Prediction on Dow Jones Stock Prices

Donggu Kim
Lisa Kim
Jing Kong
Xiqing Liu
Tianran Lu

Goal:

To find the best model of prediction on Dow Jones stock prices

Data (up to now):

-News data: historical news headlines from Reddit World News Channel
(Range from 2008-06-08 to 2016-07-01)

-Stock data: Dow Jones Industrial Average (DJIA)
(Range from 2008-08-08 to 2016-07-01)

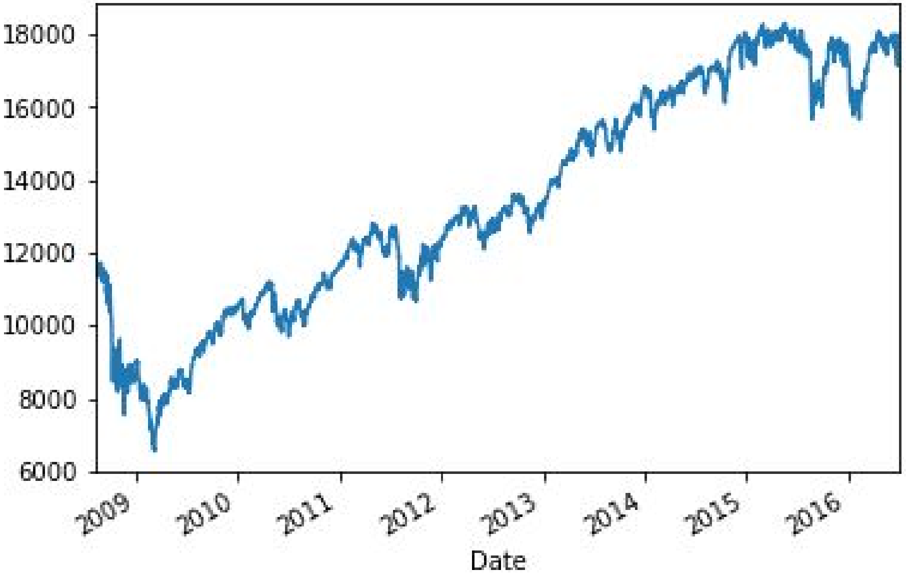
Methodology (up to now):

