}, capitol raid.

Possible problems: Natural language is inherently ambiguous. Informal posts, often contain much less data. That is why, sentiment analysis might fail in many places. We will probably not be able to detect sarcasm in comments.

2 Paper Reviews

In this section, we analyze our research specific papers.

2.1 Analyzing the Traits and Anomalies of Political Discussions on Reddit [1]

With data scraped from two popular news-subreddits and the corresponding linked news sites, the authors of this paper systematically analyze the actions of users in online discussions. They focus specifically on the sentiments of the users and want to gain insights into different archetypes of discussions to evaluate and find expected and abnormal behavior.

They categorize discussions into harmonious and controversial discussions: ones where users agree with each other and others where users strongly disagree. In an attempt to further understand what types of posts steer discussions and controversies, the authors evaluate posts in three dimenstions. First are actions, which are all interactions of users like posting text messages and voting (up/downvoting) on other users comments or posts. The other two are users sentiments relative to preceeding posts and the root post, and the third is the variation of topics.

Based on this model, they attmpted to verify or falsify 10 hypothesis related to discussion paths, discrepancy, X-posts (controversial discussions), disruptions and topic similarity. Most of their hypothesis manifested as true, on both datasets. Only less than half were false or inconclusive. All hypotheses were tested statistically against their pattern-based model of the discussion archetypes. To conclude, they found that their model of archetypes enabled to connect important elements and give insights into the relationship between the three analyzed dimensions.

2.2 The Impact of Features Extraction on the Sentiment Analysis[2]

In this paper, the authors analyze a the "SS-tweet" dataset, a sentiment strenght twitter dataset that was annotated manually. They propose a pipeline to analyze this data with pre-processing, feature extraction and classification algorithms. Lastly, they compare the performance between different variations of techniques used on metrics like precision, recall, accuracy and F-score.

Features are extracted either with either TF-IDF (frequency-inverse document frequency) or N-Grams in form of vectors, specifically 2-Grams in this paper.

Then, the extracted features are processed with 6 different classification algorithms such as KNN, SVM, Random Forest and Logistic Regession. They found that using word-level TF-IDF performed around 3-4 % better than 2-Grams, and that Logistic Regression was the best technique for both TF-IDF and N-Grams.

2.3 A review on sentiment analysis and emotion detection from text [3]

This paper aims to give a general overview of sentiment analysis and emotion detection from informal text on social media platforms. It differentiates between sentiment analysis (analysing if a piece of text is either neutral, positive or negative) and emotion detection (identifying human emotions in text like fear or happiness). The main challenge about Natural Language Processing (NLP) techniques is that language is inherently ambiguous. A working NLP system transforms this unstructured data into meaningful insights.
In their main evaluation, they categorise different approaches in the following categories:
Lexicon Based Approaches This method maintains a word dictionary in which each positive and negative word is assigned a sentiment value. It is further categorised into dictionary-based and corpus-based methods. Both methods perform well regarding both emotional and sentiment analysis. However, the dictionary-based approach is more straightforward to apply and easier to generalize.
Machine Learning Approaches They categorise their approaches into traditional Machine Learning and Deep Learning approaches. Traditional Machine Learning approaches (Support Vector machines, Random Forests) are not further evaluated, they have, however, improved recently. Deep Learning approaches like Convolutional Neural Networks (CNNs) and Long short-term memory (LSTM) tend to perform depending on the pre-processing and size of the data-set. They tend to handle large data-sets very well and are also commonly used for the task. Nonetheless, in some cases, Machine Learning models fail to extract some implicit features or aspects of the text.
They evaluated the models using a Confusion Matrix against their ground-truth data. Based on these values, researchers evaluated their model with metrics like accuracy, precision, and recall, F1 score, etc..

