Deep Pollster:
Political Orientation Prediction

Shivchander Sudalairaj
Sagar Panwar

Advisor: Anca Ralescu

Abstract

The global use of Online Social Networks to publish information in the form of blogging and microblogging to exchange opinions, by the general public, news media and popular politicians alike, has given rise to new directions of research in Computational Political Science.

In this research, we re-examine the problem of quantifying and predicting the Political Leaning of a particular Twitter user. By studying the political leaning of twitter users, it is possible to target advertisements at individuals, shape digital profiles, and deliver news, articles, views and products that are individualistic and personalized. This could also be used to predict the political outcome of an election by predicting the leaning of users in a geographical location.

Background Assumptions

- Users are consistent in their actions of tweeting, retweeting about political issues
- Users with similar political alignment tend to retweet and favorite others' tweets
- Tweet composed by the user is assumed to be one's stance and will be considered with higher weight
- Similarly retweets and favorite tweets will be considered with lesser weights
- W(tweets) > W(retweets) > W(liked tweets)
- If a user follows more politicians from one side of the spectrum than the other, then they are considered to have similar political alignment

Challenges with Existing Systems

Sentiment analysis on tweets to assess political leaning has its disadvantages :

- Does not paint a complete and holistic picture of the users' ideological views
- Cannot build a digital profile of a user from a single or even with a temporal series of tweets
- Assessing political leaning of a demography does not serve the purposes and intents of individual orientation

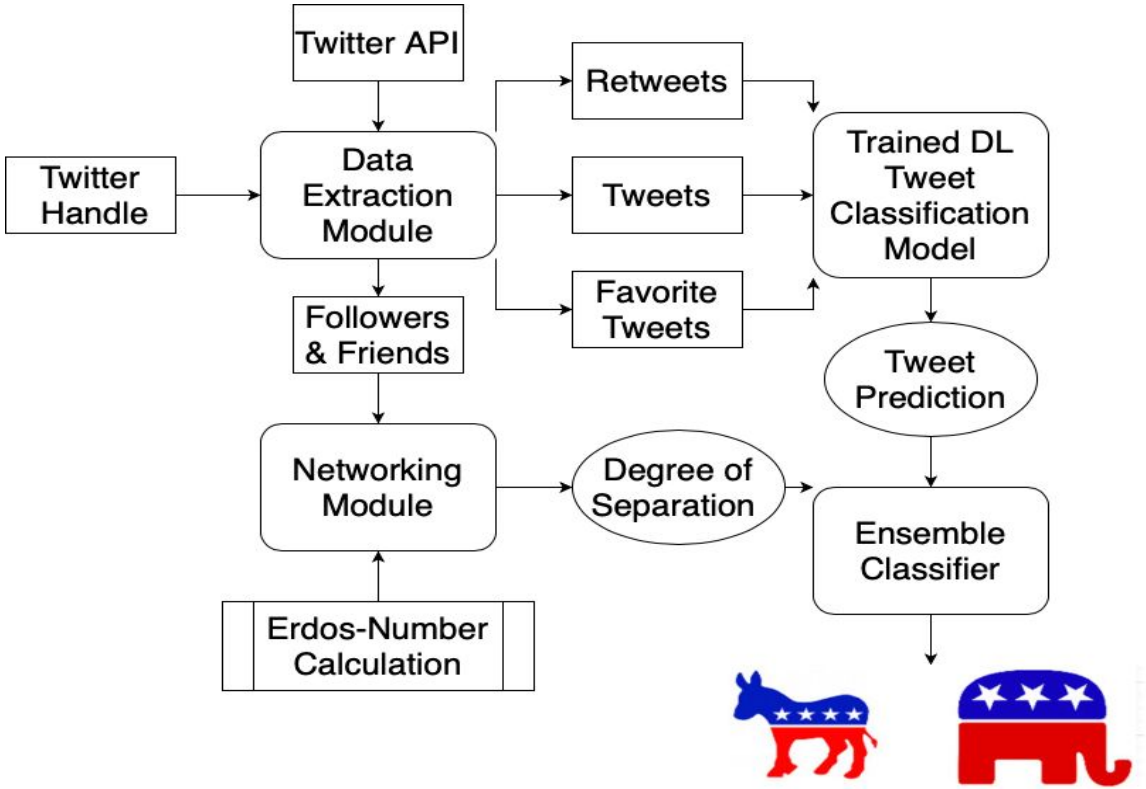
Proposed Solutions

We leverage more than just tweet-retweet maximization, or a network matrix :

- Binary classification of the latest tweets, retweets and liked tweets on the basis of political leaning
- Identifying the degree of separation between the user and politicians from both the political sides

Our proposed system is curated to provide a more rounded and holistic sense of the individual user, painting an overall picture of their digital profile, leading to potential in marketing and business spheres.

