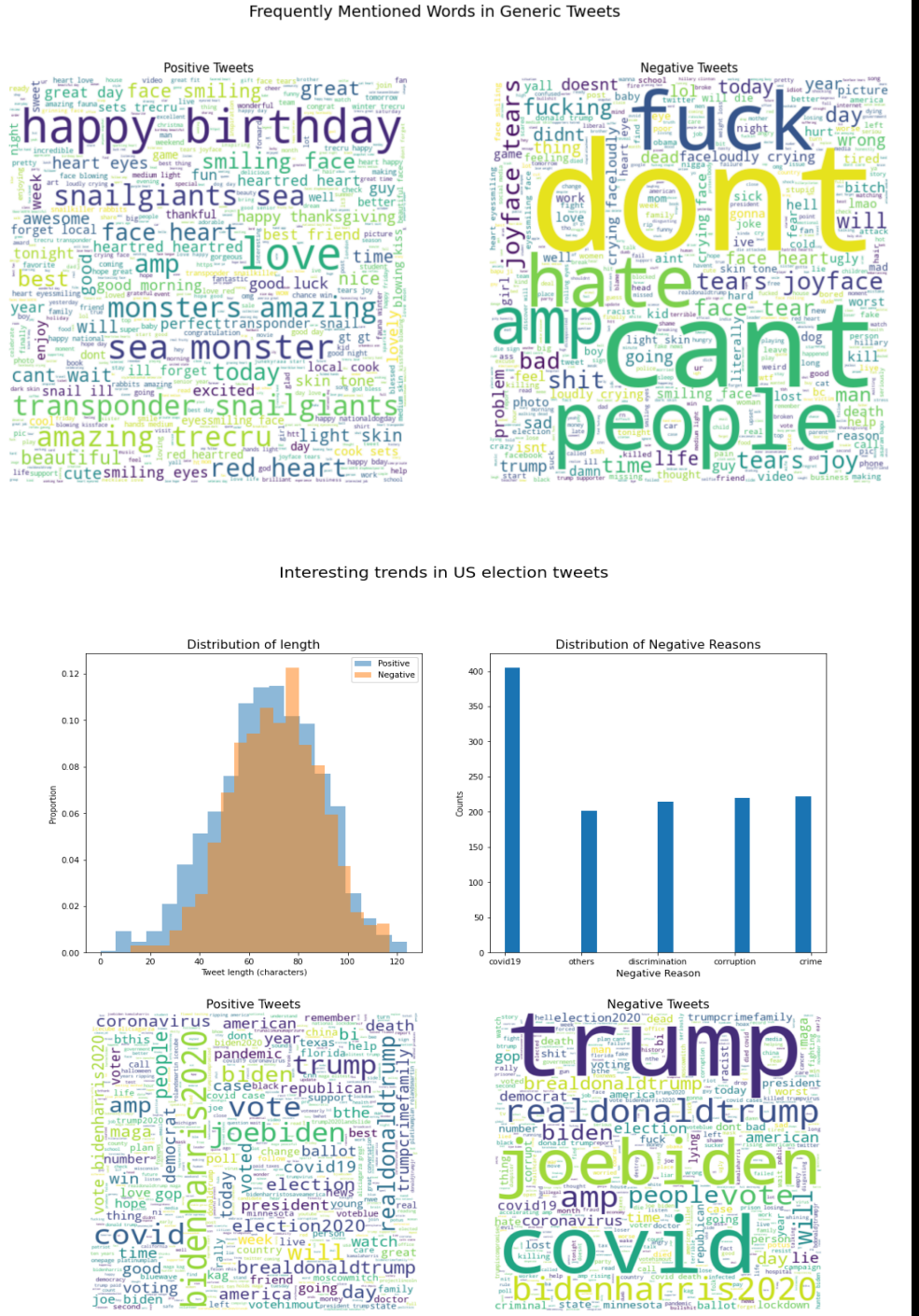
MIE1624 ASSIGNMENT 3

Paul Xie

1002905118



EXPLORATORY ANALYSIS

- Stop words, punctuations, emojis and URLs have been removed from the original tweets;
- Word cloud from generic tweets showed frequently used words in both positive and negative tweets. Positive tweets often contain birthday/holiday wishes, while negative tweets usually contain some discouraging or bad words;
- In the US election dataset, positive tweets and negative tweets have about the same average length, while the length of positive tweets is more spread out;
- The most common negative reason is COVID19, which accounts for about 1/3 of all negative tweets;
- In the negative tweets of US election dataset, the most used words are “trump” and “covid”;
- The word “trump” is not used as often in positive tweets, while “bidenharris2020” is the most used word in positive tweets.
- We may infer that Biden is not exactly well liked, but Trump is very hated by netizens