2.4 A Large-Scale Text and Network Resource of Online Political Discourse [4]

In this paper, Hofmann et al. present a large-scale data-set of political discourse covering more than 600 political discussion groups over a period of 12 years (from January 2008 to December 2019). They used the publically available Pushshift Reddit Dataset (PRD) and performed machine learning to identify subreddits with actual political discussion. A classifier identified political comments from the set. They train separate classifiers for each year, since the discussion tends to shift over the years. As positive examples, they take for each year all comments from r/Anarchism, r/Anarcho Capitalism, r/Conservative, r/Libertarian, r/Republican, r/democrats, r/progressive, and r/socialism. These subreddits were chosen since they represent different points on the ideological spectrum and thus do not bias the classifiers towards certain political ideologies. Furthermore, they classified different subreddits to their respective political ideology.

Finally, they clustered their data and extracted network data using graph theory. They found out, that Democrats and Republicans are further apart than expected. They form two connected, but very distinct clusters.

2.5 Sentiment Analysis between VADER and EDA for the US Presidential Election 2020 on Twitter Datasets.[5]

Using the respective data set provided by Kaggle, the author explores tweets about the 2020 US election and investigates changes in sentiments that were expressed towards the two candidates. Scraped by using the identifiers "#JoeBiden" and "#DonaldTrump", the data set contains 1.7 Million tweets from between October 15th and November 8th 2020, the final weeks before and the days after the election. Using VADER (Valence Aware Dictionary for Sentiment Reasoning), a widely applied dictionary-model for sentiment analysis that maps words to corresponding intensities of emotions, the author attempts to show how the proportions of positive, neutral, and negative tweets change during the period under study. Therefore, distinguished by the two candidates, all tweets that were made within one day were analysed with VADER and the proportions of positive, neutral, and negative sentiments were being estimated. Using linear regression, these data points were used to visualize trends in their sentiments, concluding that first of all most tweets expressed neutral sentiments and that tweets with negative sentiments towards both candidates proportionally lost weight, while such with neutral or positive sentiments gained importance during the run up of the election.

This paper is one of the worst I have ever read for several reasons, including the missing precision of the scientific expression, an erratic and incomprehensible argumentation, the inclusion of useless graphics (word cloud?), not to mmention spelling and grammar. Whether the sole author (who consistently refers to themselve as "we") has been taken in by a "predatory journal" or it is a case of irresponsible peer review, such an article should not be published.

2.6 Analyzing voter behavior on social media during the 2020 US presidential election campaign.[6]

By scraping over 20 million tweets using various identifiers related to the US election in 2020, the authors present the results of several analyses they conducted on their dataset, including sentiment analysis, topic modelling, and network analysis. The sentiment analysis (exact method unspecified but referred to as hybrid approach combining lexicon-based and machine learning-based methods) reveals that the majority of tweets were neutral, with a smaller proportion being positive or negative. The topic modelling analysis identifies key themes in the tweets, such as the coronavirus pandemic, racial justice, and the economy. The network analysis examines the interactions between users on Twitter and identifies clusters of users with similar interests. Demonstrating the opportunities of data analysis techniques and the role social media plays in shaping public opinions, it would have been interesting to get some more detailed insight into this (apparently quiet big) endeavour that seeks to provide a thorough summary over how contemporary computational methods can enhance political and media research.

2.7 Sentiment Analysis on Twitter Data Sample [7]

The paper Sentiment Analysis on Twitter Data from January 2015 on pages 178 - 183 in the International Journal of Innovative Research in Advanced Engineering discusses "the existing analysis of twitter dataset with data mining approach such as use of Sentiment analysis algorithm using machine learning algorithms" (Sahayak, Shete Pathan, 2015, p. 178). The area of application of such a sentiment analysis can be important in various areas, such as feedback on products or the sentiment on US Presidents Biden and Trump, relevant to this research project. An approach is treated that classifies the sentiments behind tweets from Twitter as positive negative or neutral, by the using of three models: unigram model, tree kernel model and feature based model. Two resources are used: 1) hand annotated dictionary for emoticons, 2) acronym dictionary gathered from the internet. There will be considered features (emoticons, neutralization, negation handling and capitalization). They use different machine learning classifiers (Naive Bayes, Maximum Entropy (MaxEnt), and Support Vector Machines (SVM)) and feature extractors (Unigrams, bigrams, unigrams and bigrams, and unigrams with part of speech tags). The Naive Bayes classifier assumes that the impact of a variable's value on a given class is independent of the values of other variables and is therefore class independent. The Maximum Entropy (MaxEnt model is Feature based. In this model, independence is not assumed. Support Vector Machines are a theoretically motivated algorithm. Support vector machines are supervised learning models with associated learning algorithms. Three models: the Unigram Model, tree kernel Model, and feature based model—are created in Weka using these machine learning algorithms. For feature extraction, these models will be employed. According to the survey, social media-related characteristics can be utilized to forecast sentiment on Twitter.

