Classification of Tweets of Politicians from Northern Europe

University of Rochester
Data Science Capstone Project 2023

By: Harshal Talele, Ayush Singla

I. Abstract

In today's digitally connected world, Twitter has become a big source of political expression. This research paper aims to explore various methods to predict the correct political spectrum based on user tweets from Northern Europe. An accurate classification of political views is extremely important for understanding public opinion and shaping policy decisions.

In order to achieve this objective, we have used a comprehensive dataset of tweets collected over a specific period of time. This data set includes diverse political ideologies. We utilized exploratory data analysis to better understand the data and political views of the users across various domains. This helped in gaining clarity on the nuances of political discourse on Twitter. Our study also takes into account different natural language processing techniques and machine learning algorithms to process and analyze the textual part of the data i.e. user tweets. It includes sentiment analysis, topic modeling and finally predicting the political orientation of these users.

The methods deployed helped in shedding light on the effectiveness of using Twitter data as a valuable resource for classifying individuals along the political spectrum. Further, the research provides a good benchmark model to predict the political discourse of users that can be used by organizations across the world in order to take data driven decisions and form better strategies

II. Introduction

The study delves into the realm of predicting the political inclination of users in Northen Europe by using information from Twitter's unstructured textual data. It basically addresses two fundamental questions: What are the underlying patterns and sentiment trends across different countries and gender? And what is the best NLP technique and machine learning algorithm to predict correct political spectrum labels based on user tweets.

In order to address the first question, we have used exploratory data analysis to discover meaningful insights related to political views of these politicians. An average tweet length is about 20 words with 89 being the max tweet length. This gives us an idea about the high number of words politicians have used to express their political views. 'dkpol' is the most common hashtag being used for all 7 countries with 50% being the average distribution among all the top 10 hashtags. Further, we looked at political views for each country in detail to and found that Iceland is the only country with 'independent' political view.

For the second question, we have deployed various machine learning techniques and NLP to discover the model with best accuracy that can predict the political inclination of users. Techniques like lemmatization and topic modeling were used to structure and clean the data that can help in taking key decisions for the organizations.

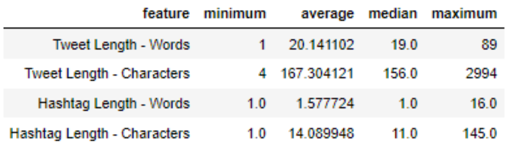
In the pages that follow, we present our methodology, findings, and implications for predicting political view labels using Twitter data. Our study contributes to the evolving field of political analysis, highlighting the potential of social media as a valuable resource for understanding political beliefs in our digital age.

III. About the Data

The data consists of all tweets posted by politicians of seven different Northern European countries: Belgium, Denmark, Iceland, Ireland, Netherlands, Norway, and Sweden which is unstructured and contains lot of stop words, punctuations etc. In addition to that, it contains gender, country,hashtags and the target variable which is political inclination of the user. Each country is associated with a different number of tweets. The test set consists of %20 of the tweets originating from each country (the remaining 80% is the training set. In total, there are 407,223 tweets in the training set (and 101,808 tweets in the test set). This makes a total of 509,031 tweets

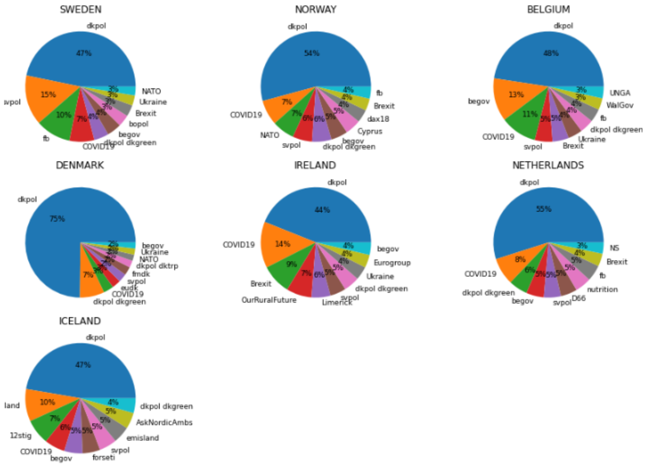
Throughout the study we identified multiple interesting patterns and trends that shaped the latter part of the research.

We looked at user tweet length and calculated metrics like minimum, average, median, and maximum for both characters and words. This helped in understanding the distribution of the tweet length data which further gave a sense of how popular social media platforms like Twitter is among these politicians. The same was performed using hashtags data.