Architecture



Technologies

Dataset

The dataset was collected from Kaggle (Political Social Media Posts) which provides us tweets from politicians' twitter accounts, along with human judgments about the purpose, partisanship, and audience of the messages.

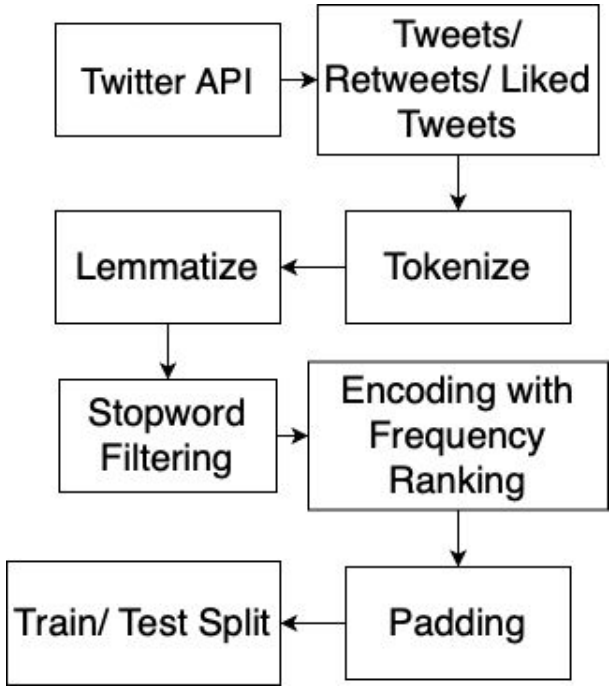
Total number of tweets = 100,000

Total number of politicians = 500 (250 Democrats/ 250 Republicans)

Total number of tweets/politician = 200

Data Extraction

- Tokenization: Tweets are split into words, split using space while handling punctuations
- Normalization: Words are then stemmed (removing affixes) and lemmatized (reduced to canonical forms)
- Encoding: Words are placed in a frequency distribution and the words are replaced by their ranks

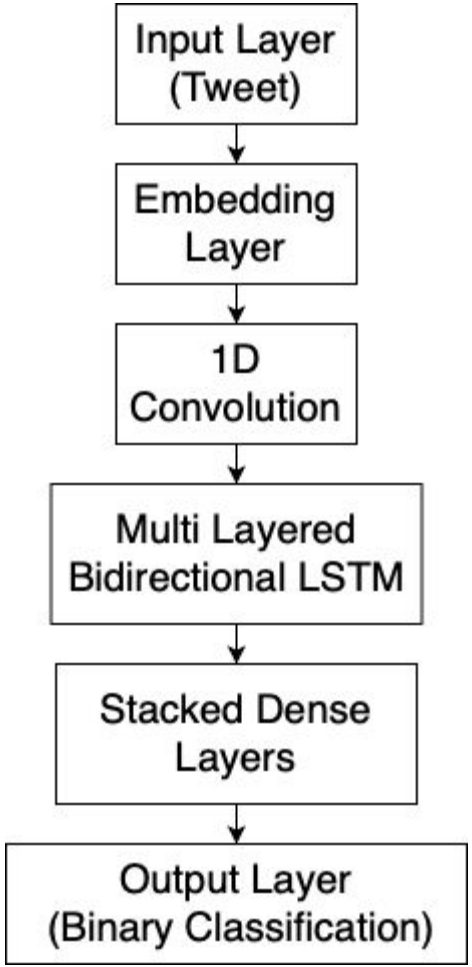


Classification Model

Since our input (tweets) are sequences of words, it was natural to use a Recurrent neural network.