2.8 Sentiment Analysis on Tweets in the 2020 US Presidential Election [8]

The Paper Sentiment Analysis on Tweets in the 2020 US Presidential Election in the Journal of High School Science from Chandak A examined the relationships between Twitter sentiment trends and election outcomes by using three various sentiment analysis models: Bag-of-Words model, trained on sentence embedding, and two Out-of-box classifiers to predict the sentiment of election-related tweets. The Valence Aware Dictionary and Sentiment Reasoner library's (out-of-the-box classifiers) offered the best level of accuracy. The same machine learning algorithms described in the paper Sentiment Analysis on Twitter Data from January 2015 by Varsha Sahayak, Vijaya Shete Apashabi Pathan were used in this study. The output of the sentiment analysis models was a collection of tweets that were arranged by the date they were posted and the political party they belonged to and categorized as positive, negative, or neutral. Only English-language tweets were used in the study and a total of around 541000 Biden-related tweets with the hashtags JoeBiden and Biden and 694000 Trump-

related tweets with the hashtags DonaldTrump and Trump were evaluated. The results based on the sentiment classification models in total: Biden tweets: about 28% negative, 29% neutral, and 43% positive; Trump tweets: about 38% negative, 33% neutral, and 29% positive. The sentiments of the tweets were also looked at over time, and the perception of the candidates changed depending on whether they won or lost the election. Trump received more negative than positive tweets before the election and lost them. Hence, the assumption that such an analysis can be used to predict the outcome of an election

2.9 Challenges of Sentiment Analysis for Dynamic Events[9]

The paper from 2017 highlights the challenges and difficulties of building a robust sentiment analysis platform to capture sentiments for predicting election results, focusing on the 2016 US presidential election. The authors propose to use a model for each individual candidate and an algorithm that classifies user's political leaning into five groups: far left-leaning, left-leaning, far right-leaning, right-leaning, and independent users.

The researchers hypothesis is the tendency of users to follow others with similar political orientations. They collected a set of Twitter users with known political orientations such as senators and then calculated the probability with a ratio and threshold towards a leaning.

The process faces several content-related and interpretation-related challenges, such as dealing with hashtags, external links, sarcasm, sentiment vs. emotion analysis, and vote vs. engagement counting. Other highlighted factors that could be a threat to the performance outcome of an sentiment analysis are user location information and the presence of manipulating social bots which create challenges regarding trust.

2.10 RAFFMAN: Measuring and Analyzing Sentiment in Online Political Forum Discussions with an Application to the Trump Impeachment[10]

The paper presents RAFFMAN (Real-time Affect Fingerprints For Measuring Narrative), a systematic approach aimed at quantifying changes in forum user sentiment towards specific topics in response to real-world events. This approach is useful for understanding how sentiment evolves over time in online discussions.

RAFFMAN consists of three phases: (a) filtering and identifying related posts, (b) detecting changes in engagement using time series, and (c) conducting sentiment analysis. To accomplish this, the authors first used a keyword-based approach to identify and gather posts relevant to the topic of interest. Next, they detected changes in user engagement by analyzing the volume of relevant posts over time. The researchers identified spikes in engagement corresponding to real-world events, which allowed them to study the impact of these events on user sentiment, providing insights into how public opinion shifts during significant occurrences. Lastly they used BERT, a state-of-the-art transfer learning model for natural language processing to classify posts into positive, neutral, and negative sentiment categories, achieving high classification accuracy.