-Time Series Analysis

-Classification Model

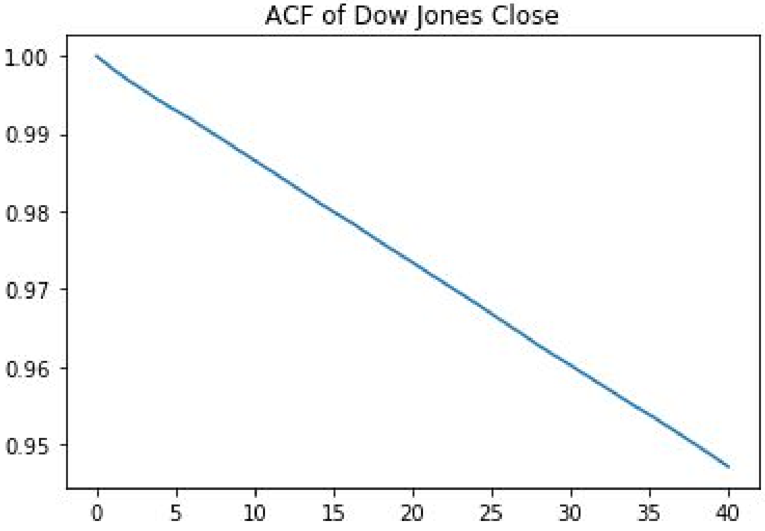
Time Series Analysis on Stock Price

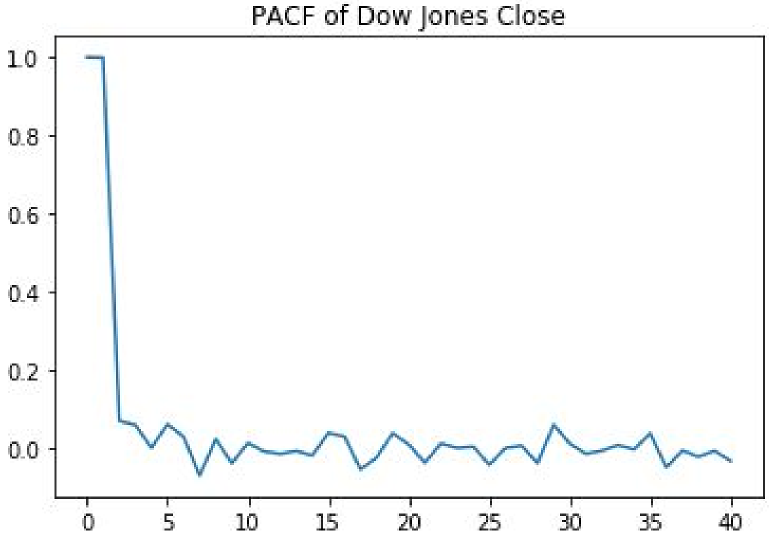
Dow Jones Close Stock Price



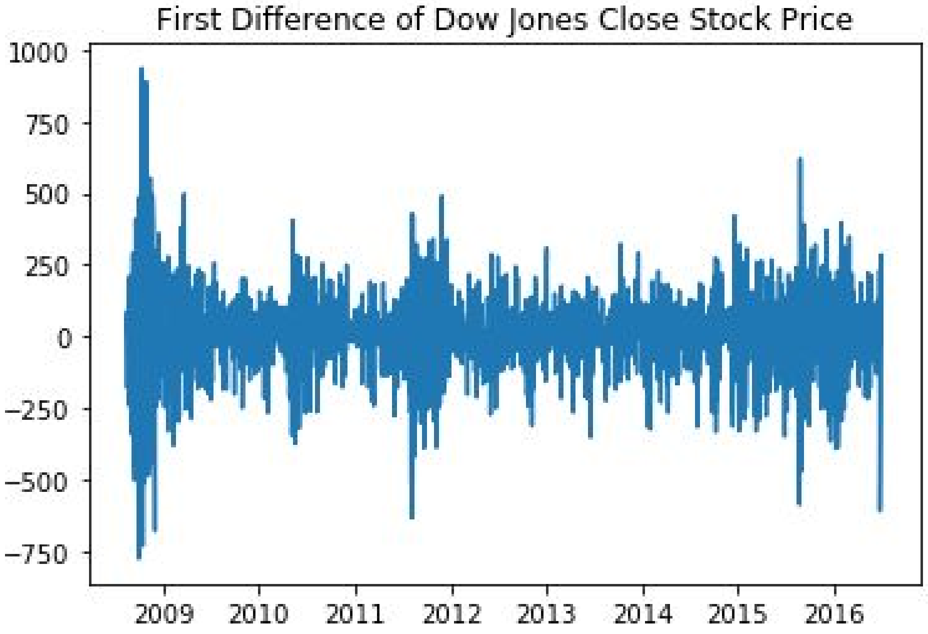
- Upward trend
- Variance is constant over time
- Not stationarity due to upward trend

ACF and PACF of Dow Jones Close Stock Price





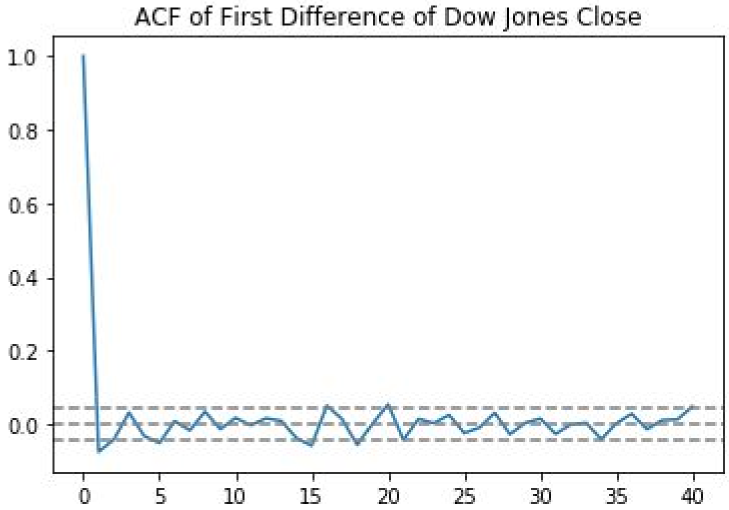
ACF is gradually decreasing and PACF goes to 0 at p = 1

