Final Project Report

Introduction

For our project we classified tweets of Donald Trump and Hillary Clinton during the 2016 elections. We created graphs, bigrams, and a prediction method that determines the probability of the tweet whether it is written by Trump or Hillary.

Dataset Description

We used the dataset: https://www.kaggle.com/benhamner/clinton-trump-tweets

This dataset provides ~3000 recent tweets and retweets from Hillary Clinton and Donald Trump, the two major-party presidential nominees during the US Election, from January until September 2016.

The dataset provides information such as: date, the tweet text, the handle, whether it was a retweet, amount of likes, amount of retweets, the device the tweet was sent from, and the location of the tweet. The information used for this project was the text of the tweet, who tweeted it, and whether or not it was a retweet.

Baseline approach description

From the dataset we selected the handle, text, is retweet data then using the regress we filtered out and cleaned each sentences. For preliminary data analysis, we cleaned the sentences using regional expressions and found the most common unigrams and bigrams for Donald Trump and Hillary Clinton (fig.1,2,3,4). For this first part, we kept in the hashtags to understand who was using more hashtags since they are important when writing tweets. We created word clouds for each candidate since they are useful in conveying and presenting text information (fig.5,6).

We also created word clouds to see who each candidate tweeted the most. In both Donald Trump (fig.7) and Hillary Clinton (fig.8), we could see that both retweeted person or organization that would benefit their personal campaign. Trump retweets most about Eric Trump and Donald J Trump Jr, which as a son promotes his father Donald Trump. There are also others such as drudge report from Matt Drudge who is a political reporter that promotes Trump's campaign, Dan Scavino who is the person in charge for Trump's Social Media and TeamTrump who tweets from his team itself. While Donald Trump has retweets from a variety of authors, Hillary Clinton focuses more on her branding campaign, which is The Briefing 2016 and HFA which is Hillary for America. There is also Tim Kaine who is running with Hillary Clinton for vice president in 2016.

When looking to create a predictor we cleared out the # signs in order to just get the words since hashtags aren't that useful when creating a predictor. Then each of the tweet for @HillaryClinton and @realDonaldTrump were filtered out from the list and given labels are 0 and 1.

Then all of the URLs, hashtags and messages were all split and lowered and combined with their respective labels. Then a final data frame was created which included labels and messages.

The messages were then converted to tokens and lemmas using TextBlob for further processing. After that we used the CountVectorizer to convert the messages into a sparse matrix of token counts and also used the TF-IDF(Term frequency-inverse document frequency) to apply term frequency inverse document frequency normalization to the sparse matrix we just created.

Method description

We then first created three different pipelines that included 3 different classification models, which included a Decision Tree, Multinomial Naive Bayes, and an ensemble Random Forest.

Using each of the 3 different pipelines we then ran a grid search using the training data to find the optimized hyper parameter setting for the dataset. The parameters we tested on were TF-IDF true or false, and bag of words conversion using lemma or split into tokens. Then this data was refitted and we used core parallelization which was running on all cores to make the processing time faster. The scoring optimization was done using accuracy. And finally a stratifiedKFold was performed using 10 folds.

We then printed out the grid search results as shown in (fig.12,13 and 14). Using the grid search parameter setting we performed prediction on the test data. The results as shown in (fig.9,10,11) were 84%, 87% and 93% respectively. With MNB scoring the highest accuracy. The precision, recall, f1-scores and support values were also calculated and are shown below (fig.9,10,11) for each classifier.

Finally a ROC was plotted to show the accuracy comparison of each of the classifier to better visualize the results. Also a confusion matrix for each classifier was also made to show the fpr, tpr for each classifier as shown in (fig.15)

Evaluation

Looking at the results we can see that out of the 3 classifiers we tested on our data set the Multinomial Naives Bayes outperformed the rest with a test accuracy of 93.61%. While both the ensemble Random forest and decision tree models were only giving us an accuracy of 87.44% and 84.57% respectively. We also printed out the grid scores which gave us precision, recall, F1 scores and support scores as shown in the figure below. (fig.9,10,11)

Inspecting the results, we see the precision scores for our best classifier (MNB) for Hillary Clinton was 0.94 and for Donald Trump was 0.92. Likewise, the recall scores were 0.93 for each

of the candidates. Therefore, based on the result we can conclude that our classifier is predicting relevant data more then 90% of the time.