The dataset used in this study comprises 32 million posts gathered from the two discussion forums, Reddit and 4chan, with focus politically-oriented subforums over six months from September 2019 to February 2020 during significant U.S. political events, depicting a case study of the Trump impeachment.

The outcomes demonstrate that RAFFMAN achieves a classification accuracy of 81.1% when focusing on posts with less than 23 words and up to 74%

accuracy with all posts. The results show the potential of capturing user affection and tracking its sentiment change in online discussions and in the future the authors plan to make it open-source.

2.11 Election 2020: the first public Twitter dataset on the 2020 US Presidential election. [11]

The researchers in this study are addressing the importance of understanding online political discourse for ensuring free and fair elections in a democracy. As Twitter has historically been a platform used by politicians to reach their base, and other online social platforms are used by the population in order to voice their opinions and engage in conversation surrounding the elections it has caused also another phenomenon, that is that social media has become an environment where misinformation and disinformation can flourish and spread. They point out that limited access to social media data often makes it difficult to study and understand online political discourse. To support researchers overcome this barriers they release a massive-scale dataset related to the US Presidential elections 2020 that was being collected for over one year starting at May 2019. This period covers the events described above and more. For their data collection they use the Twitter's streaming API through the Tweepy library and follow specific mentions and accounts related to candidates who were running to be nominated as their party's nominee for president of the United States, in addition to a manually-compiled, general election-related list of keywords and hashtags. Besides finding out some limitations that should be considered in future research like the skewness of results due to the language being limited to English users, as well as twitter not being the only platform that is used in order to reach out political followers through campaigns, the researchers aim is that by providing this dataset will help empower the Computational Social Science research community and support further study relevant scientific and social issues related to politics, such as misinformation, information manipulation, conspiracies, and the distortion of online political discourse. The dataset is available for public use on Github (see in the quotemarks).

2.12 Sentiment, we-talk and engagement on social media: insights from Twitter data mining on the US presidential elections 2020[12]

The researchers in this study wanted to understand what types of social media messages during a political event, specifically the 2020 United States presidential election, lead to more engagement from the public. They used the dual process theory to test how both affective cues (such as emotional valence and intensity) and cognitive cues (such as insight and causation) contribute to engagement. They collected a dataset of over three million tweets and assessed the affective and cognitive cues through sentiment analysis. They found that both affective and cognitive cues were important in engaging audiences, with negativity bias observed in the overall sample. However, emotionally charged content produced higher engagement in the subsample of tweets from famous users. The authors also found that collective self-representation ("we-talk") was consistently associated with more likes, comments, and retweets. The study sheds light on

the effectiveness of both affective and cognitive cues on information appeal and dissemination on Twitter during a political event, and the role of the tweet's author in moderating these relationships.

2.13 Political Discourse on Social Media:Echo Chambers, Gatekeepers and the Price of Bipartisanship[13]

Experts are increasingly concerned that echo chambers disrupt the discourse essential to democracy. The paper examines the existence and impact of so-called echo chambers within a political discourse on social media. An echo chamber occurs when social media users are consistently exposed to content that reflects their own political beliefs and values. This can create a cycle where they only see and share content that reinforces their existing views. In contrast to prior efforts, the Method takes into account all shared and created content instead of exclusively concentrating on particular forms of interactions and data that are mutually agreed upon for content and network.
By using a metric that identifies if a tweet is liberal or conservative, we can determine its polarity. This also allows us to categorize users into three groups: Partisan users who mainly post content with one-sided leaning, Bi-Partisan users who post content with both leanings and Gatekeeper users who consume content from both sides but mainly post one-sided content. The existence of echo chambers in discussions with political content can be proven through the bimodal distribution of the posted tweets of a user and the tweets received on their feed from users they follow polarities and the associated correlation. The analysis also revealed that users who share a tweet that does not align with the associated page's opinion might face the consequences known as the "price of bipartisanship". This can include criticism from the page itself, which may discourage users from taking positions in discussions that require compromise.