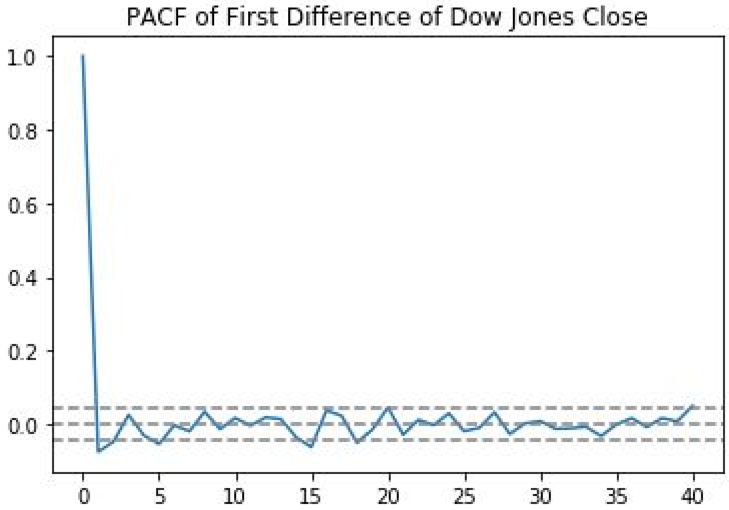


- Upward trend is removed
- Variance looks pretty constant over time

- Approximately stationarity

ACF and PACF of First Difference

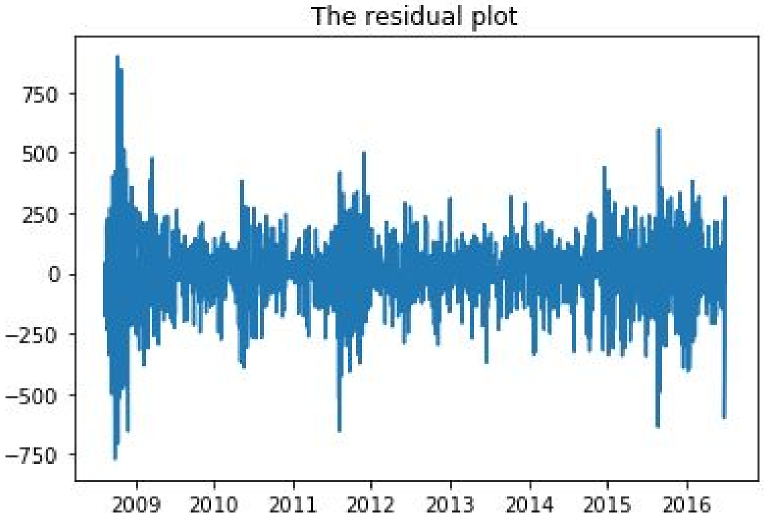


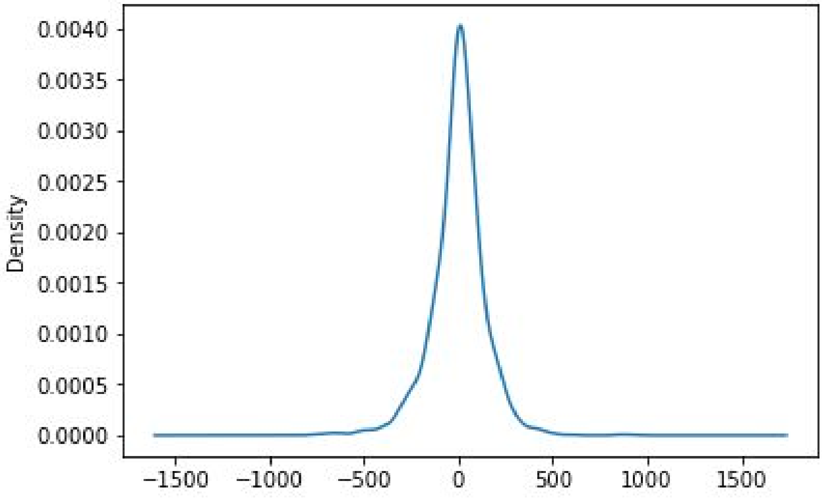


ACF goes close to 0 around q = 2 and p = 2.
Our possible ARIMA models are:
ARIMA(1,1,1), ARIMA(1,1,2), ARIMA(2,1,1), ARIMA(2,1,2)

	AIC	BIC	MSE
ARIMA(1,1,1)	25327.3	25349.7	20412
ARIMA(2,1,1)	25327.3	25355.2	20390.7
ARIMA(2,1,2)	25327.7	25361.3	20374.7

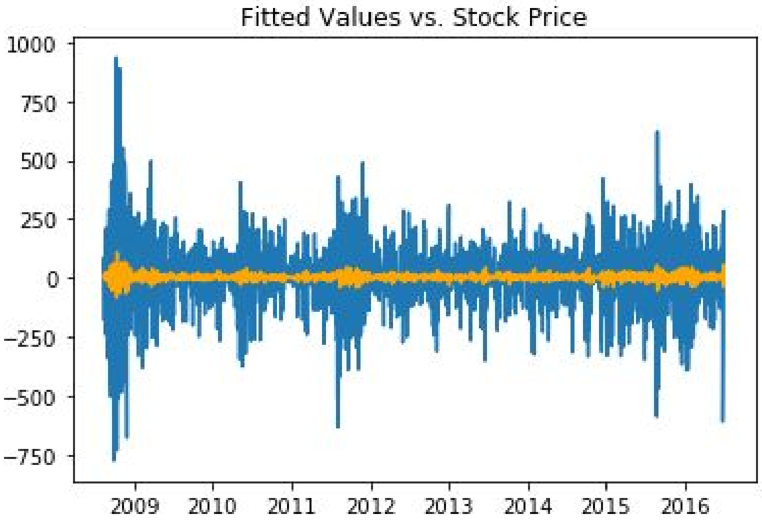
AIC, BIC and MSE are chosen as criteria for selecting the best model for forecasting.
-The higher AIC, the better
-The lower BIC, the better
- The lower MSE, the better
Our primary goal is to predict stock index with the lowest deviation from the actual value.
AIC and BIC are similar but MSE of ARIMA(2,1,2) is the lowest compared to others.
ARIMA(2,1,2) is chosen as the best model for predicting Dow Jones Close Stock Price.

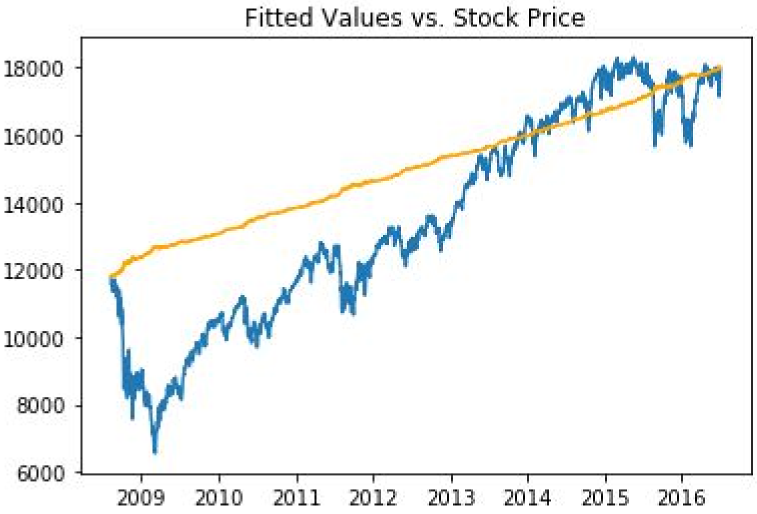




- The residuals are centered around at 0; no apparent pattern
- Mean is constant, variance looks pretty stable

