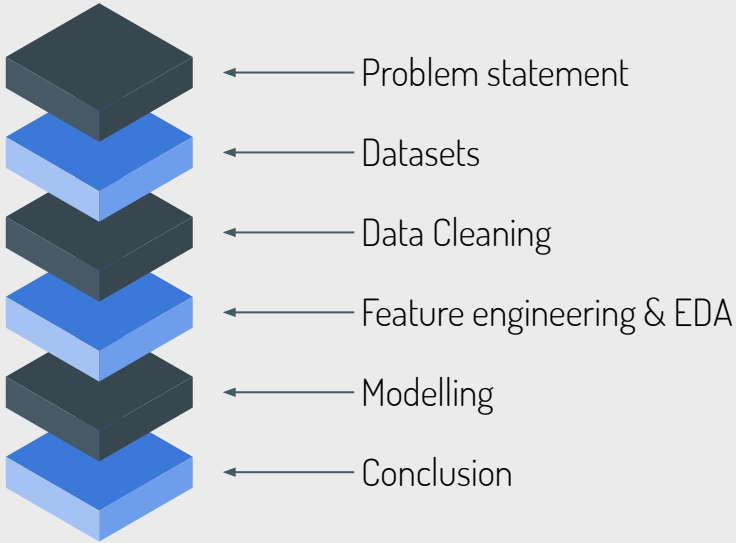
TRUMP, TWEETS AND THE STOCK MARKET

DSI Capstone Project
By: Russell Quah

TABLE OF CONTENTS



Problem statement

PROBLEM STATEMENT

The US market team at a local bank has seen literature on models that are able to predict market movement based on tweets by Donald Trump:
- JP Morgan creating a 'Volfefe index to track tweets vs bond market
- Bank of America has stated that on the days when President Trump tweets a lot, the stock market falls

They have tasked the data science team to build a classification model using Natural Language Processing to predict if Donald Trump's tweets are market moving.

PROBLEM STATEMENT

- Logistic Regression
- XGBoost
- Long Short Term Memory Neural Network
- Evaluate the models based on:
  - accuracy (% predictions the model gets correct, both a significant movement and a non-significant movement)
  - precision (% predicted significant movement when it is actually significant movement)
  - sensitivity (% predicted significant movement out of all correct predictions)
- choose the best performing model to test it on the holdout csv

Datasets

DATA SETS

@REALDONALDTRUMP TWEETS

Kaggle dataset

- 04 May '09 to 17 June '20
- 43352 tweets
- 8 columns
- Removed unnecessary features (id, link, mentions and hashtags)

S&P500 RETURNS

Yahoo finance API

- May '09 to June '20
- 2805 days
- 8 columns

TRAIN, VALIDATION, HOLDOUT

27,676 tweets | 6,919 tweets | 8,649 tweets

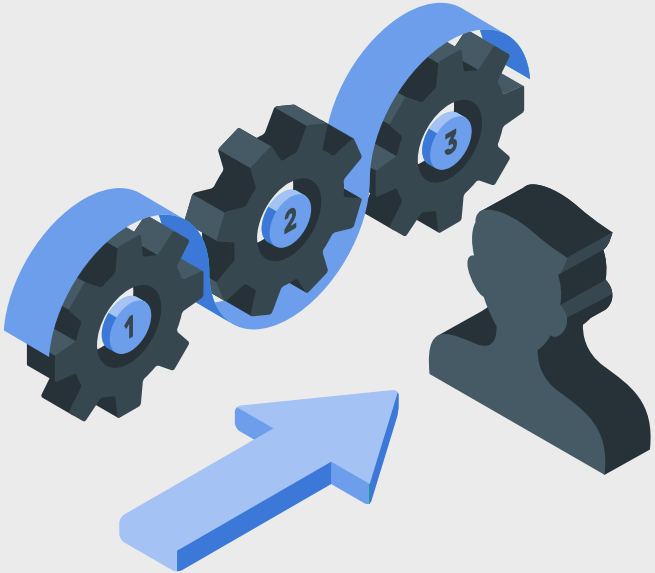
Data Cleaning

DATA CLEANING

- BeautifulSoup
- Regex
- Stop words (nltk + spacy + etc.)
- Lemmatize

Feature engineering & EDA

FEATURE ENGINEERING ON TWEETS



CYCLICAL DATA

Date was split into cyclical features:

- Month, Day, Hour, Minute
- Sin and Cosine

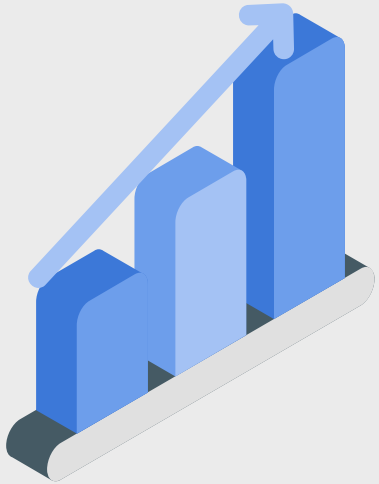
VADER SENTIMENT ANALYSIS (ON RAW TEXT DATA)

- Specifically designed to handle social media
- Compound = Positive + Neutral - Negative

LOUGHRAN MCDONALD FINANCIAL SENTIMENT ANALYSIS (PROCESSED TEXT DATA)

- Dictionary
- Counts the number of times words appear in 9 different categories

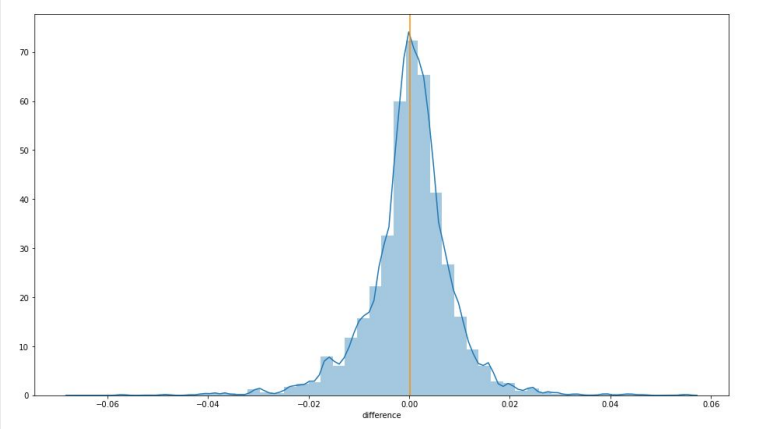
FEATURE ENGINEERING ON S&P500

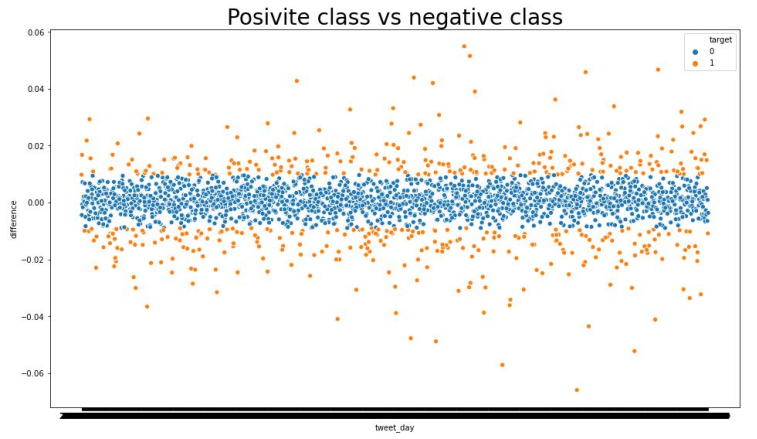


INTRADAY DIFFERENCE

Target variable: Opening price - closing price

DEFINING THE POSITIVE CLASS





Positive Class

- 1 standard deviation away from the mean

TOTAL NUMBER OF TWEETS SINCE MAY'09

34,595

19.74 MEAN WORD COUNT

TOTAL NUMBER OF TWEETS AS POTUS:

9,644

7.75 AVERAGE* TWEETS A DAY AS POTUS

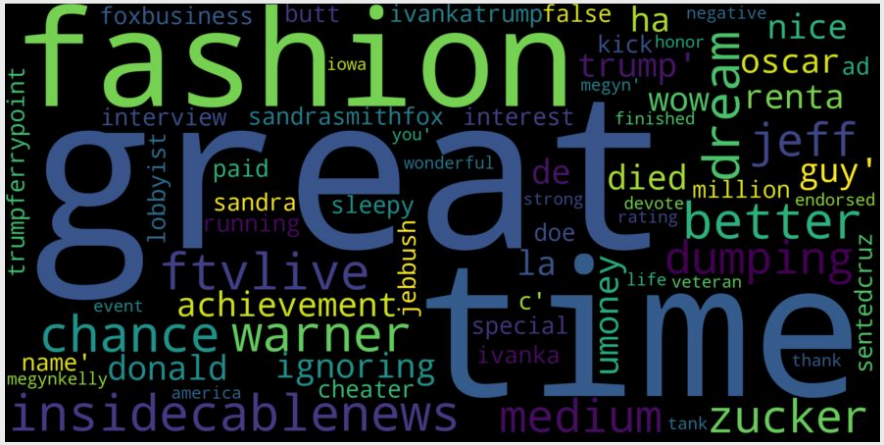
+ CLASS **9.00** AVERAGE* TWEETS A DAY AS POTUS