2.14 A Method for Predicting the Winner of the USA Presidential Elections using Data extracted from Twitter [14]

This paper aims to predict the presidential election winner in 2016 in 3 of the three major swing states. The data to make this possible was collected using the Twitter Search API, which enables the targeted collection of JSON data. In this case, the API queries have to filter US citizens eligible to vote within the respective states and who do not belong to the group of non-voters. In order to find the most relevant tweets, queries were built on relevant hashtags to find campaign-related tweets.
A Naïve Bayes classifier was used to classify the tweets as positive, neutral, or negative using sentiment analysis. This was implemented with the help of the Python library text blob, which allows the creation of user-defined classifiers. The tweets were also analyzed using "subjectivity analysis", which is also implemented in Textblob. The classification of the sentiment analysis could be further improved by the subjectivity score. A data set by N. Sanders was used to train the classifier, which is also based on already classified tweets.
The election winner in all three states was accurately predicted. This reveals that while direct voting cannot be determined through Twitter data, it is pos-

sible to observe how people talk about the presidential candidates. Drawing important conclusions about upcoming events in a society can be facilitated with the help of sentiment analysis.

2.15 Automated Pipeline for Sentiment Analysis of Political Tweets [15]

This study focuses on the topic of the 2020 presidential election. More specifically, they devote their methodology to a sentiment analysis. In order to carry out this analysis successfully, a few conditions had to be met in advance. Firstly, it was agreed to use the social media platform Twitter, the data for the analysis was exclusively obtained in English and carried out one month before the elections, as it was assumed that this is when the users' need to communicate is at its highest. Secondly, 23 countries were considered and 100,000 tweets were generated first. Finally, some classifications were made and data processing was carried out, so that 27,000 tweets could be analysed. Furthermore, these tweets were also classified as "Pro Trump", "Pro Biden", or neutral and an error label was also set up. In the results section, two visualisations were carried out, firstly a word cloud, from which one can see how the content of the tweets can be interpreted. It could be shown that Biden and Trump were mentioned equally often, with the only difference being that the mentions are to be interpreted pro Biden, as Trump is accompanied by derogatory remarks in the tweets. A visualisation through a geographical landscape was also carried out to get an overview of political sentiment by country. Using the geographical representation, it was shown that out of 23 countries, Hong Kong, Ukraine, Saudi Arabia, Nigeria, and Japan - are pro-Trump.

2.16 Biden vs Trump: Modelling US general elections using BERT language model [16]

This study is an analysis of the election forecasts for the 2020 presidential election in America. It also underpins the difficulty, relevance and chances of sentiment analysis as a reliable methodology for political forecasting in the future. It also mentions that the performance of that analysis is sought through the use of Twitter data. The challenge is to classify this content correctly, as there are also different forms of expression in terms of tone and different cultures. In the course of the study, over 1.2 million tweets were analysed from October 2020 to November 2020, exactly in the time frame of the first political debate around the final results. The BERT model was used for the data analysis of the individual sentiments. It was also shown that most tweets were from the US, Europe or India. In addition, some data was collected in order to be able to make attributions to the users. For this purpose, the date of joining twitter, tweet ID, retweet count and user follower count were taken into account. It was shown that there was a positive tendency regarding to Biden. Furthermore, the explorative data analysis showed that a few people on Twitter have a great influence on the rest. Furthermore, it felt that people who shared their geographic location were not actively involved in the election. In addition, the election was during the Covid19 pandemic, which made it more difficult to make a clear prediction as everyone's life situation deteriorated. Ultimately, the sentiment analysis showed that Biden have a higher chances of winning the election.