Fitted Values and Stock Price





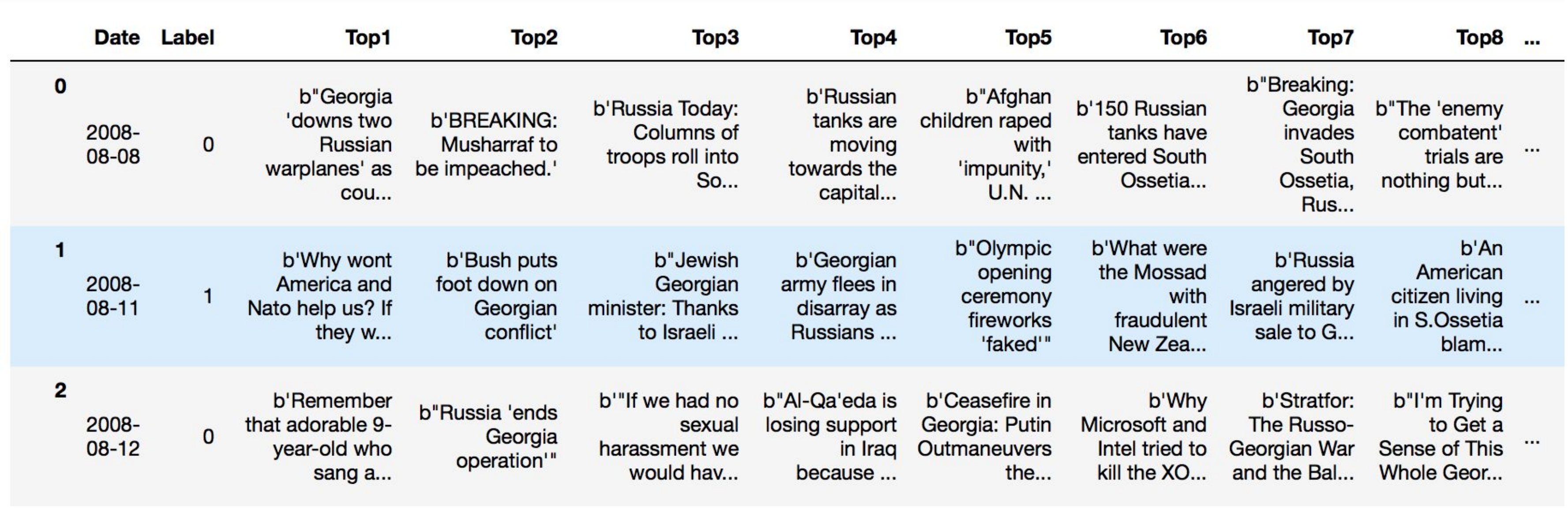
First graph describes the fitted values and actual values of first difference of DJ index. The second one describes fitted values and original close stock price. Like shown in graph, time series analysis doesn’t do a good job at predicting stock price, showing that other models will be needed in order to improve the accuracy level in prediction of DJ index.

Text Analysis on Stock Price

-Basic exploratory data analysis

-Positive and negative words ratio of daily news headlines

DJIA News Combined Dataset



	Date	Label	Top1	Top2	Top3	Top4	Top5	Top6	Top7	Top8	...
0	2008-08-08	0	b"Georgia 'downs two Russian warplanes' as cou...	b'BREAKING: Musharraf to be impeached.'	b'Russia Today: Columns of troops roll into So...	b'Russian tanks are moving towards the capital...	b"Afghan children raped with 'impunity,' U.N. ...	b'150 Russian tanks have entered South Ossetia...	b"Breaking: Georgia invades South Ossetia, Rus...	b"The 'enemy combatant' trials are nothing but...	...
1	2008-08-11	1	b'Why wont America and Nato help us? If they w...	b'Bush puts foot down on Georgian conflict'	b"Jewish Georgian minister: Thanks to Israeli ...	b'Georgian army flees in disarray as Russians ...	b"Olympic opening ceremony fireworks 'faked'"	b'What were the Mossad with fraudulent New Zea...	b'Russia angered by Israeli military sale to G...	b'An American citizen living in S.Ossetia blam...	...
2	2008-08-12	0	b'Remember that adorable 9-year-old who sang a...	b"Russia 'ends Georgia operation'"	b'"If we had no sexual harassment we would hav...	b"Al-Qa'eda is losing support in Iraq because ...	b'Ceasefire in Georgia: Putin Outmaneuvers the...	b'Why Microsoft and Intel tried to kill the XO...	b'Stratfor: The Russo-Georgian War and the Bal...	b"I'm Trying to Get a Sense of This Whole Geor...	...

