UNIVERSITY OF TORONTO

Content Table:

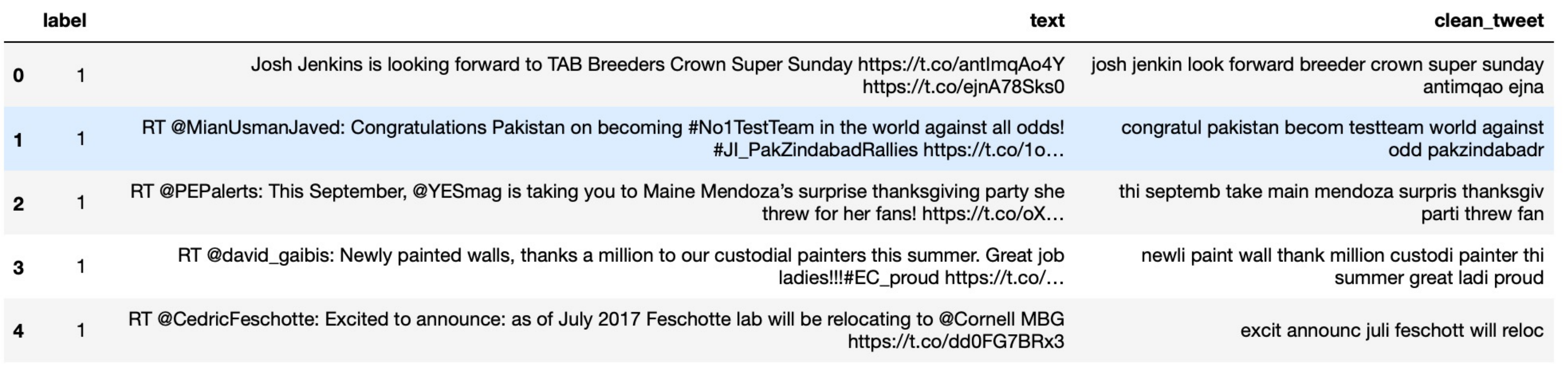
1. Data Cleaning
2. Exploratory Analysis
3. Model Preparation
4. Model Implementation & Tuning
5. Result

Name- TUHIN RANJAN
Program-MEng MIE

UNIVERSITY OF TORONTO

Data Cleaning

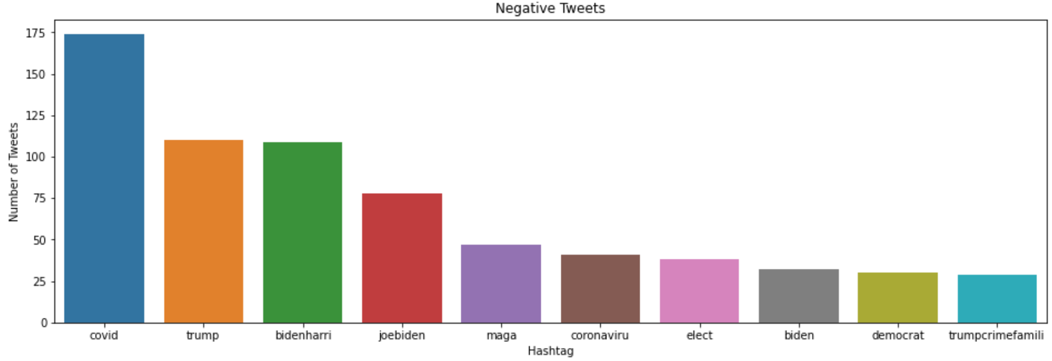
1. The main issue with text data is that it is all in text format (strings). However, Machine learning algorithms need some sort of numerical feature vector in order to perform the task. Basic text pre-processing includes:
   1. Removing Twitter Handles (@user)
   2. Removing Punctuations, Numbers, and Special Characters
   3. Removing Short Words
   4. Tokenzation/Text Normalization (It is just the term used to describe the process of converting the normal text strings into a list of tokens i.e. words that we actually want. Sentence tokenizer can be used to find the list of sentences, and Word tokenizer can be used to find the list of words in strings*)*
   5. Stemming (Stemming is the process of reducing inflected (or sometimes derived) words to their stem, base or root form - generally a written word form. For example, if we were to stem the following words: "Stems", "Stemming", "Stemmed", "and Stemtization", the result would be a single word "stem")

