**Finding sentiments in political speeches**

Statement of Work (V1)

By

Gbemisola Banjoko (100766479)

**Executive summary**

This section provides a general overview of this document. This document contains details about the Statement of Work (V1) for the project titled “Finding Sentiments in Political Speeches”. Political campaign rallies provide opportunities for candidates to be able to win their voters through their speeches. Voters can connect or disconnect from candidates based on how they feel about their sentiments towards various issues discussed by the candidate. Having a tool that can access and measure the overall sentiments in political speeches will be valuable for any political candidate.

The purpose of this project is to develop a machine learning (ML) solution that can identify and classifying positive and negative sentiments in political speeches.

Details about the background to the study, problem statement, data requirements and deliverables are provided in this document.

**Rationale statement**

Political speeches create an interaction between the speaker and the audience. This interaction is however different than the type of interaction that happens in conversations. The audience react to the speaker by acting collectively to cheer, applause, laugh or show disappointment. The audience display their approval of the candidate based on the speech. The speech is also an important feedback for the mass media to analyse. Therefore, it is important for politician to maintain a steady flow of positive sentiments in their speeches. Identification of sentiments is important to prevent against negative campaigning. Members of the audience who connects positively to the political candidate are likely to vote for them. The end product of this project will provide a tool for politicians to detect the overall sentiments of their speech, giving them the opportunity to amend their speech to a suitable level of positive sentiments, to enable positive connections to their voters.

**Statement of problem**

You are a data scientist for a campaign organisation. Your assignment on this project is to conduct text mining on political rally dataset and find positive and negative sentiments in the speeches. Develop an ML model that can be used to detect and classify future political speeches into positive or negative sentiments, as the identification of sentiments is important so as to prevent against negative campaigning. Future speeches will be run through the developed model to determine the number of positives and negative sentiments, and to determine the overall sentiment of the speech.

**Data requirements**

Political speech data in form of raw text will be needed for the project. The data can already have a sentiments column or not.

* Available data: President Donald Trump’s rally speech data is already available. The data has been downloaded from Kaggle.com (<https://www.kaggle.com/christianlillelund/donald-trumps-rallies>).

**Assumption**: the dataset will consist of more positive sentiments than negative sentiments.

**Limitation**: This project is limited because we only have dataset from one individual, therefore, the model might not be generalizable for multiple politicians.

**Data processing**

The downloaded data is in form of raw text. Therefore, this data needs to be mined and processed before feeding it to the machine learning (ML) model. the following steps needs to be undertaken to ensure that the data is fit for ML model application.

* Read in raw text data using python application.
* Clean raw text data by removing punctuations, unnecessary characters and stopwords
* Create Word Clouds
* Sentiment identification; find and identify positive and negative sentiments in dataset
* Create visualization
* Vectorize data; consider n grams and TF-IDF
* Conduct feature engineering

Once the dataset is completely processed, the data will be split into training set and test set. Based on the accuracy of the test data. 80% of the dataset will be used for training while 20% will be used for testing.

**Model selection and evaluation**

Conduct K-fold cross validation and consider at least 3 different ML algorithms. Evaluate the algorithms and select the best one for your final product. Based on the accuracy score of the algorithms, there might be a need to develop hyperparameters to improve the performance of the algorithm. This section will be updated as the project progresses

**Project plan**

|  |  |  |
| --- | --- | --- |
| **Upcoming Task** | **Deliverables** | **Due date** |
| Data Aqusition and Understanding + SOW (V2) | Perform exploratory data analysis to identify main characteristics within the data.  Determine preliminary data manipulations requirements (i.e. harmonize, rescale and/or clean)  Perform statistical analysis to identify data patterns and correlations.  Discuss the benefits of feature engineering for optimal model output.  Identify key candidate features between synthesized and real data for model input.  Update statement of work with new information. | NOV.23, 2020 |
| Modelling | Evaluate learning algorithms and frameworks to help solve the problem statement.  Assess and select appropriate software tools to successfully work with the data and model(s).  Discuss model(s) architecture and software pipeline needed to successfully create the proposed solution.  Assess dataset assumptions, limitations, and constraints in order to develop effective models. | NOV.23, 2020 |
| Prototyping | Prototype the proposed model architecture using the selected software.  Begin feeding data into model architecture and observe output. Begin necessary model tweaking in order to ensure desired results.  Evaluate model and consider alternatives if necessary. | NOV.23, 2020 |
| Deployment | Development of software pipeline to host proposed model solution (cloud-based).  Development of solution endpoint to be consumed via service by the final user.  Submit a document with only your GitHub link.  Be ready to be invited to present your project demo in a 1:1 meeting with Marcos Bittencourt and 2 other invitees. | DEC.18, 2020 |

**Note: This is an ongoing project, and this document will be updated as the project progresses.**