Predicting Political Preferences in Tweets

Yeonie Heo, Jennifer Zheng, Oyungerel Amarsanaa

Abstract

Social media provides a useful platform for understanding public opinion. In the political sphere, written posts on such channels can offer valuable insight into the shared views of different political party supporters. This paper focuses on Twitter tweets and describes an approach to identify the political stances of tweets based on their **sentiments** on a wide array of frequency of political discussion topics. By using tweets posted during the 2020 U.S. election cycle, we use a **supervised method** to train a **political preference prediction model** that identifies the political stance of tweets and partitions them into two groups: Democrats and Republicans.

1. Introduction

The use of social media has become an integral part of our daily lives. Social media is commonly used to express opinions and feelings regarding products and services, interests, politics, and more. Twitter, in particular, has proven to be a popular platform for users to express their political views. Despite its short length, there is a significant amount of information contained in tweets that allows us to identify the political perspectives portrayed in them.

In our work, we use NLP techniques to explore the microperspectives in the tweets at a greater depth. In specific, to gain a more nuanced understanding of the tweets, we analyze the perspectives and agreement patterns of the tweets on pre-defined political topics. By considering the frequency of mentions of certain political topics in each tweet and the overall emotion portrayed in the tweet, we combine them with a supervised method to accurately identify the political stances of the tweets.

In order to find politically related tweets, we leveraged Twitter hashtags to filter out tweets relevant to our research. Here, we assumed that specific hashtags included in the tweets reflect the views and content written within the tweets. We acknowledge that a person's political stance is complex and oftentimes difficult to be attached to a singular label. For the purposes of this paper, we take into account the political preference statistics of voters from the 2020 presidential election and assume that supporters of Donald Trump are likely to be Republicans and supporters of Joe Biden are likely to be Democrats.

Our paper is structured as follows: Section 2 presents prior relevant work on Twitter political party classification and sentiment analysis. Section 3 presents methods for training our Political Preference Prediction model. Our findings and evaluation of the system are illustrated in Section 4.

2. Related Works

One of the early works done on political tweet analysis is by He et al. (2010) in which they implemented a lexicon-based approach that assigns a score +1/-1 to any matched positive/negative word. For each political party, the paper counted the total number of positive and negative mentions to this specific party across all the tweets to define and calculate 'bias measure'. The slight distinction of bias measure for diverse sentiment analysis methods showed that activities on Twitter

cannot be used to predict the popularity of election parties. (Pla and Hurtado, 2014) combine the polarity of political entities mentioned in tweets with the polarity of the overall tweets to come up with a political tendency score in identifying the political orientation of users. They focus on the political stance identification of Spanish language tweets written by influential figures and classify users into four different groups: Center, Left, Right and Undefined. Published in 2017, Pietro et al. create a model to predict Twitter users' political ideology using a 7-point scale: Very conservative, Conservative, Moderately Conservative, Moderate, Moderately Liberal, Liberal, and Very Liberal. The paper is distinct in its use of data collected through primary research methods including questionnaires from Twitter users and its exploration of the relationship between language use and political ideological groups by doing a comparison in 3 ways: comparing the 2 political extremes, 2 political moderates and as well as moderates versus extremes. The paper by (Johnson and Goldwasser, 2016) takes a novel approach to model the tweets of potential presidential candidates. To determine the overall stance, they compile frequently occurring keywords for sixteen political issues and analyze their relation to one another. Our research takes inspiration from their method of identifying and compiling the political topic list. Although not a work of sentiment analysis of tweets, Walker et al. (2012) illustrates a method of constructing data features: it uses LIWC word categories, SentiWordNet, POS tagger, and self-written scripts to annotate features such as polarity, merge ratio, and passive sentence ratio to identify film characters.

3. Data Collection & Training

3.1. Collecting tweets

In order to gather the data for this study, we used a Kaggle dataset of tweets that contained the hashtags #JoeBiden and #DonaldTrump during the first three weeks of the 2020 election cycle, containing 1.72 million tweets.

We initially planned to gather our data using the Twitter API for hashtags #SupportBiden and #SupportTrump published during the one-year period of the 2020 presidential election. The tweets with #SupportBiden would be considered Democratic support, while the tweets with #SupportTrump would be considered Republican support. We discovered, however, that Twitter Elevated access, which requires a separate application and verification process, is required in order to get tweets by hashtag using the Twitter API. Our application for Elevated access was not accepted in time for this paper, thus, in the interests of time and progress, we decided to use this existing dataset from Kaggle. We acknowledge that there would be biases that arise from this shift in the dataset because the hashtag #DonaldTrump won't necessarily indicate a supportive view of Donald Trump, hence a supportive view of Republicans, and vice versa. If we gain access to the Twitter API in the future, we will be able to update the dataset accordingly to increase precision in accordance with our original assumptions.

