Predicting Political Partisanship on Social Media

Machine Learning Project Proposal

MSc Data Science & Business Analytics – ESSEC Business School | CentraleSupélec

Rafaëlle Aygalenq
B00724587

Sarah Lina Hammoutene
B00712035

Dóra Linda Kocsis
B00714326

Noémie Quéré
B00719656

MACHINE LEARNING CONCEPTS

Natural Language Processing, Logistic Regression, Support Vector Machines, Naive Bayes, Random Forests, Neural Network, Boosting

KEYWORDS

text mining, sentiment analysis, social media, politics, partisan bias, classification, machine learning models, Natural Language Processing

1 PROBLEM STATEMENT

Our project aims to apply text mining and sentiment analysis techniques to a dataset of political social media posts. Primarily, we are interested in setting up a predictive algorithm for establishing the existence or non-existence of partisan bias, specifically by examining which words contribute in what way to our classification model.

The dataset [1] in place aggregates 5,000 Twitter and Facebook messages from US Senators and other American politicians. Initially, we are provided with a breakdown by human judgement into audience (national, constituency), bias (neutral / bipartisan or biased / partisan) and message substance (e.g. informational, supportive, policy oriented, etc.).

The topic under investigation is becoming increasingly important in the age of voter microtargeting. Social Media is the number one medium of communication and information for the average consumer. Can one argue, that those digital channels used by politicians are official statements of the state administration or should they be evaluated as personal opinions? Going further, we can analyze whether these mediums are used to express political partisanship to influence the opinions of followers. These are the questions we aim to shed light on with our predictive algorithm.

[1] Source: Crowdflower – Data for Everyone Library: https://www.crowdflower.com/data-for-everyone/

Having described the dataset and the challenges to be tackled, we will now present and refine the questions investigated in this project. First, we would like to see if social media posts can be classified based on message substance and whether this can predict the existence of bias in a legislators' social media behavior (Fig.1. shows our preliminary analysis on the subject). Our secondary aim would be to establish which channel (i.e. Facebook vs Twitter) is used for what kind of purposes.

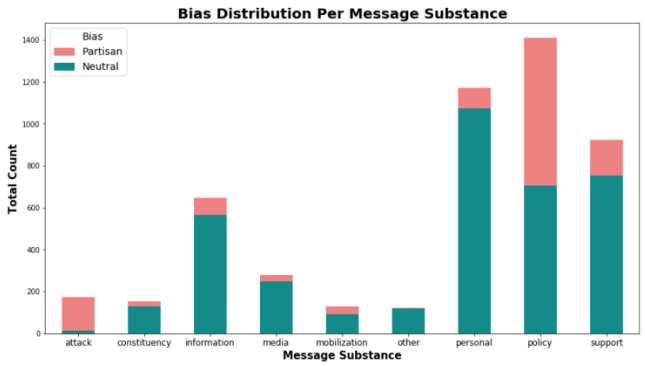

Figure 1. Distribution of bias according to message substance

2 APPROACH

After the initial analysis, we carried out a basic data cleaning process. First, we removed all the insignificant features and not specified values and changed the format of some of the features to extract more information (e.g. the politicians' names, position (Senator, Representative of the House) and states served).

In an attempt to enrich our dataset, we decided to include new features. By using a secondary dataset, we extracted and matched the party affiliation and the gender to the author of the message. This way, we hope to uncover new correlations between the features and expand the scope of our research.

Then, the focus of our work will be on cleaning and analyzing the social media posts themselves. The idea, is to create a Bag-of-Words representation of the data. From the text of the tweets and the Facebook posts, we extract all the words that appear to build our dictionary. We consider each one of these words as a new feature. Once our feature columns updated, we compute the frequency of the words in each message. By the end of the process, we obtain what is known as frequency matrix [1]. To achieve this, certain operations must be performed upstream:

- **Cleansing** [2]**:** removing URLs, names, tweets with Not Available text, special characters, numbers.
- **Text processing:** transform to lowercase, solve the misspelling errors, tokenization[2], stemming[3].
- **Build word list for Bag-of-Words** by first filtering out the most common stopwords and then building our frequency matrix.

As a next step, we divide our dataset into different samples. Usually three samples are used: the training sample on which the model will be estimated, the validation sample used to optimize the parameters and the testing sample used to test the model estimated with the training sample. The latter permits to calculate the prediction error and therefore to compare the prediction capacity of several models. This step is very important and is indispensable at least for the distinction between the training sample and the testing one to have an unbiased estimation of the prediction error. In our case, because of the small size of our dataset, it has been decided to remove the validation sample. For the validation process, cross-validation will be used. Cross-validation consists of splitting the training sample in K folds and for each fold:

- estimating the model on $K - 1$ folds
- using the last fold for the validation step.

Regarding the dataset split, we decided to take 80% of the dataset for the training sample and 20% for the testing one. The next step will be to set the number of the folds, K, by testing different values and looking at the variance of the prediction error.

Lastly, we will look at the implementation of different models (suited for similar classification problems [3 - 4]) and choose the most accurate ones. These models will be logistic regression, support vector machines, Naive Bayes, random forests, neural networks and boosting. After evaluating each of them, we might implement a more sophisticated version aggregating the best performing models that capture the true effect of political partisanship on Social Media.

3 EVALUATION

The first evaluation tool that can be used is the confusion matrix. It refers to a specific table used to visualize the performance of a model where each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class. This matrix leads to rates of false positive, false negative, true positive and true negative used to compute accuracy scores.

Another evaluation method consists of computing the ROC (Receiver Operating Characteristic) curve which highlights the true positive rate against the false positive rate. This curve leads to the AUC (Area Under Curve) score giving an idea of the model's performance.

Considering the existing research related to the topic, we could find relevant work on the text mining techniques implemented on Social Media posts and we aim to apply these during our assignment. Our work primarily falls into two parts: text mining and machine learning algorithms. With the aim to tackle both we hope to implement a model with sufficient predictive power.

REFERENCES

[1] Calvo, R. and Mac Kim, S. (2012). EMOTIONS IN TEXT: DIMENSIONAL AND CATEGORICAL MODELS. Computational Intelligence, 29(3), pp.527-543.

[2] Gokulakrishnan, B., Privanthan, P., Ragavan, T., Prasath, N. and Perera, A. (2012). Opinion mining and sentiment analysis on a Twitter data stream. Advances in ICT for Emerging Regions (ICTer), 2012 International Conference on, pp.182 - 188.

[3] Patodkar, V. and I.R, S. (2016). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. IJARCCE, 5(12), pp.320-322.

[4] A., V. and Sonawane, S. (2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. International Journal of Computer Applications, 139(11), pp.5-15.

[2] Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called tokens, perhaps at the same time throwing away certain characters, such as punctuation. Source: Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

[3] Stemming usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes. Source: Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.