2.17 A large-scale sentiment analysis of tweets pertaining to the 2020 US presidential election[17]

In order to gauge popular opinion of the candidates, this study examined 7.6 million tweets that mentioned the 2020 US Presidential Elections between October 31 and November 9, 2020. The favorability of each presidential contender and how it evolved as the election-related events played out were studied by the authors using sentiment analysis. The emotion held for each candidate could be seen across different groups of users and tweets thanks to an innovative method used to identify deleted or suspended tweets and user accounts. According to the study, tweets that were removed before Election Day were more supportive of Donald Trump than those that were deleted afterward of Joe Biden. Additionally, it was discovered that older Twitter accounts posted more supportive tweets about Joe Biden. The study emphasises how crucial it is to analyse sentiment on all posts—even those that are no longer accessible—to ascertain the actual feelings people had at the time of an occurrence.

2.18 Sentiment Analysis of before and after Elections: Twitter Data of U.S. Election 2020[18]

The data used for this article's sentiment analysis was gathered before, during, and after the 2020 US presidential election on Twitter. To extract features and ascertain if there was positive or negative sentiment towards the candidates, the authors employed the TF-IDF and Naive Bayes classifier. Twitter sentiment generally matched the election outcomes, demonstrating the high level of accuracy and precision of the sentiment analysis. Four outliers were found, though, and these were Arizona, Wisconsin, Georgia, and Pennsylvania. These states showed differences between voter sentiment and election results. Further research found that long-term patterns before and after the election showed an increase in favourable attitudes towards the victor and a reduction in favourable attitudes towards the loser. According to the article's findings, social media sites like Twitter can be helpful for forecasting election results, and significant concerns like the economy, the coronavirus, immigration law, the selection of the Supreme Court, and health care systems affected voters' choices.

2.19 Sentiment analysis of political communication: combining a dictionary approach with crowdcoding [19]

The authors faced the problem that Computer-based approaches dominate the field of sentiment analysis, which have a strong language bias as they are developed and validated predominantly with textual data in English language. In this paper, the authors describe how they create a German language sentiment dictionary for the analyses of party statements and media reports. They used crowdcoding and the services of online coders to produce the sentiment ratings of dictionary words.

At the time of writing the the political sentiment dictionary was used in two applications. Parties' use negative campaigning and the tone of media coverage with data from the Austrian National Election Study. With this article the authors showed how to create a dictionary-based measurement procedure for negative sentiment in a language of choice that is cheap, fast, reliable and valid when compared to human coding.

2.20 Regrexit or not Regrexit: Aspect-based Sentiment Analysis in Polarized Contexts[20]

Looking at polarized and polarizing context and contents emotion analysis is a challenge for Natural Language Processing modeling. In this paper, the authors the authors present a methodology to extend the task of Aspect-based Sentiment Analysis (ABSA) toward the affect and emotion representation in polarized settings.

For the first step they focused on the task of detecting emotions in the news context. In addition to that an intensity-level was also attached to the emotion recognised. As a regression problem the authors compared two standard emotion recognition approaches. Namely DepecheMood++ and and the RNN model.

As a context scenario "Brexit" was taken. The procedure consisted of:

- Dataset collection
- Key-concepts
- Results and discussion

They have been able to capture stereotypical aspect-based polarization from newspapers regarding the Brexit scenario using biased key-concepts with their approach.

2.21 Quantifying polarization across political groups on key policy issues using sentiment analysis[21]

The above paper is a study measures the level of polarization among different political ideologies in the United States regarding four key policy issues using sentiment analysis. The writers collected tweets related to immigration, climate change, gun control, and abortion posted between January 2016 and March 2017 and filtered them to ensure that they were related to the identified four key policy issues (pre-processing). Next, they classified the tweets to political groups based on the Twitter bios of the users who posted them. To measure the level of polarization, the authors performed sentiment analysis on the tweets using the VADER lexicon, which is a rule-based approach that assigns positive, negative, or neutral sentiment to text. The sentiment scores were compared between Democrats and Republicans, and the authors analyzed the degree of polarization on each policy issue. The results of the study show that there is significant polarization across political groups regarding the four policy issues analyzed. The authors found that tweets posted by Democrats and Republicans differed significantly in sentiment, and the polarization was more pronounced for some policy issues than others. Specifically, there was a higher degree of polarization regarding immigration and gun control than abortion and climate change. The study has several implications for understanding political polarization in the United States. The findings suggest that sentiment analysis can be a useful tool for quantifying the degree of polarization across political groups on key policy issues, including elections. Furthermore, the study highlights the need for policymakers to address the increasing polarization in the political landscape and to find ways to bridge the divide between different political groups which are key considerations in determining the future leaders of a powerful nation.