In addition, to improve the efficiency of model training, 100,000 tweets from the Kaggle Twitter dataset were randomly selected as representative data for training. 80% of all tweets were used for model training and development, while 20% were used to test the performance of our Political Preference Identifier.

3.2. Identifying Political Topics List

To identify political stances in a multidimensional way, we develop a "Political Topic List" with 17 topics that are popular among Democrats and Republicans, including “abortion”, “sustainability”, “nuclear”, “military”, “religion”, “education”, “transportation”, “covid”, “corruption”, “media”, “racism”, “health”, “economy”, “vote”, “peace”, “democracy”, “tax” (Figure 1).

Originally, the list was constructed following the method of (Johnson and Goldwasser, 2016), in which the political issues on the “Political Topic List” were obtained from the 2020 Presidential Election Quiz on ISideWith.com, an independent, nonpartisan election-related website. The key nouns of each relevant question were manually identified and added to the “Political Topic List”. For instance, the word “abortion” was identified to be the key topic from the question “Do you support abortion?”. In the process of topic identification, we considered the repetitions in quiz

question subjects and merged repetitive sub-topics into larger topics. A total of 15 topics were finalized for the "Political Topic List" (Figure 2).

Tax	Corruption	Economy	Abortion	Religion
Education	Media	Vote	Sustainability	
Transportation	Racism	Peace	Nuclear	
Covid-19	Health	Democracy	Military	

Figure 1. Final Political Topic List

Topics	Sub-topics	Quiz Questions from ISideWith.com
Abortion	Abortion	Do you support abortion?
	Planned Parenthood	Should the government continue to fund Planned Parenthood?
Drug	Drugs	Do you support the legalization of Marijuana?
Sustainability	Environment	Should the federal government continue to give tax credits and subsidies to the wind power industry?
	Climate Change	Should the government increase environmental regulations to prevent climate change?
	Paris Climate Agreement	Should the U.S. withdraw from the Paris Climate Agreement?
Guns	Guns	Do you support increased gun control?
Immigration	Border Security	Do you support stronger measures to increase our border security?
	Border Wall	Should the U.S. build a wall along the southern border?
	Immigrant Healthcare	Should illegal immigrants have access to government-subsidized healthcare?
	Homeless Shelter	Should homeless individuals, that have refused available shelter or housing, be allowed to sleep or encamp on public property?
Nuclear	Nuclear Weapon	Should the U.S. conduct targeted airstrikes on Iran's nuclear weapons facilities?
	Nuclear Energy	Do you support the use of nuclear energy?
Military	ISIS, War	Should the U.S. formally declare war on ISIS?
	Military	Should the government increase or decrease military spending?
Safety	Police	Should funding for local police departments be redirected to social and community based programs?
	Criminal Conviction	Should convicted criminals have the right to vote?
	NSA	Do you support the Patriot Act?
Marriage	Same Sex Marriage	Do you support the legalizatiaon of same sex marriage?
Income	Pay	Should employers be required to pay men and women, who perform the same work, the same salary?
	Minimum Wage	Should the government raise the federal minimum wage?
	Taxation	Should the U.S. raise taxes on the rich?
Healthcare	Medicaid	Should the federal government increase funding of health care for low income individuals (Medicaid)?
	Childhood Vaccination	Should the government require children to be vaccinated for preventable diseases?
	Affordable Care Act	Do you support the Patient Protection and Affordable Care Act (Obamacare)?
Religion	Religion	Should a business, based on religious beliefs, be able to deny service to a customer?
Education	Student Loan	Would you support increasing taxes on the rich in order to reduce interest rates for student loans?
Transportation	Public Transportation	Should the government increase spending on public transportation?
Covid-19	Covid-19 Vaccination	Should the government require employees of large businesses to be vaccinated from COVID?

Figure 2. Initial Political Topic List

However, during the training stage, the low accuracy of our original model indicated the possibility that the original topics identified were not highly relevant to our Twitter dataset. The original list (Figure 2) was updated by removing topics with low total frequency and adding new topics by manually inspecting tweets in the dataset. To achieve this, we counted and visualized the total frequency of each topic and excluded the topics with a frequency under 500 counts (Figure 3), which appear in fewer than 5% of the total processed tweets. This led to the elimination of “drug”, “gun”, “immigration”, “safety”, “marriage”, “income”, and “healthcare” from the list. In the case of adding more relevant topics, after a manual inspection of 1% of the total collected tweets and insights summary, the following 10 new topics were added to the Political Topic List: “corruption”, “media”, “racism”, “health”, “economy”, “vote”, “peace”, “pandemic”, “democracy”, and “tax”.