An ROC plot was plotted to give a visualization of each of the accuracies. Which shows that curve is inclined towards the true positive values. Meaning good accuracy. (fig.16)

Lastly, we put an input line where we can test if our input tweet is how much related to each trump or Hillary. This will predict the users tweet and tell you how much it is accurate to each candidate.

Discussion

This project shows that it is possible to predict who tweeted a certain tweet, based on their past tweets. By analyzing the language of the past tweets, it is possible to predict whether a tweet is written by Hillary Clinton or Donald Trump. These individuals have words that they use frequently, which makes it easier to categorize whether a tweet is from Clinton or Trump. For example, Trump is more likely to use the words “Make America Great Again,” or “Crooked,” then Clinton.

Our results show that the Multinomial Naive Bayes (MNB) is the best at accurately predicting whether a tweet was written by Donald Trump or Hillary Clinton, with a test accuracy of 93.61%. For analyzing language, this classifier is likely going to perform better than other classifiers. Looking at the Confusion Matrix (fig.15) the MNB was able to correctly classify 855 of Hillary’s tweets (out of 923) and correctly classify 742 of Donald’s tweets (out of 794). This outperformed the confusion matrices for the Decision Tree and the Random Forest Classifiers.

As a future application of this work, this predictor could be used to help create a tweet generator. An individual could choose if they want a tweet from Donald Trump, and the application could give a randomized Donald Trump tweet based on his most used unigrams and bigrams.

To improve the model accuracy we could also have used neural networks but due to lack in computing power we were not able to. Also, some of the hyperparameters for the tree models could have been changed to find the most optimized parameter. Which also we could not perform due to lack of computer power.

Conclusion

In conclusion, for the twitter dataset we used the best classification model, which was the Multinomial Naive Bayes with the default hyperparameters. Although the decision tree and random forest did fairly good, it did not perform as well as the MNB. The reason behind this could be due to the other two models being tree models whereas MNB is for multinomial models. So the multinomial classifier models are suitable for classification with discrete features (example, word counts for test classification). Multinomial models normally require feature counts. However, in practice fractional counts such as TF-IDF also work which we have used in our model.