1: DJIA Close value rose or stayed as the same

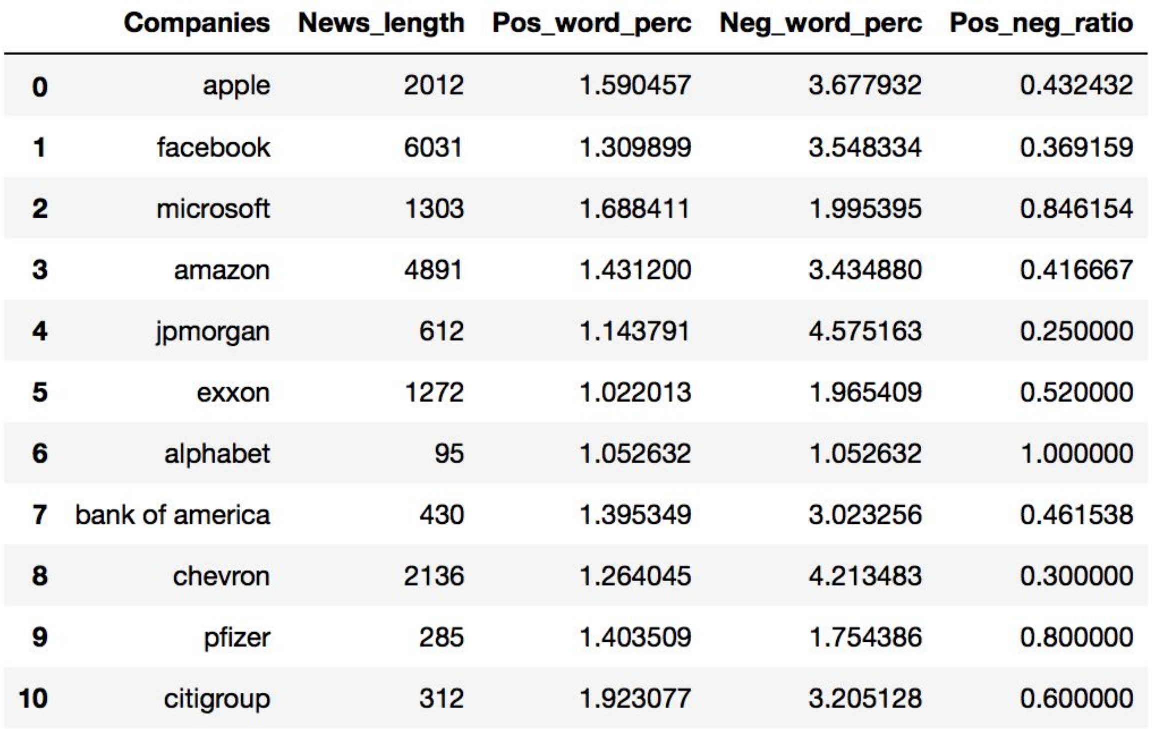
0: DJIA Close value decreased

Positive-Negative Words Ratio of Companies

Inspiration:

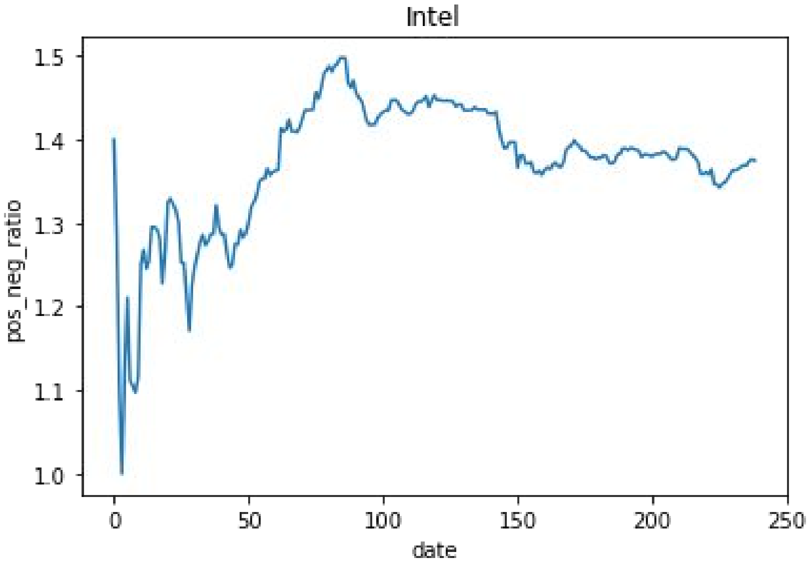
Positive negative word ratio in overall company-related news headlines

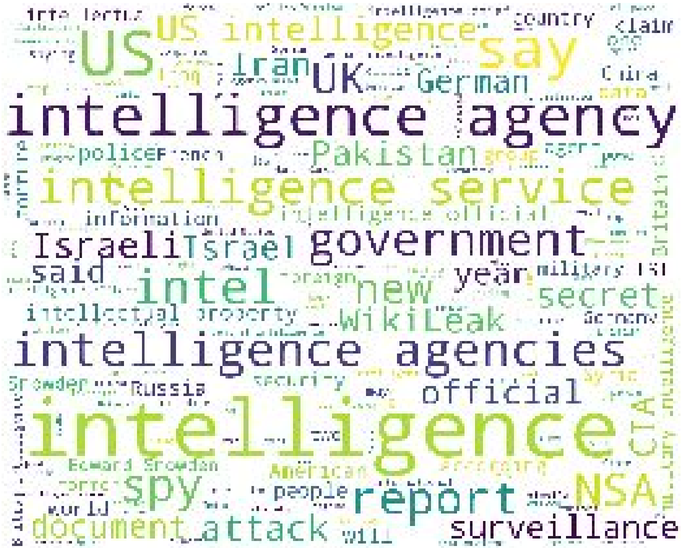
High ratio may indicate an increase in the overall stock price