- CLASS **7.77** AVERAGE* TWEETS A DAY AS POTUS

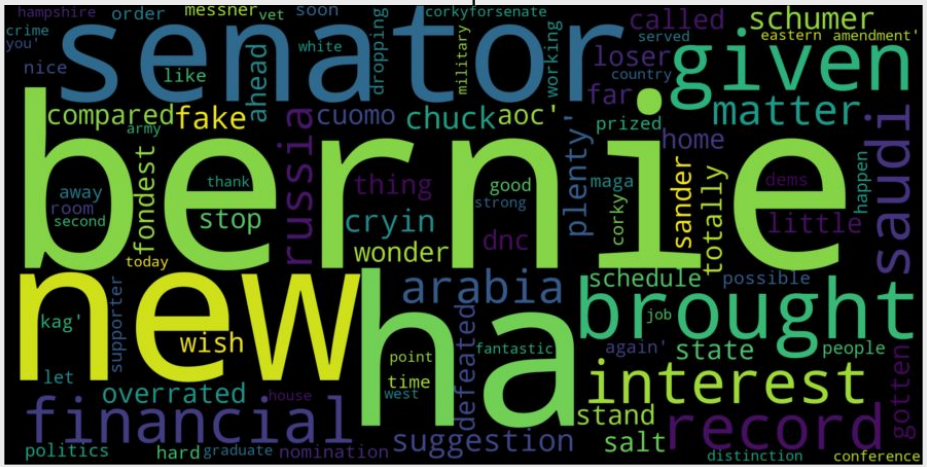
*Excludes days where he does not tweet at all

