COGS 118A- Project Proposal

Project Description

Peer Review

You will all have an opportunity to look at the Project Proposals of other groups to fuel your creativity and get more ideas for how you can improve your own projects.

Both the project proposal and project checkpoint will have peer review.

Names

- Cameron Faulkner
- Nikhil Hegde
- Qianxi Gong
- Atul Nair

Abstract

This section should be short and clearly stated. It should be a single paragraph <200 words. It should summarize:

We are trying to assess what keywords and tweet sentiments result in a viral tweet amongst politicians. We will be using tweets from congresspeople from twitter using a scraper. We would have to clean the raw data and adapt it to a readable data frame so that we can manipulate the data for our data science workflows (EDA, feature selection, etc). We will be creating a performance metric based on the number of likes, retweets and followers to assess virality.

Background

Previous research has demonstrated that sentiment analysis of Members of Parliament (MPs) s' tweets can be used to predict their popularity. In particular, studies have shown that factors such as negative sentiments is a key factor that influences the level of retweets received by MPs [1]. These findings suggest that analyzing the words and language used by MPs in their tweets can provide insights into their strategies for gaining popularity on social media platforms.

Given our real-world political knowledge, we suspect that the use of certain words and language patterns may have different impacts on the popularity of tweets from MPs of different parties.

To explore this further, we are interested in investigating whether the distribution of language use and popularity among different parties can be modeled using a mixture of supervised and unsupervised methods. By examining the use of specific words and phrases across tweets from different political parties, it is possible to score a tweet from a MP given their party, by template matching language patterns from that party and model how they will be reacted upon by the public [2].

Potential footnotes:

Problem Statement

When politicians make public statements, they often seek to maximize the reach of their message and build their personal brand. The appeal of doing so is obvious and certain members of Congress have made careers, and presidencies, out of being the metaphorical loudest voice in the room. However, some statements fail to gain traction and leave their originators looking out of touch with their constituents and the political atmosphere.

We are attempting to solve the uncertainty of whether a Tweet made by a sitting member of Congress will receive high engagement, defined in terms of likes, replies, and quote Tweets. By training a model with sentiment analysis of Tweets, word frequency analysis, partisan affiliation, and engagement metrics, we believe that we will be able to accurately predict the extent of engagement a potential Tweet will have.

Data

We will be collecting the Tweets of all 534 American members of the 117th session of Congress with Twitter accounts over the duration of the Congress: January 3, 2021- January 3, 2023. To do so, we will be using the snscrape scraper package which can neatly collect a specified user's Tweets.

Each Tweet will be an observation in the dataframe that includes the text content of the tweet, as well as its likes, retweets, and quote tweets. We will also be labeling the tweets as coming from either a Republican or a Democrat (there were no Independents in the 117th congress) as we theorize there may be a difference in the factors that lead to viral tweets based on party affiliation.

- Link to the source of Congressional Twitter accounts

Proposed Solution

One potential solution to the problem of predicting tweet engagement among members of Congress is to use a machine learning algorithm that incorporates sentiment analysis, word frequency analysis, and partisan affiliation.

One specific approach to this problem could involve using a supervised learning algorithm, such as logistic regression or random forest, to predict the likelihood of a tweet going viral based on its features. The features could include sentiment analysis scores for each tweet, as well as the frequency of certain keywords or phrases. Other features could include the partisan affiliation of the member of Congress who tweeted the message, the number of followers they have, and the time of day the tweet was posted.

To implement this solution, one could use a variety of libraries and functions. For example, the NLTK library could be used for sentiment analysis, while scikit-learn could be used for machine learning.
To test the accuracy of the model, one could split the data into training and testing sets and evaluate its performance using metrics such as accuracy, precision, recall, and F1 score. It may also be helpful to compare the performance of the model against a benchmark model, such as a simple baseline that predicts the most common class (e.g. non-viral).

Overall, the success of this solution will depend on the quality of the data and the features selected.

Evaluation Metrics

One evaluation metric that can be used to quantify the performance of both the benchmark model and the solution model is the F1 score. In the context of predicting tweet engagement, precision represents the proportion of predicted viral tweets that are actually viral, while recall represents the proportion of actual viral tweets that are correctly predicted as viral. The F1 score balances these two metrics, giving equal weight to precision and recall, and provides a single score that summarizes the overall performance of the model.

F1 = 2 * (precision * recall) / (precision + recall)

where precision = true positives / (true positives + false positives)

and recall = true positives / (true positives + false negatives)

True positives are the number of tweets that are correctly predicted as viral, while false positives are the number of tweets that are predicted as viral but are actually non-viral, and false negatives are the number of tweets that are actually viral but are predicted as non-viral.

For example, if the solution model correctly predicts 80% of viral tweets with a precision of 75%, and correctly predicts 70% of non-viral tweets with a precision of 90%, then the F1 score would be:

precision = 0.75

recall = 0.8

F1 = 2 * (0.75 * 0.8) / (0.75 + 0.8) = 0.77

This indicates that the model has a relatively good balance between precision and recall, with a reasonable overall performance. By comparing the F1 score of the benchmark model and the solution model, we can assess the effectiveness of the solution in improving predictive accuracy.

Ethics & Privacy

One potential ethical concern with this project is the potential for unintended bias in the data, particularly in labeling tweets as coming from either a Republican or a Democrat. This could lead to the perpetuation of stereotypes and reinforce political divisions.

In addition, the potential for highlighting nationalistic and nativist speech due to its ability to generate high engagement may preserve some of the worst political tendencies in this country.

Team Expectations

Put things here that cement how you will interact/communicate as a team, how you will handle conflict and difficulty, how you will handle making decisions and setting goals/schedule, how much work you expect from each other, how you will handle deadlines, etc...

- Cameron Faulkner: Twitter Data Collection
- Nikhil Hegde: training model
- Qianxi Gong: Implementation of sentiment analysis and feature parsing
- Atul Nair: Model evaluation and selection

Project Timeline Proposal

2/28: finish data collection
3/5/23: finish model construction
3/8: complete checkpoint requirements
3/12: conduct model selection and analysis
3/15: begin final write up

Footnotes

[1] Antypas D. , Preece A., Camacho-Collados J. *“Politics, Sentiment and Virality: A Large-Scale Multilingual Twitter Analysis in Greece, Spain and United Kingdom”.* Online Social Networks and Media. August 22, 2022. https://arxiv.org/pdf/2202.00396.pdf

[2] Gangwar, A. and Mehta,T.. *“Sentiment Analysis of Political Tweets for Israel using Machine Learning”.* LearnByResearch. April, 2022. https://arxiv.org/pdf/2204.06515.pdf