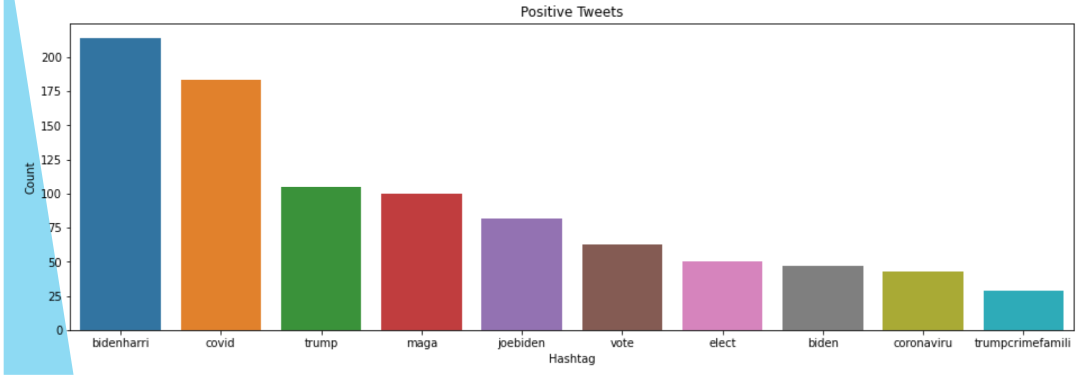


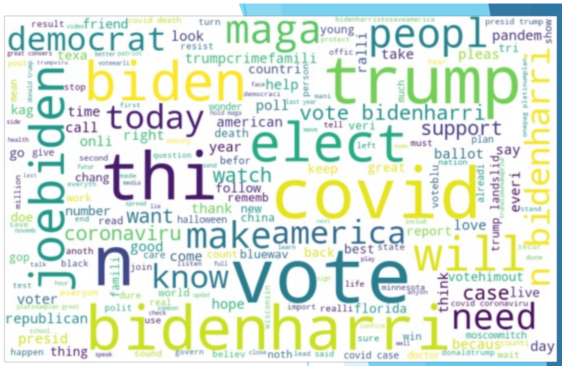
	label	text	clean_tweet
0	1	Josh Jenkins is looking forward to TAB Breeders Crown Super Sunday https://t.co/antImqAo4Y https://t.co/ejnA78Sks0	josh jenkin look forward breeder crown super sunday antimqao ejna
1	1	RT @MianUsmanJaved: Congratulations Pakistan on becoming #No1TestTeam in the world against all odds! #JI_PakZindabadRallies https://t.co/1o…	congratul pakistan becom testteam world against odd pakzindabadr
2	1	RT @PEPalerts: This September, @YESmag is taking you to Maine Mendoza's surprise thanksgiving party she threw for her fans! https://t.co/oX…	thi septemb take main mendoza surpris thanksgiv parti threw fan
3	1	RT @david_gaibis: Newly painted walls, thanks a million to our custodial painters this summer. Great job ladies!!!#EC_proud https://t.co/…	newli paint wall thank million custodi painter thi summer great ladi proud
4	1	RT @CedricFeschotte: Excited to announce: as of July 2017 Feschotte lab will be relocating to @Cornell MBG https://t.co/dd0FG7BRx3	excit announc juli feschott will reloc

Exploratory Analysis

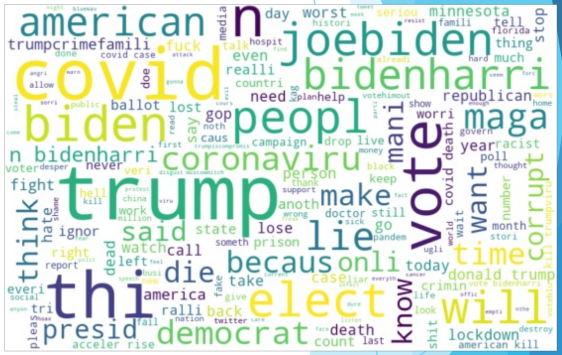
1. EDA or Exploratory Data Analysis can be done on multiple approach as of:
    1. Word of Clouds – Where larger the font, more the frequencies of that word
    2. Hashtag Analysis



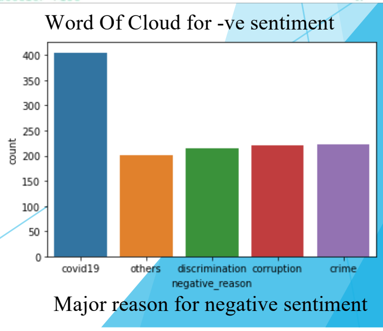




Word Of Cloud for +ve sentiment



Word Of Cloud for -ve sentiment



Major reason for negative sentiment

UNIVERSITY OF TORONTO

Model Preparation

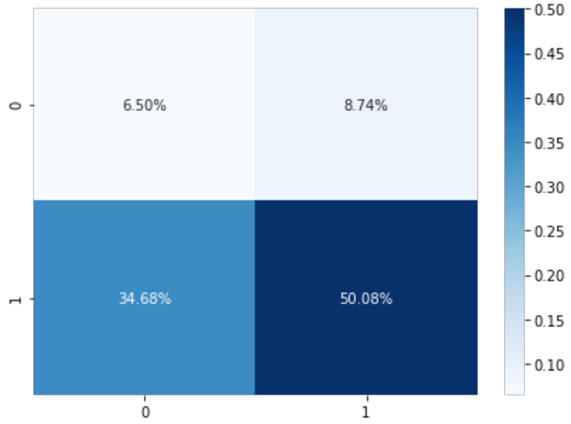
1. **Feature Extraction:**
   1. Bag Of Words Features (In this model, a text (such as a sentence or a document) is represented as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity)
   2. TF-IDF Feature (Some words have high frequency in large text corpus but makes very little meaningful information. If we use the count data directly to train a model, the very frequent tokens with little meaning will have a significant influence over the low frequency but far more interesting terms. So it is necessary to normalize and re-weight the word count result)