Graphs

Fig.1

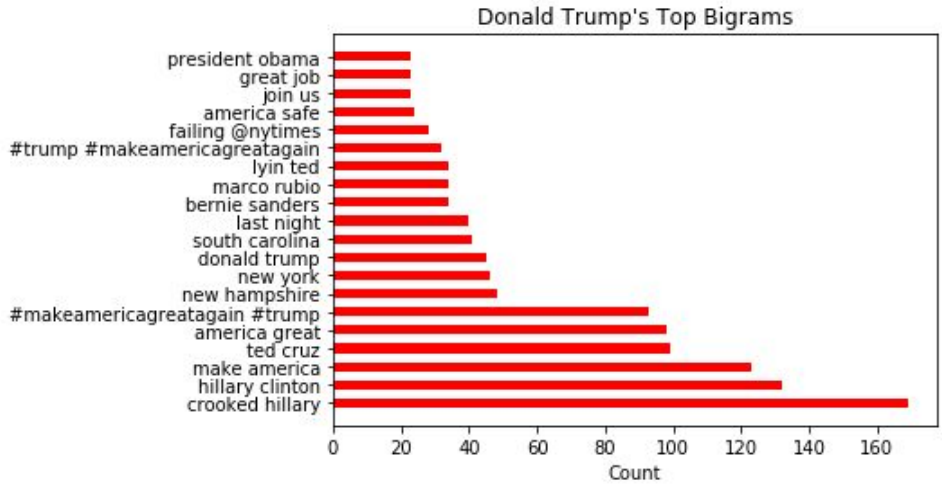


Fig.2

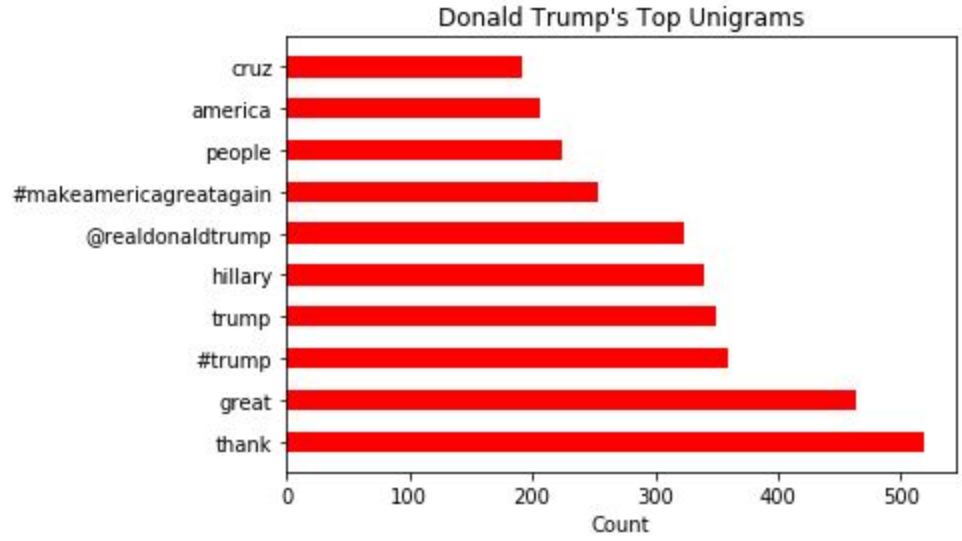


Fig.3

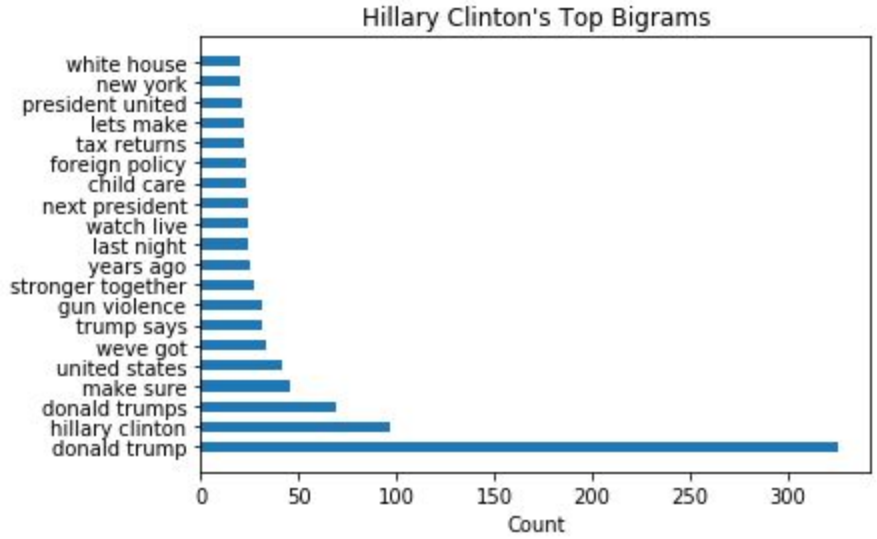

Hillary Clinton's Top Bigrams
white house
new york
president united
lets make
tax returns
foreign policy
child care
next president
watch live
last night
years ago
stronger together
gun violence
trump says
weve got
united states
make sure
donald trumps
hillary clinton
donald trump
0
50
100
150
200
250
300
Count

