**Predicting Twitter Sentiment of the US public about 2016 Presidential Candidates**

**Hemanth Inukonda Kamalakannan   - A20369096**

**Sreekumar Selvam                                 - A20372235**

**Srilekha Napa Ugandhar                       - A20370759**

**ABOUT THE PROJECT**

As all of us know that the interesting and most sensational topic currently running in the United States is the Presidential Election,2016. So, we decided to take the most sensational topic in the field and to predict the sentiment of the public after carefully examining, experimenting and analyzing them. We are excited about how the public views the highest election candidates, namely Donald Trump, Hillary Clinton. Feelings toward candidates fluctuate swiftly as interviews, debates, responses to world routine, and different disorders come to the front. To reap a colossal, diverse dataset of current public opinions on the candidates, we made up our minds to use Twitter. Twitter supplies us with residing entry to opinions about the election across the globe. We're examining contemporary tweets related to the highest candidates in the 2016 election to predict the general public sentiments towards each candidate. In addition, we are aiming for a better, extra primary evaluation of sentiments toward candidates than the media and polls have. We are expanding on the Positive, negative and neutral emotions.

**OBJECTIVE:**

* Predicting sentiments of the US public for Hillary Clinton and Donald Trump in the Presidential election 2016.
* Multi­class classification allows us to categorize the tweets into multiple distinct classes to handle a more specific range of emotions. We modelled the task as a multi­class classification problem to account for an array of different sentiments that people could realistically feel regarding the election candidates.
* This will be performed using algorithms “SVM” , “KNN” and “Naïve Bayes classifiers” to come up with the most efficient approach for analysing.
* In addition word clouds are formed using various emotions of the public towards the candidates and from which various sentiments of public can be analysed and produced.

**APPROACH:**

1. **COLLECTION OF DATA:**

* Using Twitter Search API keywords like “TRUMP, DONALD TRUMP, CLINTON, HILLARY CLINTON, PRESIDENTIAL ELECTION 2016” will be collected for Trump and Clinton respectively.
* To get our info information, we downloaded tweets utilizing the Twitter API found from a hunt of our watchwords or the full name of a 2016 presidential competitor. For both testing and preparing input, we utilized as information the content of tweets
* The data output is in a JSON format of which we are interested only in the TWEET text and Date.
* We will be using an N-Fold Cross Validation to train the data set in the below steps.

1. **CLASSIFICATION OF TOKENS:**

* Once the sentences are normalized and tokenized, the tokens are assigned an F-Score.
* Based on the F-Score we plan to use the following classification techniques:

a) KNN Classifier.

b) Naive Bayes.

c) Support Vector Machine (SVM).

* Finding the most accurate and efficient classification using MAE ,precision and recall.

1. **SENTIMENT ANALYSIS:**

* The tweet texts of Hillary and Trump are then analysed into ‘POSITIVE’, ‘NEGATIVE’ or ‘NEUTRAL’ sentiments.
* Dividing the words of the groups POSITIVE, NEEGATIVE and NEUTRAL for Hillary, Trump separately, we have 6 unique groups.
* For Example – Hillary and POSTIVE, Trump and NEGATIVE, etc.
* For each of the 6 groups create a WORD CLOUD using WORD FREQUENCY.
* The word clouds are then analysed to predict the most influential topics with their attached sentiments for each candidate.

**CHALLENGES AND SOLUTIONS:**

* Potential challenge in the project could be the precision rate. The rate of precision based on the algorithm could be low due to the complexity involved in feeding the data. The algorithm basically depends on the data which is fed into it and it predicts the result from it. But there are chances where some data could be found in the training data hence the prediction rate would become low and hence the precision rate will become low.
* As already stated the algorithm predicts the result based on the data fed into it. These data which are being provided would also include the general stop word from the general classification. These stop words thought are limited is not something which could always be fixed. They either keep changing or gets added up. Hence to inclusion of stop words for the analysis is a complex and exhaustive technique. Also the complete inclusion of stop words for the prediction based is also a difficult technique.
* The data that we collected could be in unstructured format. The text data present in the outside world will not be present in the structured format. Hence cleaning the data and converting it into structured format is a tedious and complex process.
* One of the major challenges in this project could be the language complexity and multilingual problems

**SOLUTIONS:**

* In order to deal with the problems faced due to stop word, we could come up with a solution in such a way that after carefully analyzing the high frequency used word( most commonly used English words) we could feed them into the algorithm so that there are chances of better prediction.