Model Preparation and Implementation

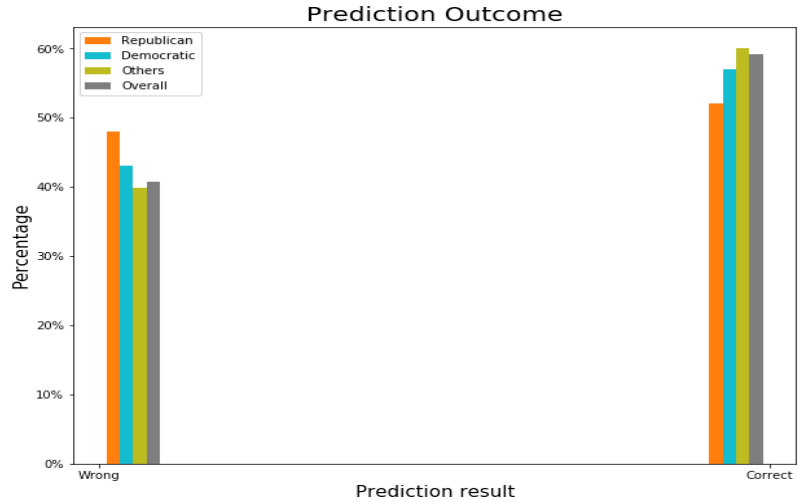
- Want to train classification models to predict the sentiments of tweets;
- Need to have a feature space so we can implement classification algorithms;
- Tried two different techniques to create feature space, bag of words (word frequency) and TF-IDF;
- Train logistic regression, SVM and naïve Bayes models using each technique;
- Tuned all the models so they had the best performance, tested all the models on US election tweets.

Model Comparison and Prediction Results

- Compared models based on their testing accuracy;
- All the models achieved over 85% testing accuracy. Logistic regression and linear SVM are slightly more accurate than naïve Bayes models;
- For a specific model, the performance on both feature spaces were very similar;
- Further tested every model's performance on US election tweets, logistic regression using BOW feature implementation had the highest prediction accuracy. Visualized the prediction accuracy for each political party with this combination;
- Tweets from Republican party supporters had the lowest prediction accuracy (~50%), while our model predicted the correct sentiments for over 60% of the apolitical tweets.

Testing Accuracy			
Feature type	Logistic Regression	Linear SVM	Naive Bayes
Bag of Words	0.8774694460930971	0.8768274809530154	0.8597790670914134
TF-IDF	0.8779418355357986	0.8780023982848629	0.8590341452779224

Prediction results on US election tweets			
Feature type	Logistic Regression	Linear SVM	Naive Bayes
Bag of Words	0.5920846394984326	0.5807210031347962	0.5697492163009404
TF-IDF	0.5877742946708464	0.5865987460815048	0.5681818181818182



Negative Reasons Prediction

Prediction accuracy on negative reasons				
	Logistic Regression	SVM	KNN	Naive Bayes
Training accuracy	0.3382352941176471	0.38235294117647056	0.43552036199095023	0.38122171945701355
Testing accuracy	0.2789473684210526	0.3078947368421053	0.24210526315789474	0.29736842105263156

- Tried to predict the reasons of negativity for negative tweets in the US election dataset;
- Prepared the dataset using TF-IDF as feature space, number of features set to 100 due to small sample size;
- All four models had relatively low testing accuracy, the KNN model seems to be overfitting our data since testing accuracy is a lot lower than training accuracy;
- All four models had low training accuracy too, suggesting our model is undertrained, thus more training samples are needed.

Discussion

- To improve the prediction accuracy, we can change the numbers of features, need to find the right number so that models have low bias and low variance;
- In general, public opinion is more positive towards the Democratic party (Joe Biden) than the Republican party (Donald Trump);
- Supporters of Republican party is less vocal on social media compare to supporters of Democratic party;
- COVID-19 remains the most concerned issue in America, Trump is disliked partly because of the way he handled the pandemic;
- Campaign makers can use NLP to analyze the public sentiment towards their political party, and focus more on the swing states whose opinion is 50-50 split;
- Further, campaign makers can analyze the negative reasons of their political party, so they will know where to improve in their campaign strategy.