Fig.4

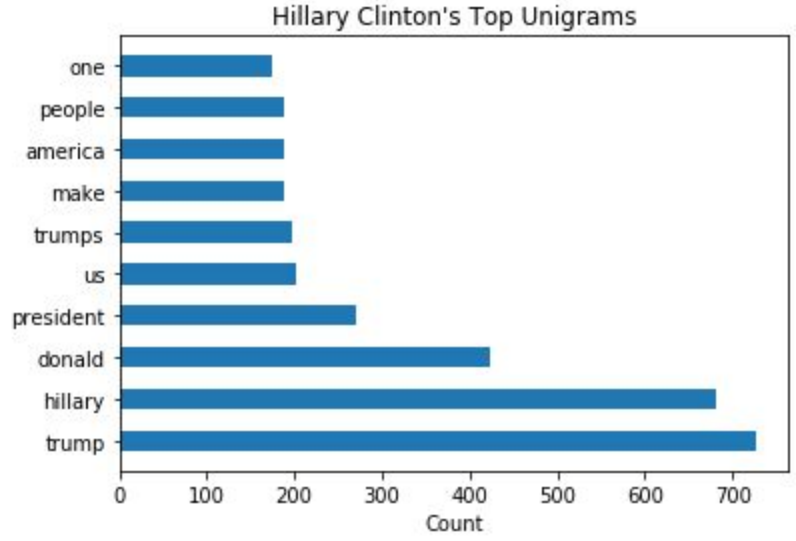

Hillary Clinton's Top Unigrams
one
people
america
make
trumps
us
president
donald
hillary
trump
0
100
200
300
400
500
600
700
Count

Fig.5

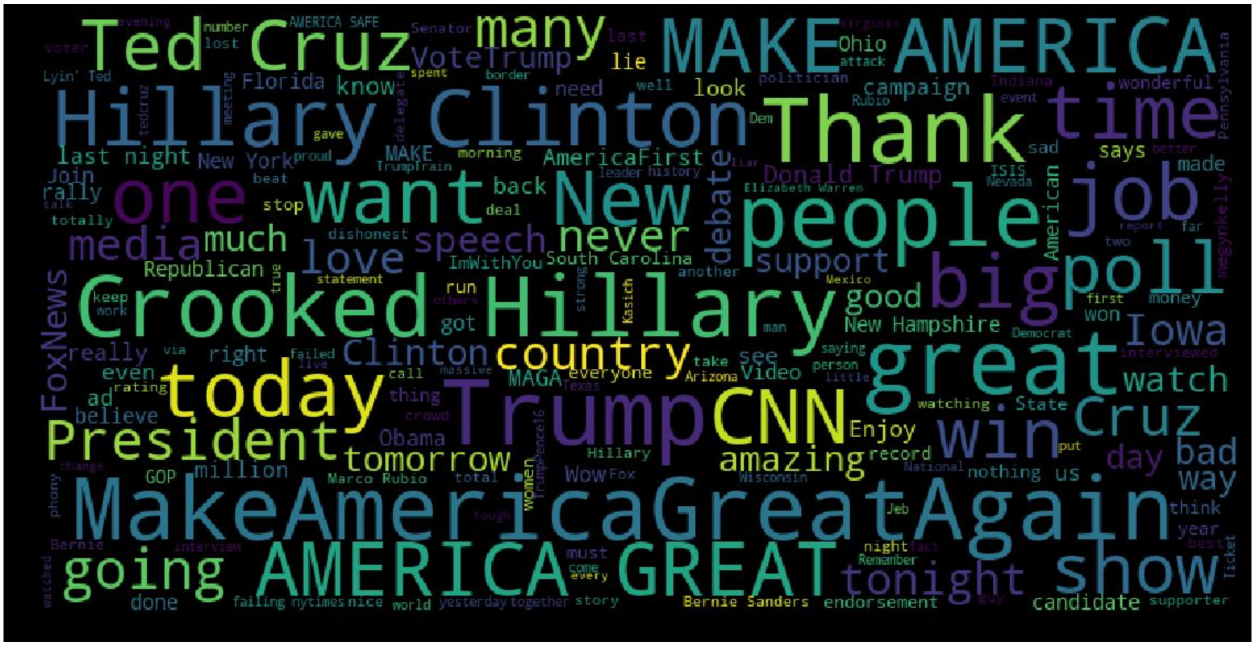

Ted Cruz
many
MAKE
AMERICA
VoteTrump
Hillary Clinton
Thank
time
one
want
New
people
job
media
much
love
speech
never
debate
support
big
poll
Crooked Hillary
FoxNews
today
Trump
CNN
great
country
Clinton
Iowa
watch
President
win
Cruz
tomorrow
amazing
bad
MakeAmericaGreatAgain
going
AMERICA GREAT
tonight
show
Florida
know
last night
New York
AmericaFirst
campaign
Donald Trump
American
Republican
South Carolina
ImWithYou
good
New Hampshire
really
right
even
MAGA
everyone
Video
Enjoy
Obama
million
GOP
Marco Rubio
Bernie Sanders
endorsement
candidate
done
day
way
Ohio
look
back
rally
Join
says
made
stop
deal
got
run
see
believe
nothing
us