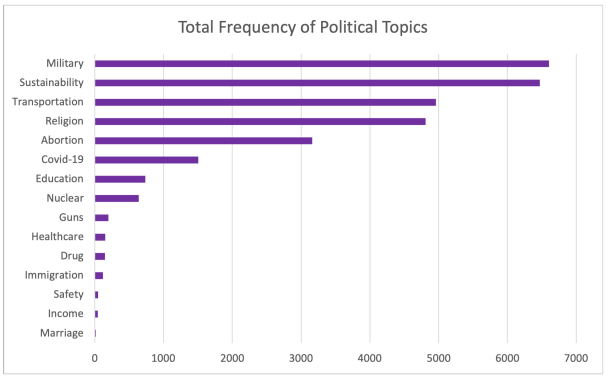

Figure 3. Total Frequency of Political Topics

3.3. Track Political Topics Activities and Involvement

Given the 17 political topics listed in the "Political Topic List", the next step would be to determine the frequency of each topic word in each tweet. In this case, **spaCy**, an open-source library, is used to search for semantically similar words to each topic in the tweet and increment its frequency accordingly. To more effectively capture base topic words regardless of variations in tense, affixes and forms, all pre-processing steps such as stemming were performed on the topic words as well as words in each tweet.

After tracking the frequencies of each topic in the tweets, feature vectors were formed based on the the frequencies of topics. For example, if there are two frequencies of "abortion" and one of "religion" in a tweet, 2 is assigned to topic 'abortion', 1 to topic 'religion', and 0 for the rest of the topics for this tweet.

3.4. Sentiment Analysis

For the purpose of tracking the sentiments towards each topic in a tweet, the overall sentiment score for each tweet is calculated through a popular open-sourced package **VADER**. Among all the tweets, a substantial portion of the tweets – 43% to be exact – had zero sentiment scores. As zero sentiment scores don't indicate an opinion over topics, it is possible that it will not contribute to the model training and rather weaken the model's accuracy. To decide if zero-sentiment-score tweets should be discarded, the model accuracy scores were calculated and compared between the dataset with zero-sentiment-score tweets and the dataset without. We witnessed a decrease in the accuracy from 57.22% to 51.58% when zero-sentiment-score tweets were included. This has led us to build the training model solely with the 62,325 tweets with non-zero sentiment scores.

After the sentiment score is finalized, it is applied to weigh the frequency vector, in which the weighted vector will be used in the final training stage to identify the sentiment on each political topic for each tweet.

$$\text{Sentiment Score of a Political Topic} = X * Y$$

X = Overall Sentiment Score of a tweet
Y = Proportion of frequency of a topic to the total frequency

For instance, if the frequency vector for a given tweet is [1, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1] ordered accordingly to the topics in the Political Topic List and the overall sentiment of the tweet is 0.49, the weighted sentiment vector becomes the following: [0.49, 0.98, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.49, 0, 0, 0.49].

For model training, we use **SVC, Naive Baynes, Logistic Regression** and **K-nearest Neighbor algorithms**, which are the three of the most popular binary classification algorithms. The best-fit model was determined by comparing metrics derived from the models.

4. Political Preference Prediction model Evaluation

4.1. Testing the Political Preference Prediction model

The remaining 20% of tweets were used to test the Political Preference Prediction model.

1. The same sentiment analysis was performed to obtain the sentiment score features and the topic-related word frequency features were counted for each test tweet. Once we have prepared the features from the training stage, we plug them into the model and predict the political preference of each tweet.

2. The political preference of the testing tweets is already explicitly known from the data collection, so we calculated the accuracy of our system output against the answers. In this way, our Political Preference Prediction model was able to be objectively evaluated.

3. Besides the internal evaluation, an external evaluation was conducted by comparing the accuracy of our system with existing research in the literature as well as verifying the means to improve the Political Preference Prediction Model.