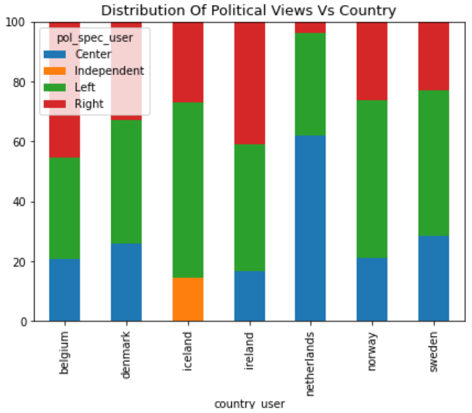


feature	minimum	average	median	maximum
Tweet Length - Words	1	20.141102	19.0	89
Tweet Length - Characters	4	167.304121	156.0	2994
Hashtag Length - Words	1.0	1.577724	1.0	16.0
Hashtag Length - Characters	1.0	14.089948	11.0	145.0

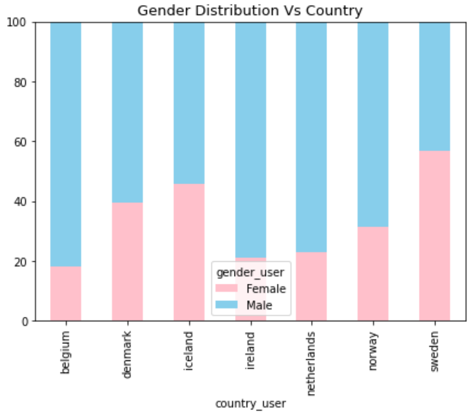
The above table provides two useful insights: The first is the average length of tweets in terms of words is 20 with 89 being the maximum showing that Twitter is a popular platform for expressing one's views. The second is the average length of tweets in terms of characters is 167 which means that the data is not clean and there is a lot of scope to make it better to read and understand.



Here, we looked at top 10 most commonly used hashtags and calculated the distribution of these hashtags across these 7 countries. 'dkpol' is the most commonly used hashtag across all the 7 countries. In Denmark, 75% of the top 10 hashtags is dkpol which makes sense as 'dkpol' means 'Danish Politics' or 'Politics of Denmark'. Interestingly politicians from other countries are talking about Denmark. Other prominent hashtags are COVID19, svpol etc.



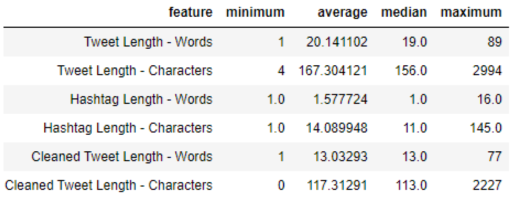
The distribution of political views across all the 7 countries gave interesting insights on the political inclinations of these users. Iceland is the only country with users having independent political view. Its the same country where we observed max percentage of users having left political views. The maximum percentage of users with center political view was observed in Netherlands while it is Belgium for right political views.



Here, we have created the distribution of gender across all the 7 countries to understand gender ratio for these countries which might help in further predicting their political inclinations. Clearly, Belgium has maximum percentage of users being male (80% approx). On the other hand, Sweden has maximum percentage of users being female (60% approx). It is also the only country among all the 7 who has majority of users being female.

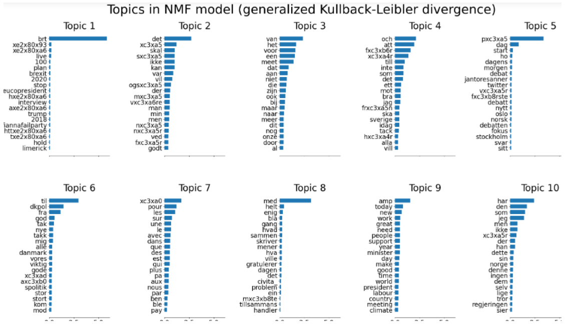
IV. METHODS

Now that we have performed the exploratory data analysis and gained some interesting insights from the data, this is where we started cleaning the tweets data and applied topic modeling that enabled us to look through multiple topics and organize, understand and summarize them at scale. We could quickly and easily discover hidden topical patterns that were present across the data, and then used them as part of our model development process.