Fig.6

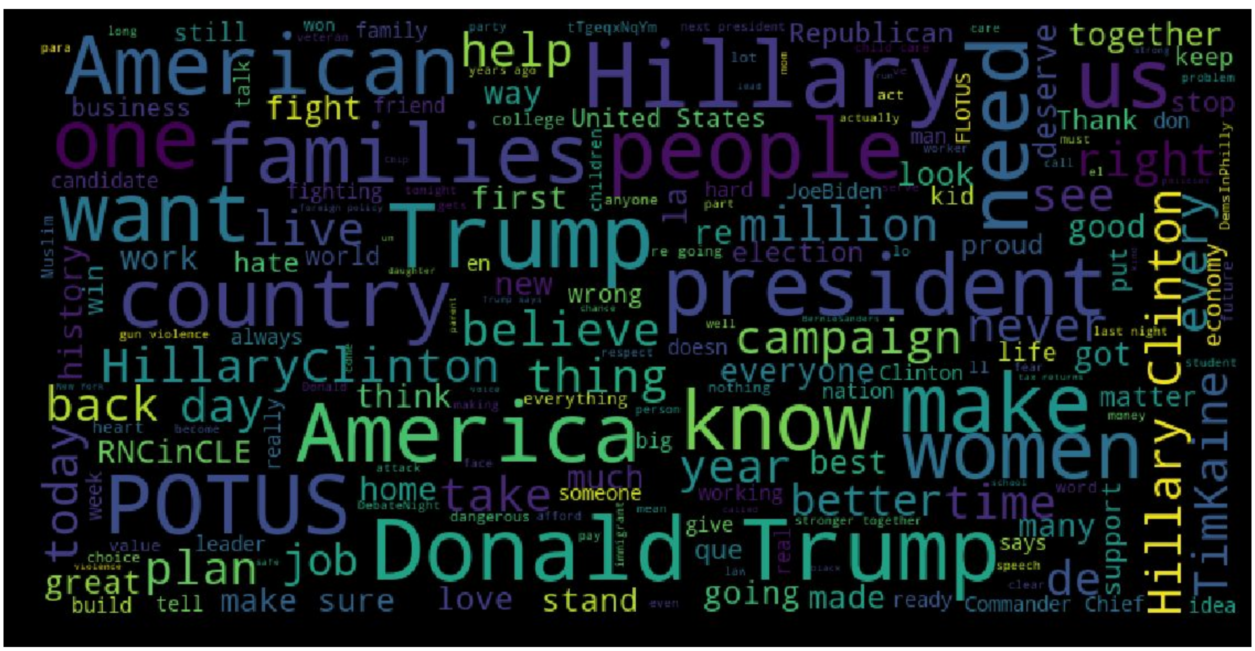

American
help
Hillary
us
need
together
one
families
people
Republican
want
live
Trump
first
million
country
president
election
HillaryClinton
believe
campaign
never
know
make
America
women
POTUS
Donald Trump
HillaryClinton
TimKaine
back
day
thing
everyone
think
today
RNCinCLE
year
best
better
time
many
plan
job
home
take
much
great
make sure
love
stand
going
made
ready
Commander Chief
idea
fight
still
family
business
way
United States
Thank
right
see
look
good
proud
history
win
work
hate
world
new
wrong
life
got
matter
support
economy
JoeBiden
FLOTUS
deserve
keep
stop
kid
put
every
plan
leader
build
tell
give
que
says