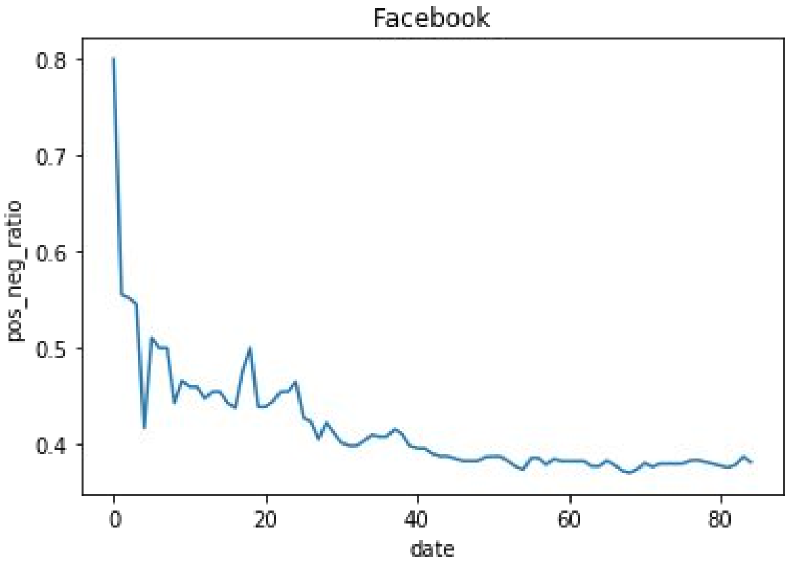
	Companies	News_length	Pos_word_perc	Neg_word_perc	Pos_neg_ratio
0	apple	2012	1.590457	3.677932	0.432432
1	facebook	6031	1.309899	3.548334	0.369159
2	microsoft	1303	1.688411	1.995395	0.846154
3	amazon	4891	1.431200	3.434880	0.416667
4	jpmorgan	612	1.143791	4.575163	0.250000
5	exxon	1272	1.022013	1.965409	0.520000
6	alphabet	95	1.052632	1.052632	1.000000
7	bank of america	430	1.395349	3.023256	0.461538
8	chevron	2136	1.264045	4.213483	0.300000
9	pfizer	285	1.403509	1.754386	0.800000
10	citigroup	312	1.923077	3.205128	0.600000

Intel


Intel
pos_neg_ratio
date
1.5
1.4
1.3
1.2
1.1
1.0
0
50
100
150
200
250


US
US intelligence
say
country
claim
Iran
UK
German
China
intelligence agency
police
French
Pakistan
intelligence service
information
intelligence official
Britain
Israeli
Israel
government
said
intel
foreign
new
year
military
secret
intellectual property
WikiLeak
intelligence agencies
Snowden
Russia
security
official
intelligence
CIA
Edward Snowden
American
spy
people
report
NSA
world
document
attack
will
surveillance

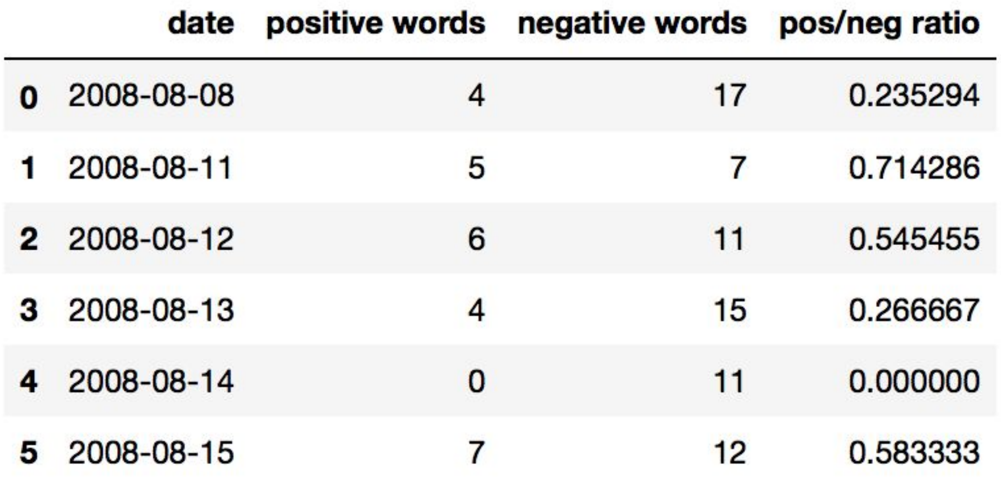
Facebook



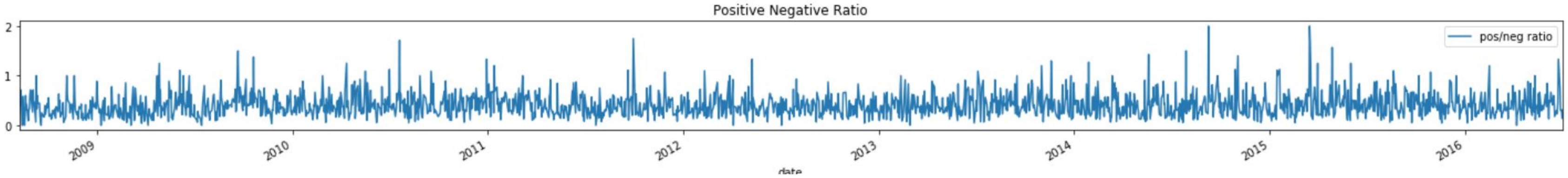
Facebook
Google
user
Twitter
privacy
Facebook page
Facebook post
Internet
government
arrested
police
court
student
photo
protest
India
UK
Turkish
new
group
year
anti
say
Israeli
death
world
Germany
Palestinian
data
post
Egypt
public
people
ban
site
jailed
posted
access
social media
law
page
face
YouTube
service
insulting
three years prison
content
web
Yahoo
Iranian
two
Russian
girl
account
man
million
teenager
secret
German
Turkey
EU