2. **Model Building /Preparation (MODEL-1):**
   1. Here we have build up 7 models/classifier to test our training dataset and out of that we will select best model with maximum accuracy to test our test data set. Both BOW and TF-IDF is being used as features to train & test models.
   2. The 7 models are as follows:
      1. Logistic Regression
      2. Naive Bayes
      3. Support Vector Machine
      4. Random Forest
      5. Decision Tree
      6. XGBoost
      7. K-NN

3. **Model Building /Preparation (MODEL-2):**
   1. We have build up 3 models/classifier to test our training negative sentiment dataset and out of that we will select best model with maximum accuracy. BOW is being used as features for the models.
   2. The 3 selected models are as follows:
      1. Random Forest, 2. Decision Tree, 3. K-NN

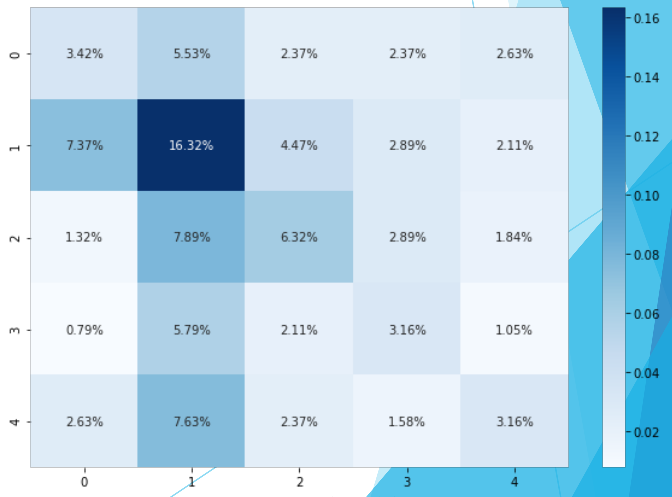
Model Implemention & Tuning

1. Since the performance accuracy of the models like logistic regression, SVM, XGboost and Randon Forest is approximatley same ~97%, thus performing Hyper parameter tuning only on the random forest model as the accuracy is quite high and need not to perform on multiple hyper parameter tuning on different models.
2. Got the accuracy rate of ~57% on the testing data after tuning. Further the confusion matrix and the heat map suggest that the maximum weightage lies into the True Positive section thus reflecting that most of the prediction for the positive sentiment is correct. The reason being the initial training data which has maximum positive tweets which trained the model to execute effectively on the positive sentiment data.



First Model

1. For MODEL-2, out of 3 selected classifier, Random Forest is giving the best accuracy rate for the negative sentiment sub-dataset. After tuning the model, the accuracy is about ~34%.
2. Second model approximately able to predict the correct reason behind the negative sentiment. Referring the "Major reason for negative sentiment" graph in slide 2 , COVID is the main reason for the negative sentiment while the model also predicted the higher proportion for the COVID in the heat map



Second Model

UNIVERSITY OF TORONTO

Result

1. Understanding the first model and its result, it can be seen that the True Positive Value is occupying the maximum percentage thus reflecting that the positive sentiment is on the much higher side. Also we did the the EDA, then we found that the negative sentiments word of clouds is comprising of the words like Trump which indicated that the people are not happy with the Trump administration.

2. The word of clouds also mentions some major negative words as of Kill which may be the reflection of the black community being targeted or killed thus creating the negative impression on the Trump administration.

3. Republic party is getting some bad impression while the Democrat party is in advantage due to COVID mishandling as this comes by analysis of the test data set. Joe Biden and Kamla Harris combined hashtag "bidenharris" is in major proportion in positive sentiments values while Trump and COVID is in major negative sentiment values.

4. Understanding the second model and its result, it can be seen the selected model predicted an accuracy of ~34% while COVID is comprising of maximum weightage in the confusion matrix ~18%. So, behind the negative sentiment COVID is the major reason followed by crime and discriminations.

5. NLP or Natural language Processing either through tweeter or any social media platform gives a great idea to analyze the sentiment of the large demography across any given region, thus helps to predict possible correct consequences of any upcoming event.

6. So answering the research question: The Democrats will be in major advantage and the Republican will probably loose the coming election due to hurting the sentiment of the people because of mishandling of COVID and other racial injustices.

7. Second model approximately able to predict the correct reason behind the negative sentiment. Refering the negative hashtag graph above, COVID is the main reason for the negative sentiment while the model also predicted the higher proportion for the COVID in the heat map.

8. First model accuracy can be improved by considering the larger training dataset with approximately equal number of both positive and negative sentiment data so that it can train the model well with respect to both sentiments. In above selected training set, the major proportion is positive sentiment thus affecting the accuracy of the test data set. Second model accuracy can be improved by taking more rows on the training set data.