Fig.7 Retweet Cloud Donald

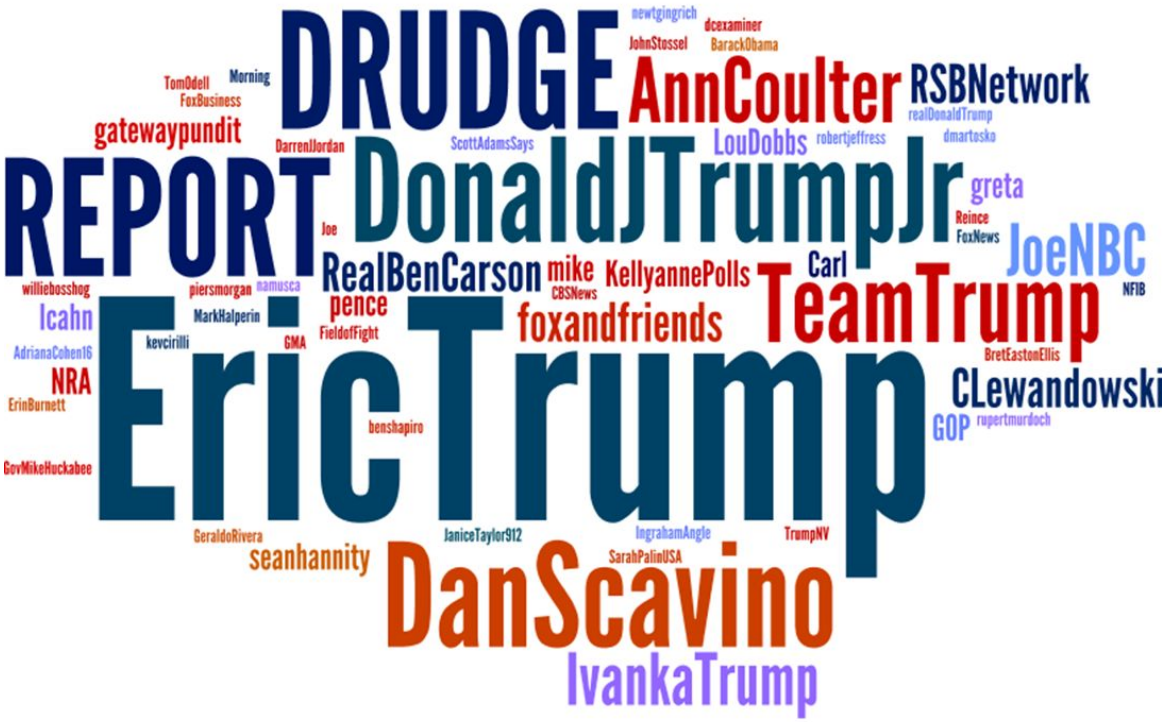


Fig.8 Retweet Cloud Hillary

Fig.9

	precision	recall	f1-score	support
0	0.94	0.93	0.93	923
1	0.92	0.93	0.93	794
avg / total	0.93	0.93	0.93	1717

Accuracy on test set: 93.61%

Fig.10

	precision	recall	f1-score	support
0	0.85	0.85	0.85	923
1	0.82	0.82	0.82	794
avg / total	0.84	0.84	0.84	1717

Accuracy on test set: 84.57%

Fig.11

	precision	recall	f1-score	support
0	0.88	0.89	0.89	923
1	0.88	0.86	0.87	794
avg / total	0.88	0.88	0.88	1717

Accuracy on test set: 87.44%

Fig.12

```
[mean: 0.84569, std: 0.01872, params: {'bow__analyzer': <function split_into_lemmas at 0x1a1c139f28>, 'tfidf__use_idf':
True}, mean: 0.83346, std: 0.02567, params: {'bow__analyzer': <function split_into_lemmas at 0x1a1c139f28>,
'tfidf__use_idf': False}, mean: 0.83296, std: 0.02051, params: {'bow__analyzer': <function split_into_tokens at
0x1a1c139e18>, 'tfidf__use_idf': True}, mean: 0.83571, std: 0.01506, params: {'bow__analyzer': <function
split_into_tokens at 0x1a1c139e18>, 'tfidf__use_idf': False}]
```