2.22 Techniques for sentiment analysis of Twitter data: A comprehensive survey[22]

The paper "Techniques for sentiment analysis of Twitter data: A comprehensive survey" discusses several different techniques for sentiment analysis of Twitter data, including:

1. Lexicon-based approaches: Lexicon-based approaches use sentiment lexicons or dictionaries, which contain a list of words or phrases and their associated sentiment scores, to perform sentiment analysis. These approaches assign sentiment scores to individual words or phrases in a tweet and aggregate them to compute an overall sentiment score for the tweet. 2. Machine learning-based approaches: Machine learning-based approaches use algorithms such as Naive Bayes, Support Vector Machines (SVMs), and Random Forests to train a model on a labeled dataset of tweets and then use the model to predict the sentiment of new, unlabeled tweets. 3. Hybrid approaches: Hybrid approaches combine lexicon-based and machine learning-based techniques to improve the accuracy of sentiment analysis. For example, a hybrid approach may use a sen-

timent lexicon to identify words with known sentiment and a machine learning algorithm to classify the remaining words in a tweet. 4. Deep learning-based approaches: Deep learning-based approaches use neural networks to perform sentiment analysis. These approaches have shown promising results in recent years, particularly for tasks such as sentiment classification of short texts like tweets.

References

[1] Guimaraes, A., Balalau, O., Terolli, E., Weikum, G. 2019. Analyzing the Traits and Anomalies of Political Discussions on Reddit. Proceedings of the International AAAI Conference on Web and Social Media, 13 (1), 205-213. https://doi.org/10.1609/icwsm.v13i01.3222

[2] Ravinder A., Aakarsha C., Shruti K., Shaurya G., Pratyush A. 2019. The Impact of Features Extraction on the Sentiment Analysis, 2019 International Conference on Pervasive Computing Advances and Applications - PerCAA, Procedia Computer Science. https://doi.org/10.1016/j .procs.2019.05.008

[3] Nandwani, P., Verma, R. 2021. A review on sentiment analysis and emotion detection from text. Social Network Analysis and Mining, 11 (81). https://doi.org/10.1007/s13278-021-00776-6

[4] Hofmann, V., Schütze, H., Pierrehumbert, J. B. 2022. The Reddit Politosphere: A Large-Scale Text and Network Resource of Online Political Discourse. Proceedings of the International AAAI Conference on Web and Social Media, 16 (1), 1259-1267. https://doi.org/10.1609/icwsm.v16i1.19377

[5] Endsuy, R. 2021. Sentiment Analysis between VADER and EDA for the US Presidential Election 2020 on Twitter Datasets. Journal of Applied Data Sciences 2 (1), 8-18. https://doi.org/10.47738/jads.v2i1.17

[6] Belcastro, L., Branda, F., Cantini, R.. Marozzo F.. Talia D.. Trunfio P. 2022. Analyzing voter behavior on social media during the 2020 US presidential election campaign. Soc. Netw. Anal. Min, 12 (83). https://doi.org/10.1007/s13278-022-00913-9

[7] Varsha S., Vijaya S., Apashabi P. 2015. Sentiment Analysis on Twitter Data. International Journal of Innovative Research in Advanced Engineering, 1 (2), 178-183

[8] Chaudhry, H.N., Javed, Y., Kulsoom, F., Mehmood, Z., Khan, Z.I., Shoaib, U., Janjua, S.H. 2021. Sentiment Analysis of before and after Elections: Twitter Data of U.S. Election 2020. Electronics 2021, 10 (17). https://doi.org/10.3390/electronics10172082