WORDCLOUD

Positive Class

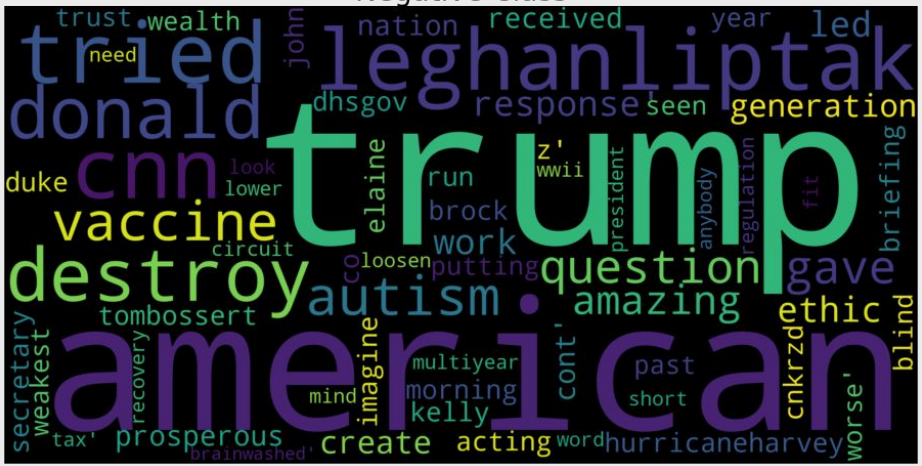


POTUS Status positive class

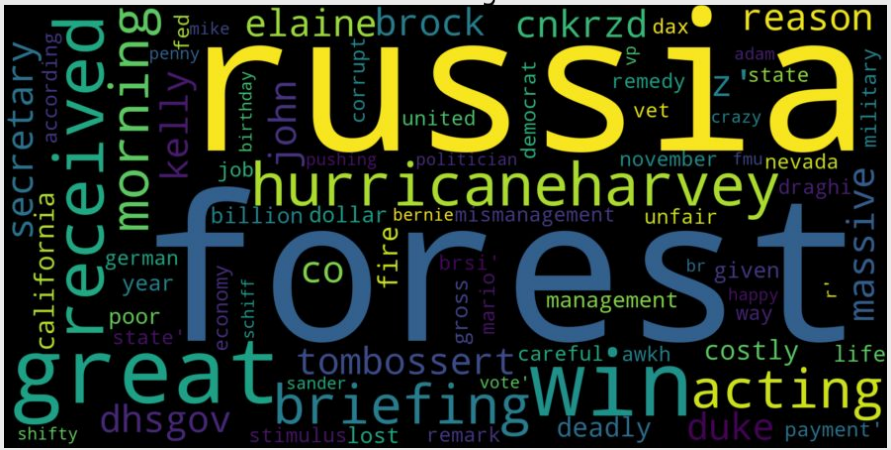


WORDCLOUD

Negative Class

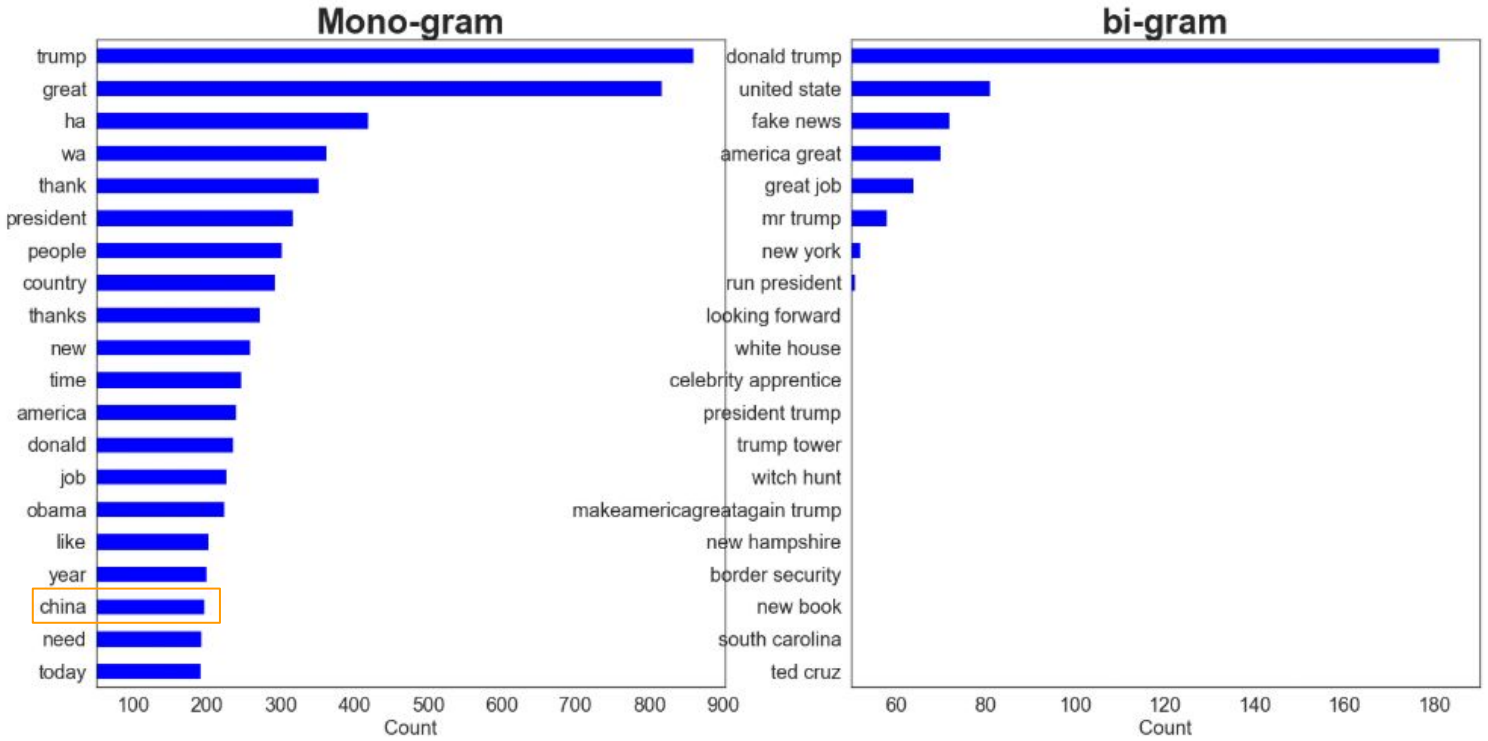