Next Step: compare ratio with company's’ stock price; still need more data

Positive-Negative Word Ratio in Everyday News



	date	positive words	negative words	pos/neg ratio
0	2008-08-08	4	17	0.235294
1	2008-08-11	5	7	0.714286
2	2008-08-12	6	11	0.545455
3	2008-08-13	4	15	0.266667
4	2008-08-14	0	11	0.000000
5	2008-08-15	7	12	0.583333

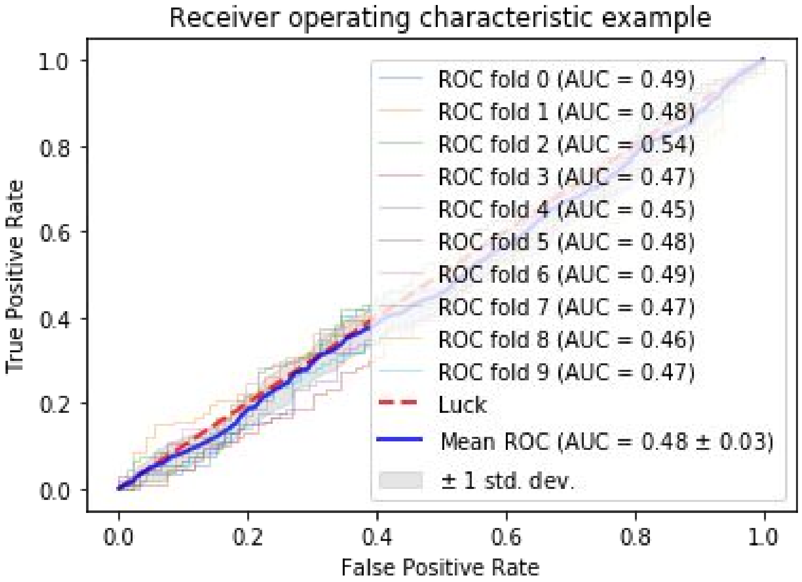


Methodology and Models

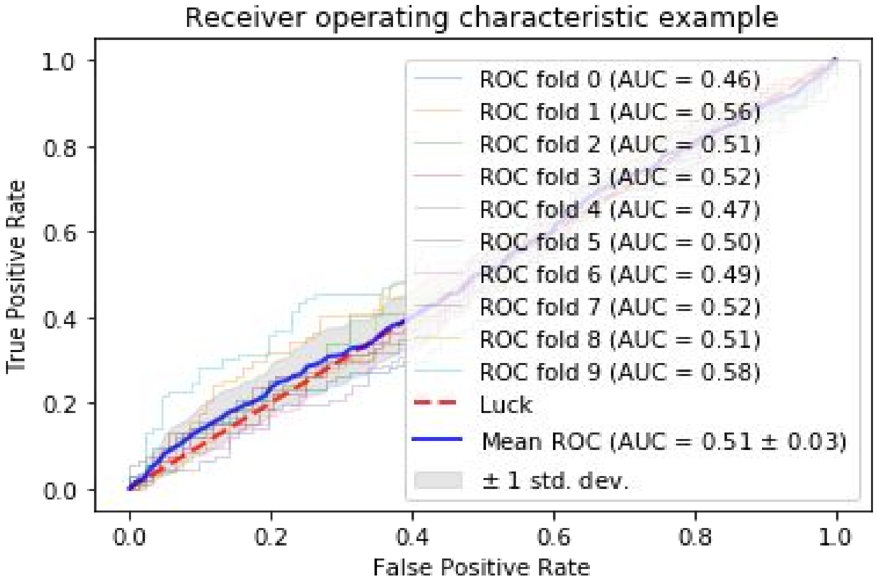
- Clean and check the dataset
- Use 10-folds Cross Validation
- Apply different classification models to fit the data and compare different models with AUC