4.2. Findings and Improvements

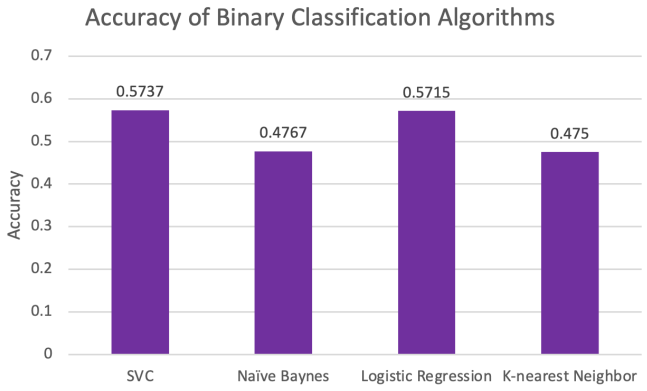

Figure 4. Accuracy of Binary Classification Algorithms

Four binary classification algorithms mentioned above – SVC, Naive Baynes, Logistic Regression, and K-nearest neighbor – are used to test the performance of our Political Preference Prediction model. The algorithms yielding the highest accuracy were **SVC** and **Logistic Regression** – between 56% to 57%. This result was significantly lower than our expectations and demonstrated that our system perform only slightly better than a random guess. During this first iteration, we identified factors that may have influenced the accuracy score. The paper describes some successful changes to methods for building the model, as well as a few unsuccessful explorations that failed to enhance performance accuracy.

Our first concern was the low accuracy of our predictive model as a result of using a small dataset of only 10,000 tweets. As such, we increased our dataset size by roughly 6 times, into around 62,325 tweets, and reran the training model program. The accuracy dropped by 4% to 5% to become 51.71% when the same SVC and Logistic Regression algorithms were used as before. Based on this trial, it has been demonstrated that the pivotal reason for the low accuracy does not stem from the sample size. A decrease in accuracy with a larger dataset suggests that the features need to be revised.

Furthermore, concerning the associated limitations of our novel way of calculating the weighted vectors, we proceeded by rebuilding our training model with the unweighted original vectors of topic frequency. After processing under identical machine learning algorithms of SVC and Logistic Regression, the accuracy still remained low at 51.71%. This trial presented the inefficacy of weighting the vectors of frequencies in the training stage.

The political ideology prediction system produced by Pietro et al. (2017) took into account the levels of political engagement and found out that the system accuracy highly improves for tweets with more extreme political views, achieving near 90% accuracy, whereas for moderate groups, the classification problem was a lot harder where the accuracy score was between 58% and 68%. This clearly highlights the limitation of using tweet data without further information on users and writers of corresponding tweets. With the absence of data on political engagement, our prediction model lacks a significant feature for the training model. However, considering that our system uses all tweets alike irrespective of the political engagement level of the user, the accuracy of our prediction model at 58% which is within the range produced by previously published classification models strongly suggests an overall reasonable performance.

5. Conclusion

This paper built a Political Preference Prediction model and tested its accuracy. In contrast to most previous work, we attempted to train a model by exploring a number of supervised machine-learning techniques. While SVC and Logistic Regression show the highest accuracy, there are some improvements to be done for higher performance of the prediction model. As highlighted above, additional features regarding political engagement level would have influenced the accuracy. Researchers may also take into consideration of characteristics and style of speech in tweets. Given an increasing number of concatenated or minted words and expressions in text data of tweets, our selected machine-learning tools could have had limitations in understanding the overall flow of the context, leaving a possibility of misinterpretation of the sentiment score of the tweet. Next, the political topics we revised could have been outdated and questionable for their objectivity in covering a wide range of neutral political topics. Despite the explicit statement of objectivity from the source of our generated political topics, some topics may be more or less favored or discussed by certain political parties, building bias in the training phase of our research.

While our study focused solely on tweets posted during the election in 2020, follow-up work can use other modalities such as broadcasting companies with more accurate and clearer usage of speech to improve prediction performance. Predicting the preference of broadcasting companies with the corresponding binary classification algorithms may produce intriguing insights for followers of each broadcast to reevaluate their political stances. Another direction of future study will look at the political variances in other countries and cultures, where much less work has been done universally.

References

[1] Daniel Preoţiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. 2017. Beyond Binary Labels: Political Ideology Prediction of Twitter Users. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 729–740, Vancouver, Canada. Association for Computational Linguistics.

[2] Ferran Pla and Lluís-F. Hurtado. 2014. Political Tendency Identification in Twitter using Sentiment Analysis Techniques. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 183–192, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

[3] Kristen Johnson and Dan Goldwasser. 2016. Identifying Stance by Analyzing Political Discourse on Twitter. In Proceedings of the First Workshop on NLP and Computational Social Science, pages 66–75, Austin, Texas. Association for Computational Linguistics.

[4] Marilyn Walker, Grace Lin, and Jennifer Sawyer. 2012. An Annotated Corpus of Film Dialogue for Learning and Characterizing Character Style. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 1373–1378, Istanbul, Turkey. European Language Resources Association (ELRA).

[5] Yulan He, Hassan Saif, Zhongyu Wei, and Kam-Fai Wong. 2012. Quantising Opinions for Political Tweets Analysis. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 3901–3906, Istanbul, Turkey. European Language Resources Association (ELRA).