POTUS Status negative class



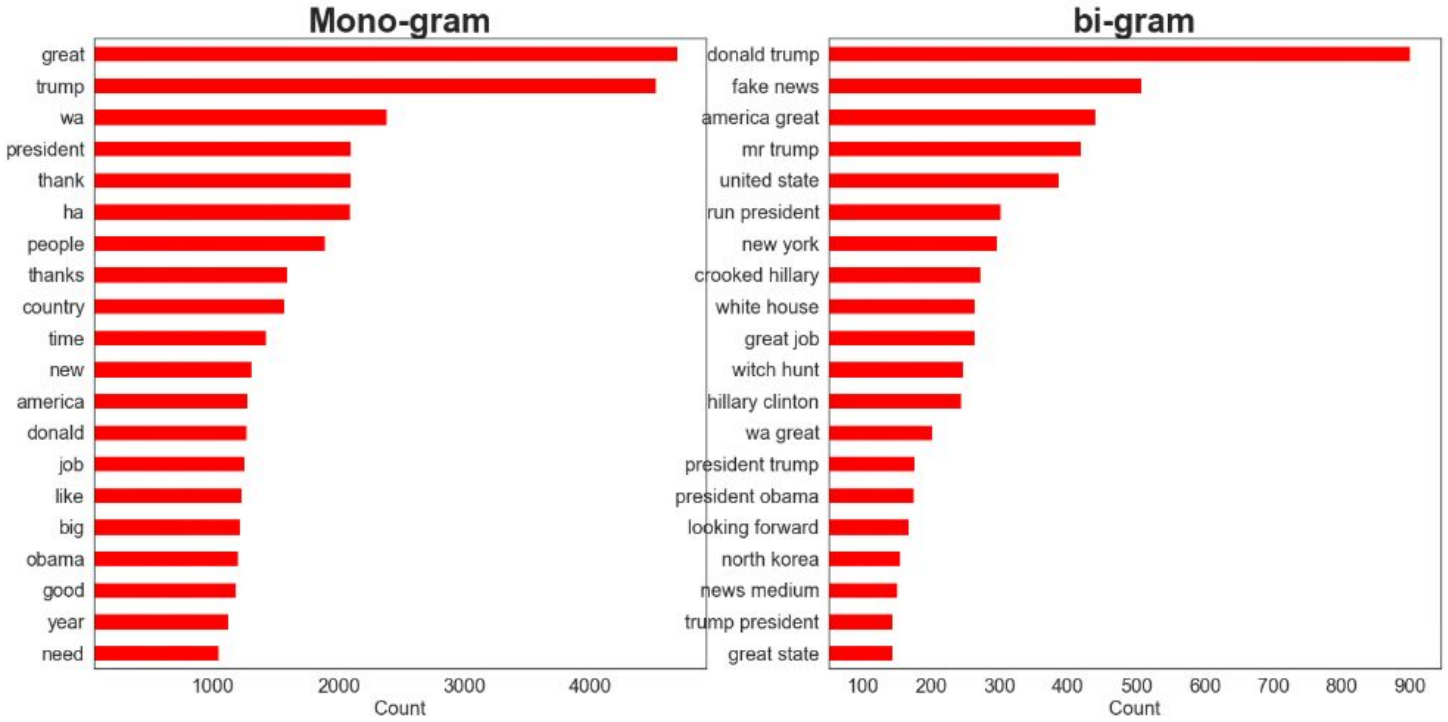
POSITIVE CLASS MOST USED WORDS

```
In [107]: #positive class
          barplot_cvec(1, 'Positive Class', ['Mono-gram', 'bi-gram'], 'blue', 50)
```