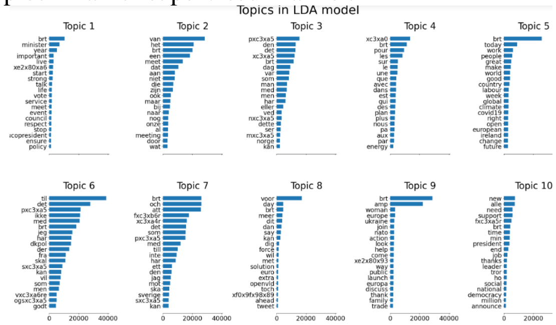


feature	minimum	average	median	maximum
Tweet Length - Words	1	20.141102	19.0	89
Tweet Length - Characters	4	167.304121	156.0	2994
Hashtag Length - Words	1.0	1.577724	1.0	16.0
Hashtag Length - Characters	1.0	14.089948	11.0	145.0
Cleaned Tweet Length - Words	1	13.03293	13.0	77
Cleaned Tweet Length - Characters	0	117.31291	113.0	2227

As part of cleaning process, we removed stopwords, all words that are shorter than 3 characters,all links (starting with http), emojis and punctuations from the tweets using WordNetLemmatizer. Calculated the minimum, average, median, and maximum for the newly created cleaned tweets data to get a better understanding of the data. The average tweet length in terms of words dropped from 20 to 13 and the maximum length from 89 to 77.



The final step before building machine learning model was to perform topic modeling and therefore, we used LDA and Non-negative Matrix Factorization techniques for topic analysis. Set the number of topics to 10 and extracted the topics in an unsupervised manner.



Compared both the techniques and found that in contrast to LDA, NMF is a decompositional, non-probabilistic algorithm using matrix factorization and belongs to the group of linear-algebraic algorithms. NMF works on TF-IDF-transformed data by breaking down a matrix into two lower-ranking matrices. The results are very different which can be seen through clusters formed above.

The next section of the study talks about the model building process where we applied various text classification algorithms. This paper though, will cover only the top 3 models we build and their accuracy scores.

In addition to the text classifiers, we have used CountVectorizer to transform a given text into a vector on the basis of the frequency of each word that occurs in the entire text and TfidfTransformer that systematically computes word counts using CountVectorizer and then compute the Inverse Document Frequency (IDF) values and only then compute the Tf-idf scores.

In order to build and test the model, the given data set was split into 80:20 ratio having train and test sets respectively. The model was trained using the train set and tested against the test set.

A. Multinomial Naive Bayes:

A popular technique in Natural Language Processing, it guesses the tag of a text using the Bayes theorem and then calculates each tag's likelihood for a given sample and outputs the tag with the greatest chance.

Combinations of parameters like max_features in CountVectorizer and alpha in MultinomialNB were used to identify the best combination for the model that would give the highest accuracy

B. Logistic Regression:

A simple and most effective technique in machine learning, it estimates the probability of an outcome based on input features using a logistic function. It's widely used in machine learning for its effectiveness in classification tasks.

Parameters like max_iter and the regularization parameter were used to get the best possible combination for the logistic regression model

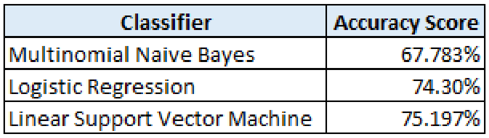
C. Linear Support Vector Machine:

Finally, we used Linear SVM which turned out to be the best classifier for this task. It separates data points into different classes by finding the hyperplane that maximizes the margin between them. It is very effective for high dimensional data.

Regularization parameter was used to identify the best model.

V. RESULTS

Accuracy_score method was used to calculate the accuracy of each model.



Classifier	Accuracy Score
Multinomial Naive Bayes	67.783%
Logistic Regression	74.30%
Linear Support Vector Machine	75.197%

The highest accuracy was obtained for Linear SVM which is 75.964%. The accuracy for Logistic Regression was close to Linear SVM i.e. 74.3% and therefore, could also be a good option to classify this kind of textual data.

VI. CONCLUSION

In conclusion, our study underscores the significance of predicting 'political spectrum' labels by using the power of Twitter's tweet data. Through natural language processing and machine learning techniques, we have delved into the realm of text classifiers to understand the political inclinations of users. These insights hold immense value for organizations globally. As we work towards refining our models, we are mindful of the uncertainties that politics bring and can play a huge role in shaping these predictions. Our ongoing efforts have the potential to shape and enhance the accuracy of these predictions