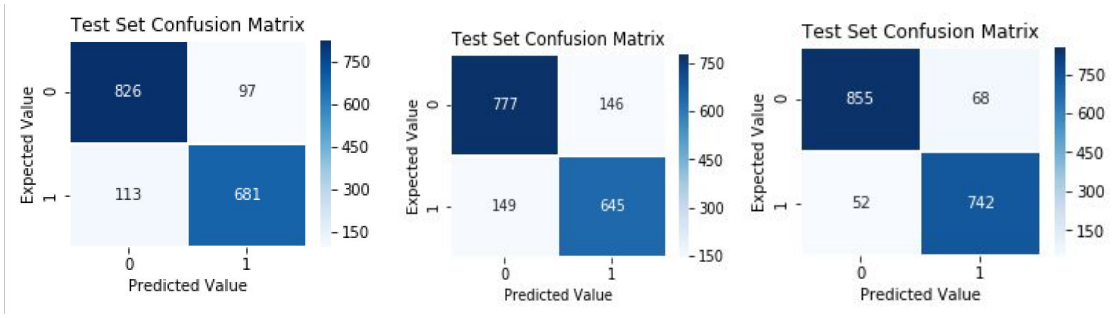
Fig.13

```
[mean: 0.93608, std: 0.01612, params: {'bow__analyzer': <function split_into_lemmas at 0x1a1c139f28>, 'tfidf__use_idf':
True}, mean: 0.92684, std: 0.01353, params: {'bow__analyzer': <function split_into_lemmas at 0x1a1c139f28>,
'tfidf__use_idf': False}, mean: 0.93358, std: 0.01554, params: {'bow__analyzer': <function split_into_tokens at
0x1a1c139e18>, 'tfidf__use_idf': True}, mean: 0.92684, std: 0.01402, params: {'bow__analyzer': <function
split_into_tokens at 0x1a1c139e18>, 'tfidf__use_idf': False}]
```

Fig.14

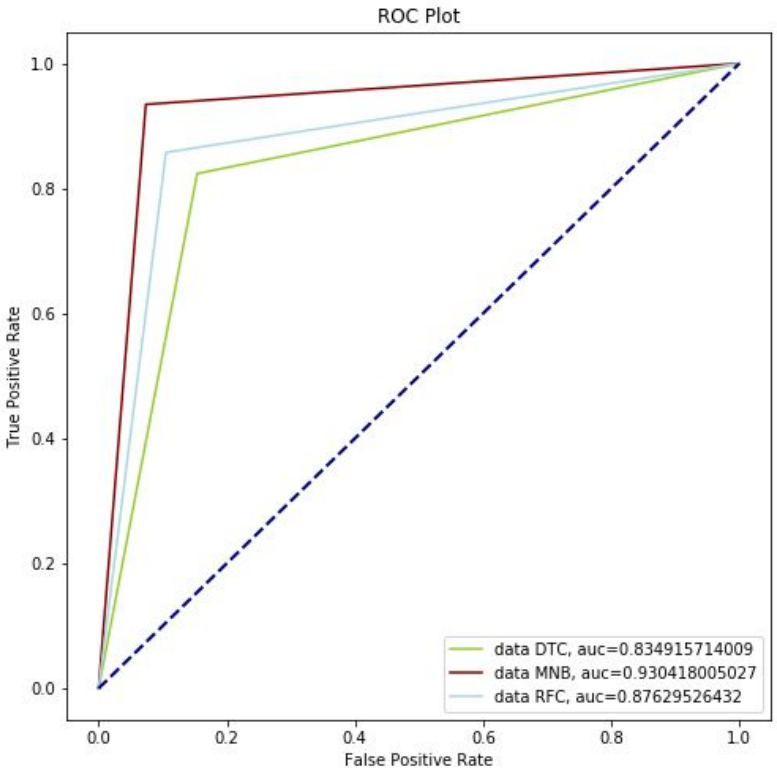
```
[mean: 0.87441, std: 0.01308, params: {'bow__analyzer': <function split_into_lemmas at 0x1a1c139f28>, 'tfidf__use_idf':
True}, mean: 0.87191, std: 0.02313, params: {'bow__analyzer': <function split_into_lemmas at 0x1a1c139f28>,
'tfidf__use_idf': False}, mean: 0.86542, std: 0.01989, params: {'bow__analyzer': <function split_into_tokens at
0x1a1c139e18>, 'tfidf__use_idf': True}, mean: 0.87266, std: 0.01919, params: {'bow__analyzer': <function
split_into_tokens at 0x1a1c139e18>, 'tfidf__use_idf': False}]
```

Fig.15



(RF) (DTC) (MNB)

Fig.16



Appendix

Team Member	Contribution
Sameep Shah	
Ryan Keaveny	
Albert Sugianto	
Faaleh Sayeed	