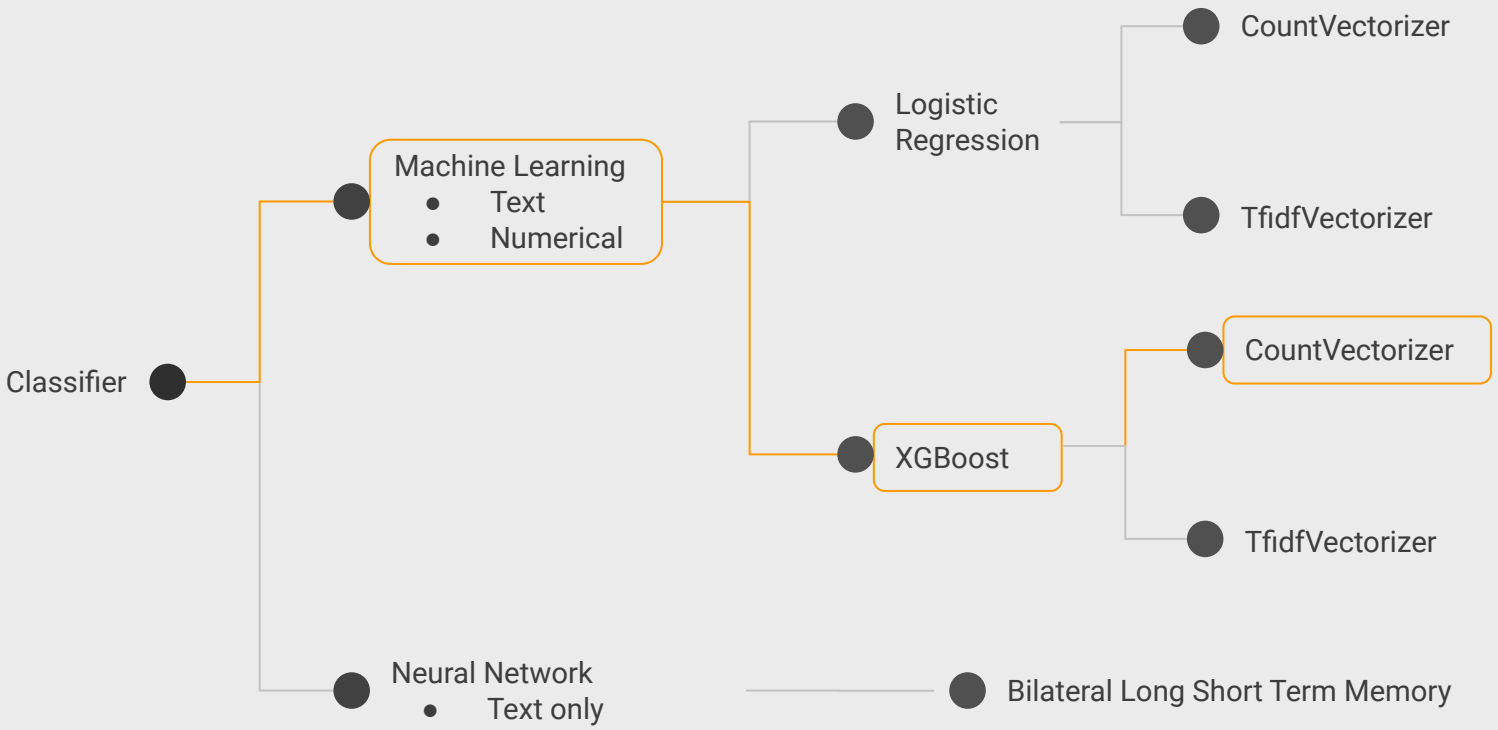
NEGATIVE CLASS MOST USED WORDS

In [108]:
```
#negative class
barplot_cvec(0, 'Negative class', ['Mono-gram', 'bi-gram'], 'red', 50)
```



Modelling

MODELLING


Classifier
Machine Learning
Text
Numerical
Logistic Regression
CountVectorizer
TfidfVectorizer
XGBoost
CountVectorizer
TfidfVectorizer
Neural Network
Text only
Bilateral Long Short Term Memory