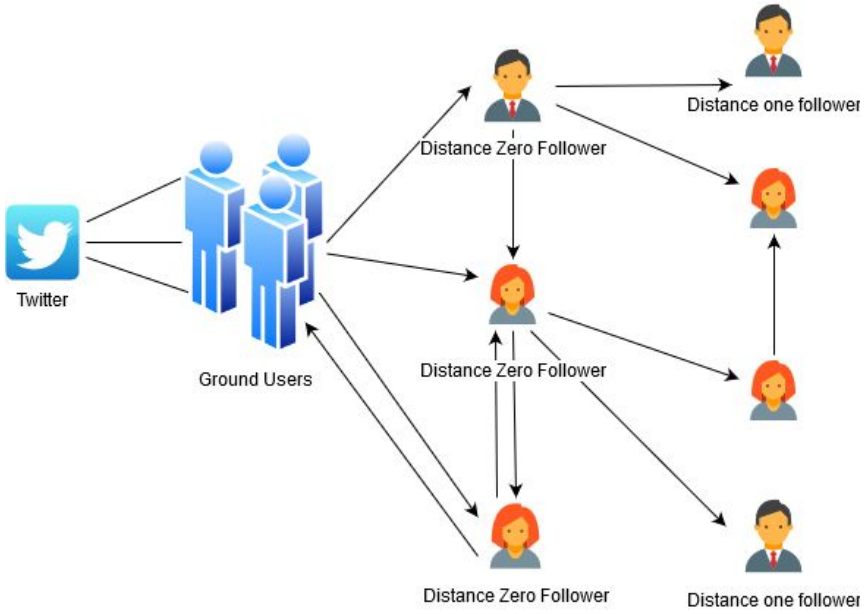
After many experiments we found that a bidirectional LSTM (Long Short Term Memory), a type of RNN, gave us the best results. This model can pick up context from the tweets.

After feeding the input through the Bidirectional layers, we use stacked fully-connected layers to project this vector to one-dimension. Finally we pass it through a sigmoid activation to get a (0,1) output.



Networking Module

For an optimal solution, we carry out two breadth first searches; one from the source (user) and the other from the destination (politician). Once a node is found at the intersection, we calculate the number of hops required which will be the degree of separation.

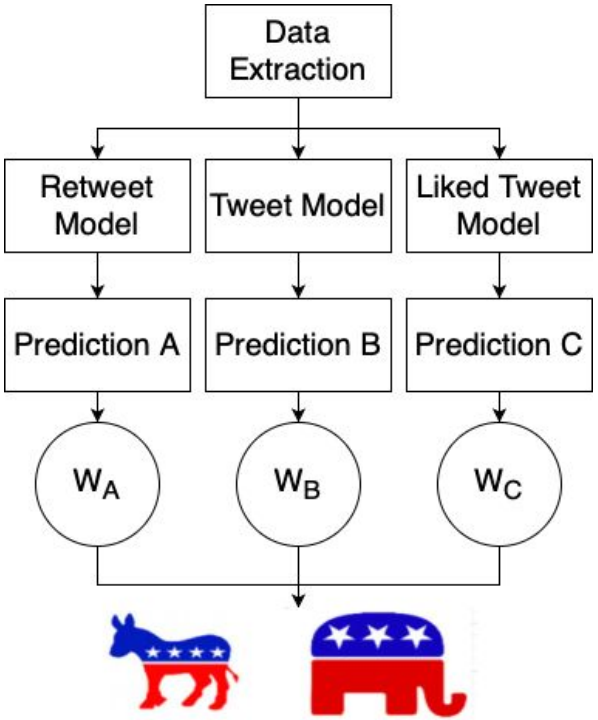


Ensemble Classifier

The individual models are trained using bagging (Parallel training with different training sets)

The ensemble classifier learns independently from each other in parallel and combines them following results using a deterministic averaging process with specific weights

W (tweets) > W (retweets) > W (liked tweets)