[9] Ebrahimi M., Yazdavar A. H., Sheth A.,2017. Challenges of Sentiment Analysis for Dynamic Events. IEEE Intelligent Systems, 32 (5), 70-75. https://doi.org/10.1109/MIS.2017.3711649

[10] Tachaiya, J., Gharibshah, J., Esterling, K. E., Faloutsos, M. 2021. RAFFMAN: Measuring and Analyzing Sentiment in Online Political Forum Discussions with an Application to the Trump Impeachment. Proceedings of the International AAAI Conference on Web and Social Media, 15 (1), 703-713. https://doi.org/10.1609/icwsm.v15i1.18096

[11] Chen, E., Deb, A., Ferrara, E. 2022. #Election2020: the first public Twitter dataset on the 2020 US Presidential election. J. Comput. Soc. Sc., 5, 1–18. https://doi.org/10.1007/s42001-021-00117-9

[12] Hagemann, L., Abramova, O. 2023. Sentiment, we-talk and engagement on social media: insights from Twitter data mining on the US presidential elections 2020. Internet Research, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/INTR-12-2021-0885

[13] Garimella, K., De Francisci Morales, G., Gionis A., Mathioudakis, M. 2018. Political Discourse on Social Media: Echo Chambers, Gatekeepers, and the Price of Bipartisanship. Proceedings of the 2018 World Wide Web Conference (WWW '18), 913–922. https://doi.org/10.1145/3178876.3186139

[14] Oikonomou, L., Tjortjis, C. 2018. A Method for Predicting the Winner of the USA Presidential Elections using Data extracted from Twitter. 2018 South-Eastern European Design Automation, Computer Engineering, Computer Networks and Society Media Conference (SEEDA_CECNSM), 1-8. https://doi.org/10.23919/SEEDA-CECNSM.2018.8544919

[15] Das, A., Gunturi, K. S., Chandrasekhar, A., Padhi, A. Liu, Q. 2021. Automated Pipeline for Sentiment Analysis of Political Tweets. 2021 International Conference on Data Mining Workshops (ICDMW). https://doi.org/10.1109/icdmw53433.2021.00022

[16] Chandra, R., Saini, R. 2021. Biden vs Trump: Modeling US General Elections Using BERT Language Model. IEEE Access, 9, 128494–128505. https://doi.org/10.1109/access.2021.3111035

[17] Ali, R.H., Pinto, G., Lawrie, E. Linstead, E.J. 2022. A large-scale sentiment analysis of tweets pertaining to the 2020 US presidential election. J. Big Data, 9 (79). https://doi.org/10.1186/s40537-022-00633-z

[18] Chaudhry H.N., Javed Y., Kulsoom F., Mehmood Z., Khan Z.I., Shoaib U., Janjua S.H. 2021. Sentiment Analysis of before and after Elections: Twitter Data of U.S. Election 2020. Electronics 2021, 10 (17). https://doi.org/10.3390/electronics10172082

[19] Haselmayer, M., Jennym M. 2017. Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. Qual Quant, 51, 2623–2646. https://doi.org/10.1007/s11135-016-0412-4

[20] Vorakitphan, V., Guerini, M., Cabrio, E., Villata, S. 2020. Regrexit or not Regrexit: Aspect-based Sentiment Analysis in Polarized Contexts. Proceedings of the 28th International Conference on Computational Linguistics, 219-224. https://doi.org/10.18653/v1/2020 .coling-main.19

[21] Bor, D., Lee, B., Oughton, E. 2023. Quantifying polarization across political groups on key policy issues using sentiment analysis. https://arxiv.org/abs/2302.07775

[22] Desai, M., Mehta, M. A. 2016. Techniques for sentiment analysis of Twitter data: A comprehensive survey. 2016 International Conference on Computing, Communication and Automation (ICCCA), 149-154. https://doi.org/10.1109/CCAA.2016.7813707