MODELLING

Pipeline 1 - Gridsearch

Countvectorizer → XGBoost

+

Pipeline 2 - Gridsearch

Smote → Standard Scaler → XGBoost

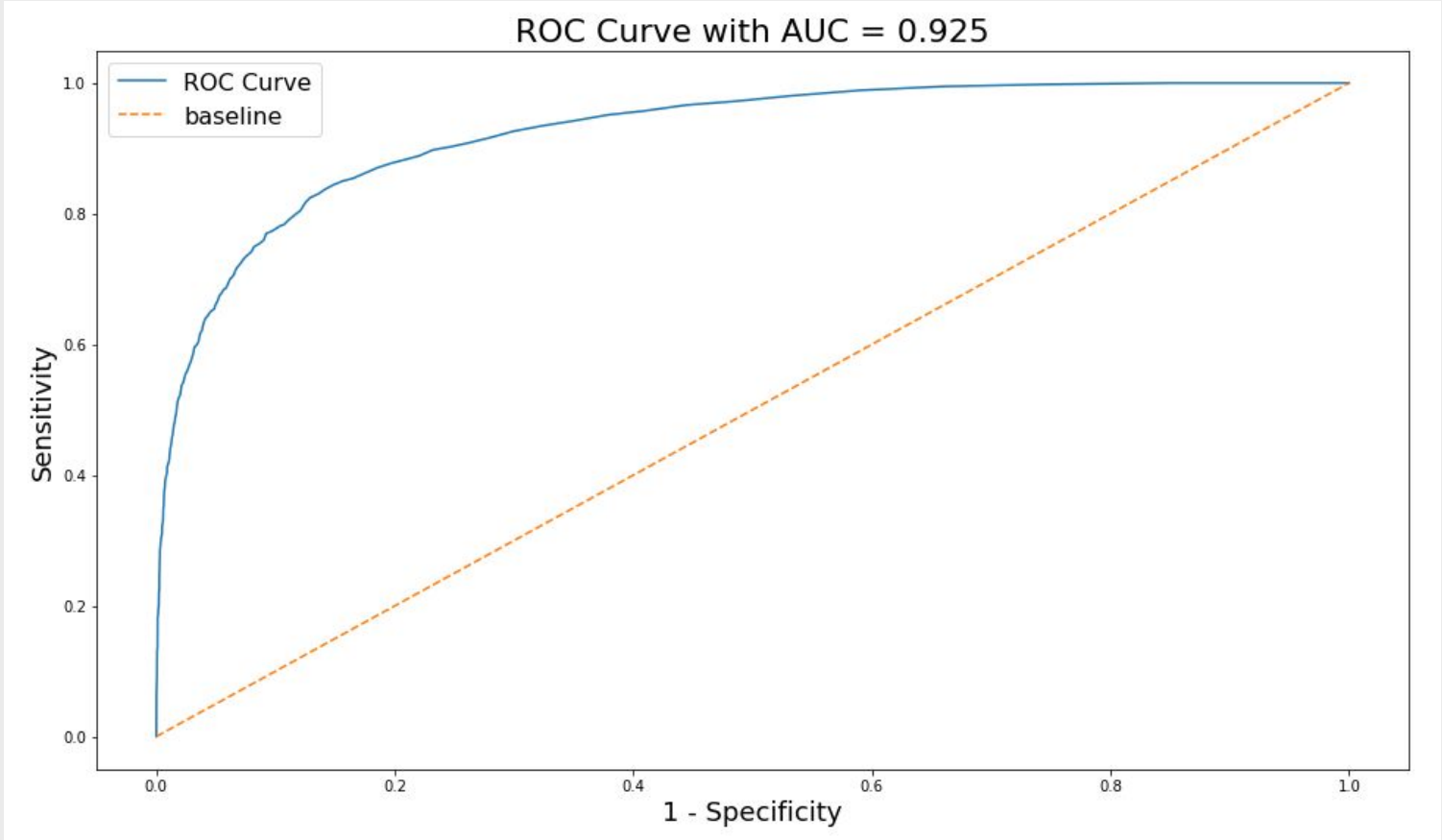
MODEL PERFORMANCE

	gs1 (cvec->Ls)	gs2 (tfidf->Ls)	gs3 (cvec->XGB)	gs4 (tfidf->XGB)	gs6 (Bilateral LSTM)	best_model (cvecc->xgboost)
accuracy(train)	0.661	0.600	0.999	0.992	0.877	0.998
accuracy(val)	0.665	0.591	0.905	0.895	0.806	0.906
precision	0.331	0.455	0.462	0.401	0.087	0.469
sensitivity	0.180	0.177	0.854	0.830	0.200	0.857
F1	0.233	0.255	0.600	0.541	0.122	0.606
roc_auc	0.547	0.552	0.925	0.908	0.512	0.925

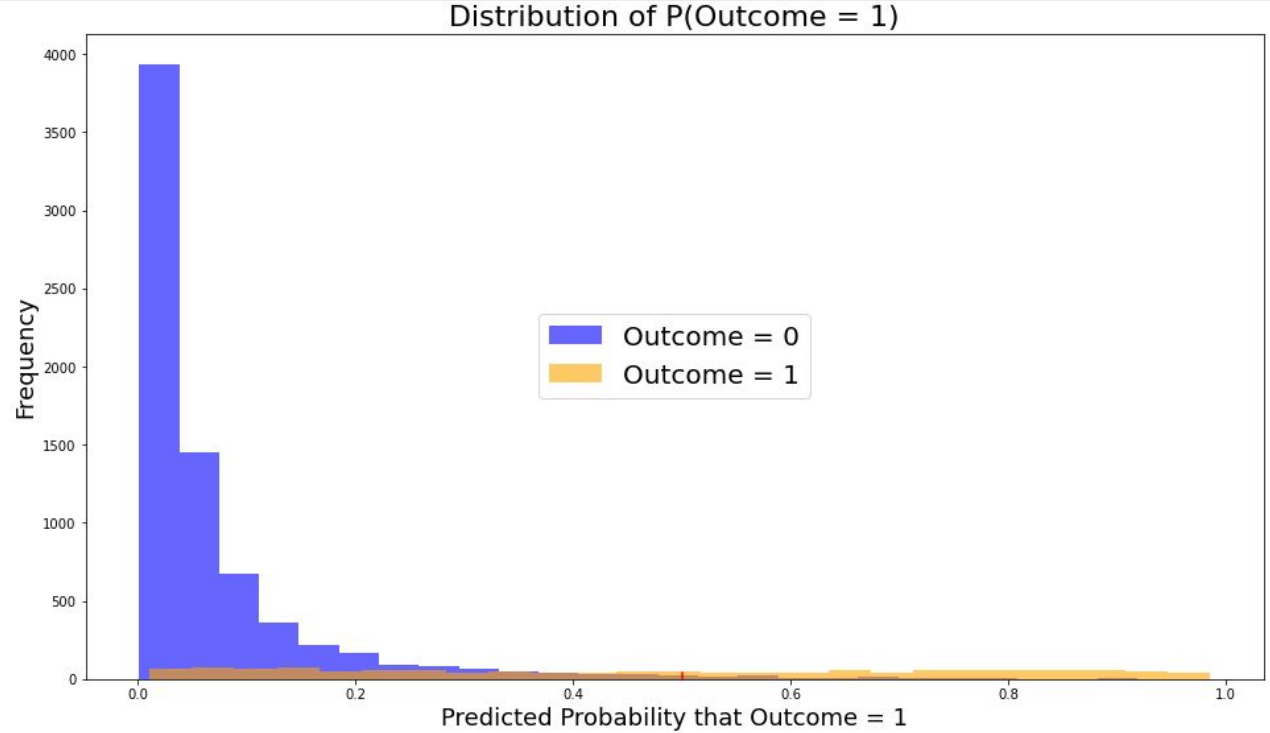
Baseline:

- 0.846 of negative class

ROC AUC

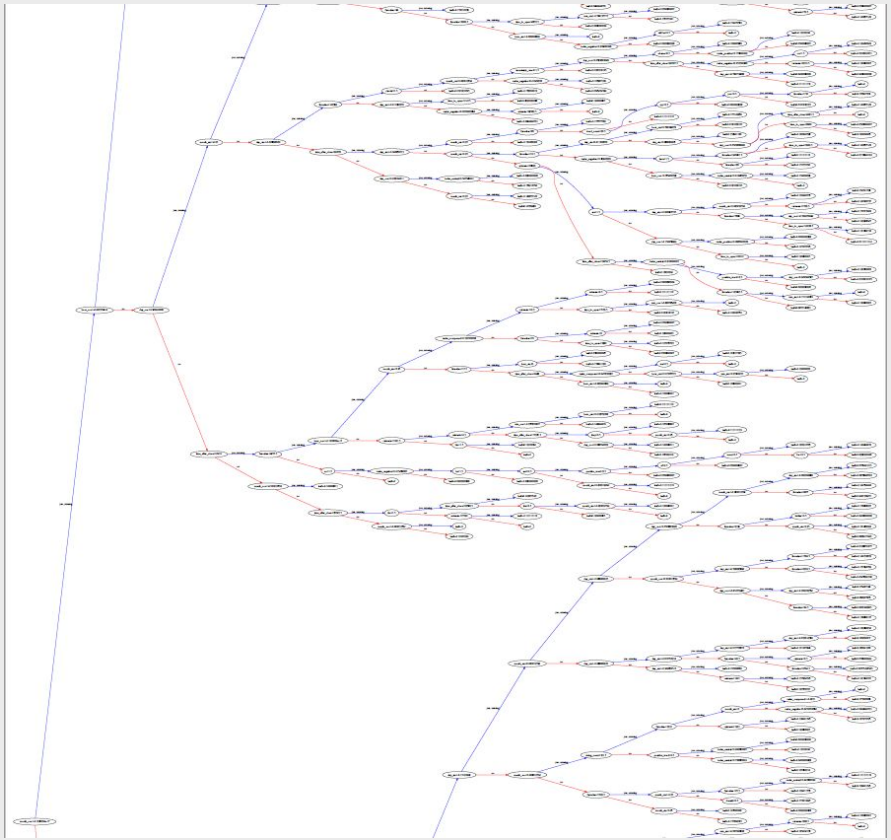

ROC Curve with AUC = 0.925
ROC Curve
baseline
Sensitivity
1 - Specificity
0.0
0.2
0.4
0.6
0.8
1.0

ROC AUC



The model is able to accuractely predict with a high degree of certainty the negative class outcome, however, it is unable to perform as well on the positive class.

MAP OF XGBOOST TREE



Conclusion

CONCLUSION

Best model is CVEC into XGBoost,

- **Accuracy of 0.906**
- **Precision of 0.469**
- **Sensitivity of 0.857**

Recommendation for the US market team:

- On days of positive class, Trump tends to tweet more on average
- The model is able to accurately predict when his tweets **will not move** the market

Limitations

- Model falls short on precision
- It is not a parametric model
- More work on neural network

THANKS

Does anyone have any questions?