Testcase

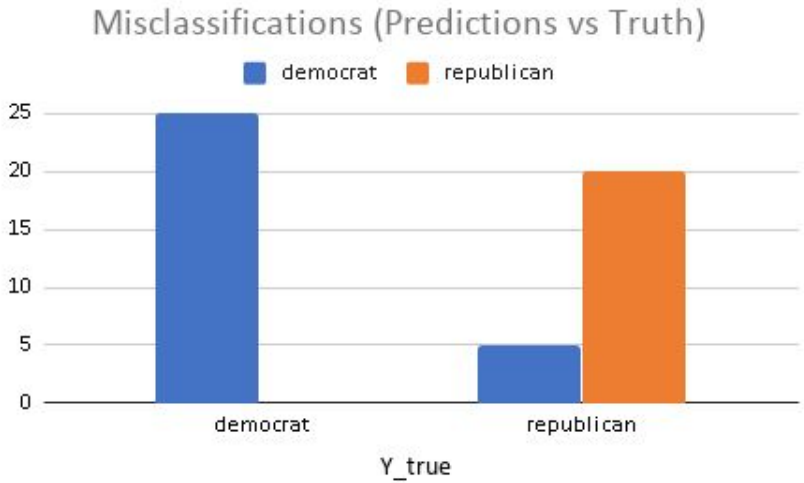
- The model was tested with multiple sets of test cases to eliminate any innate bias
- Each sets of test cases consists of a set of 50 previously untested politicians' twitter handles

Hypothesis : *Politicians are relatively consistent with the language models that they follow while publishing tweets on twitter*

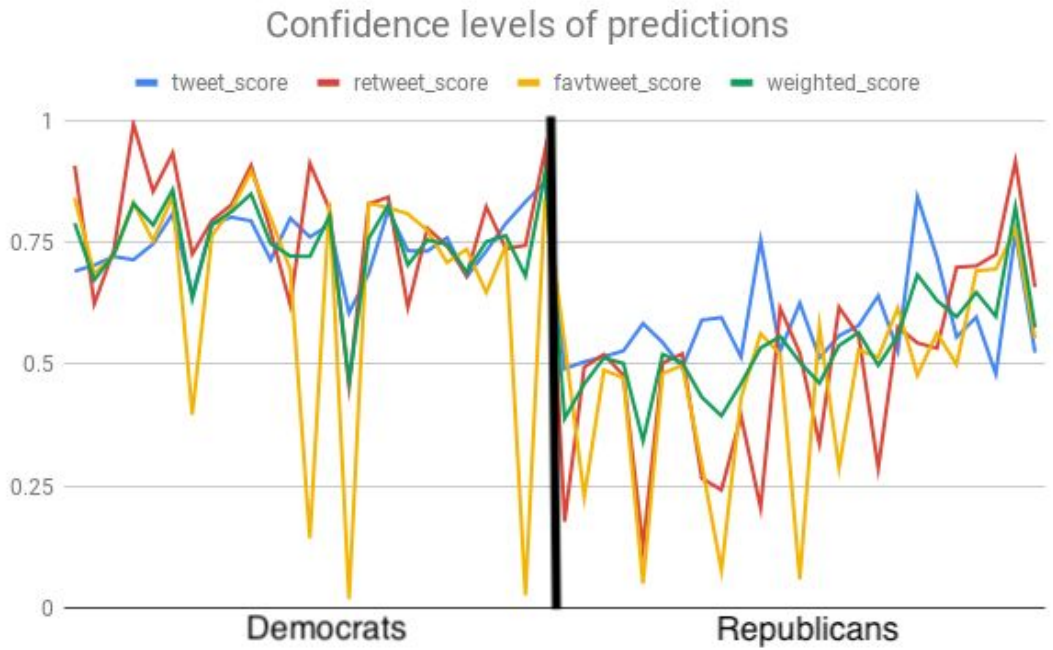
Results

After running our tests, we observed that our model was able to predict democratic politicians with 100% accuracy

While the misclassifications arose with republican politicians resulting with an accuracy of 80%



Results



After running our tests, we observed that our model was able to identify democrats with significantly higher confidence than republicans.

We also observed that predictions from tweets model was more stable than retweets and liked tweets. This supports our initial hypothesis for variable weights

Inference

- Democratic tweets consistently fall on the political left
- Republican tweets falls more on political centre and centre-right
- Democrats are more consistent with their language and in turn political ideologies used on twitter
- Republicans use language which is more ambiguous and tend to waver between left and right of the political spectrum
- Naively extrapolating this model to general public users will cause inherent bias

Future Work

- The handling of tweet content analysis and classification can further be improved to handle spam accounts. It is also to be noted that sarcasm is yet to be handled by our model and additions could be made to account for this
- To develop a generalized model to be applicable for general public users, we would need to survey users and find out their political orientation to develop a more general dataset and retrain the model using the dataset

Thank You