Models

Logistic Regressions with single word

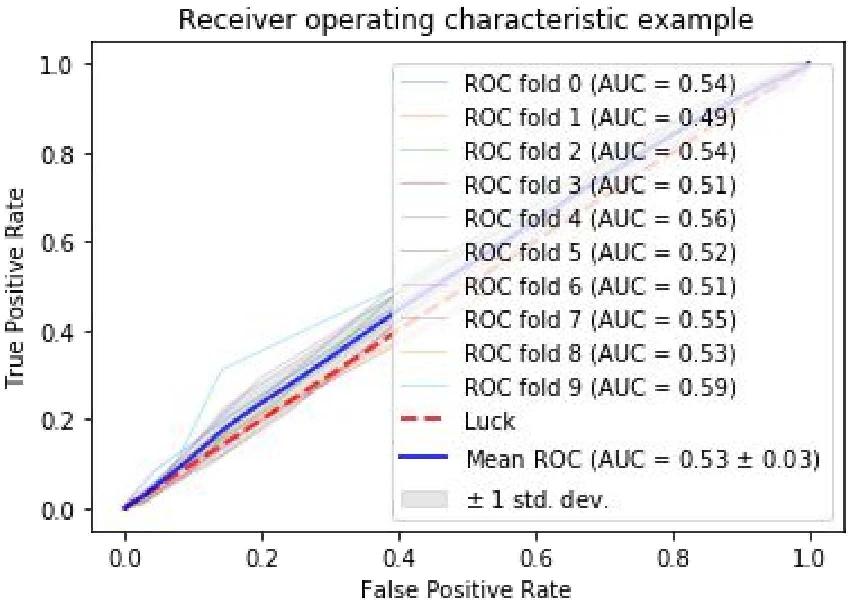


Logistic Regressions with two connected words

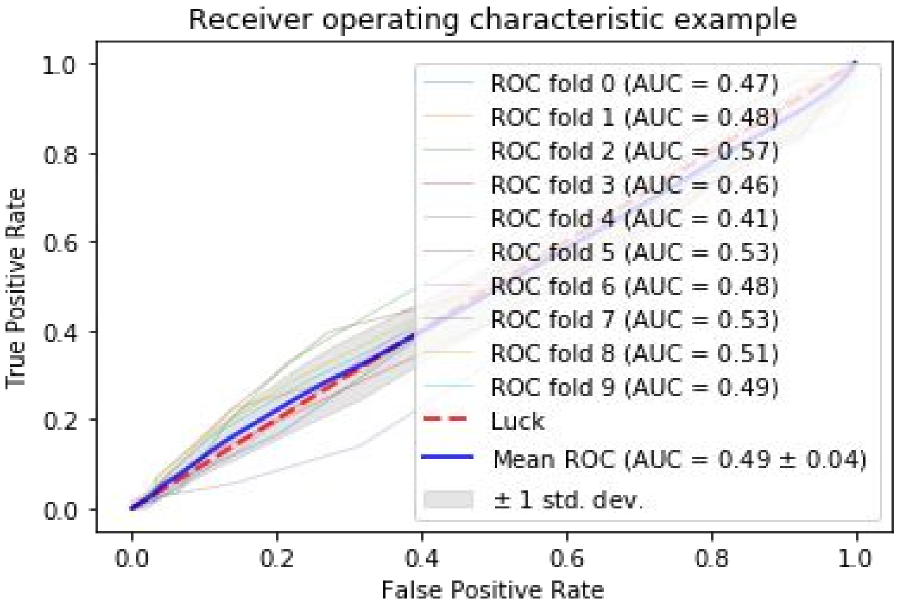


Models

Random Forest with single word

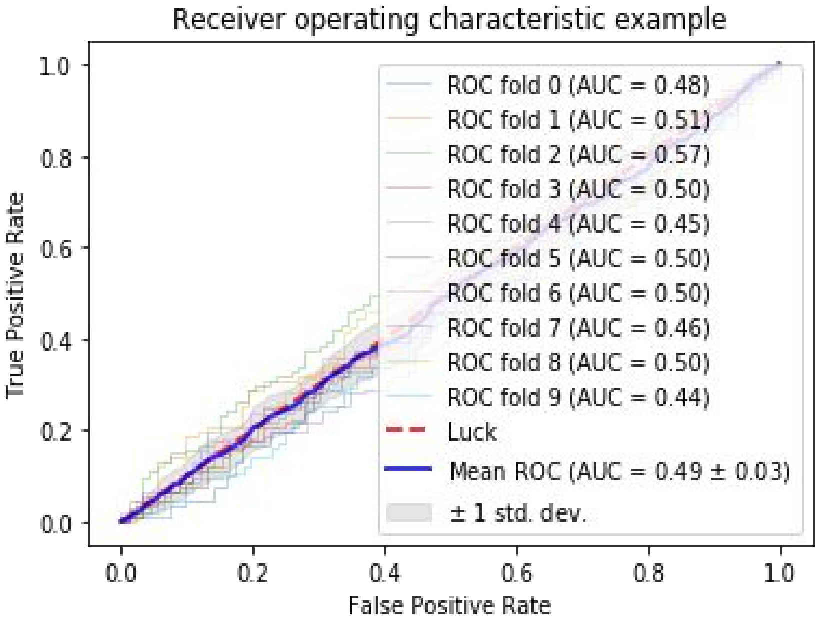


Random Forest with two connected words

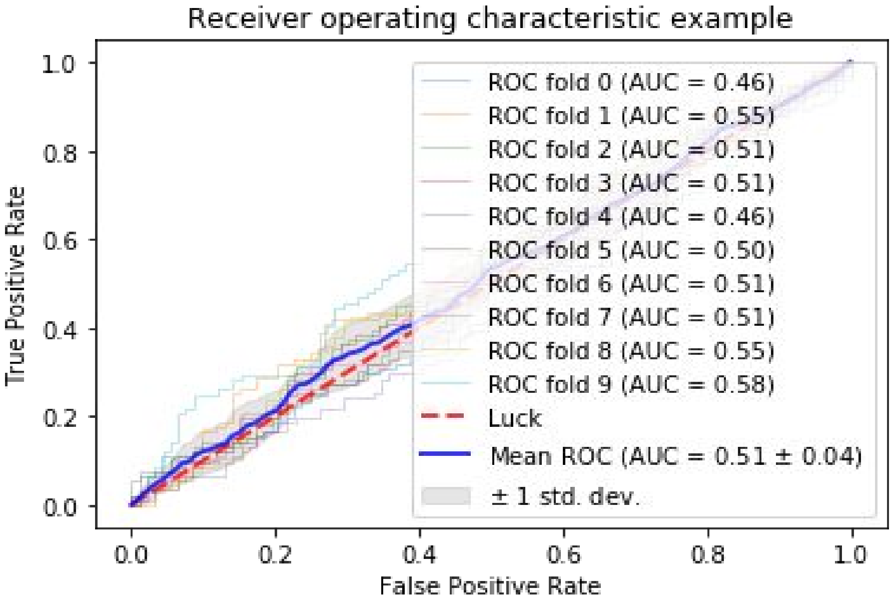


Models

Multinomial Naive Bayesian with single word

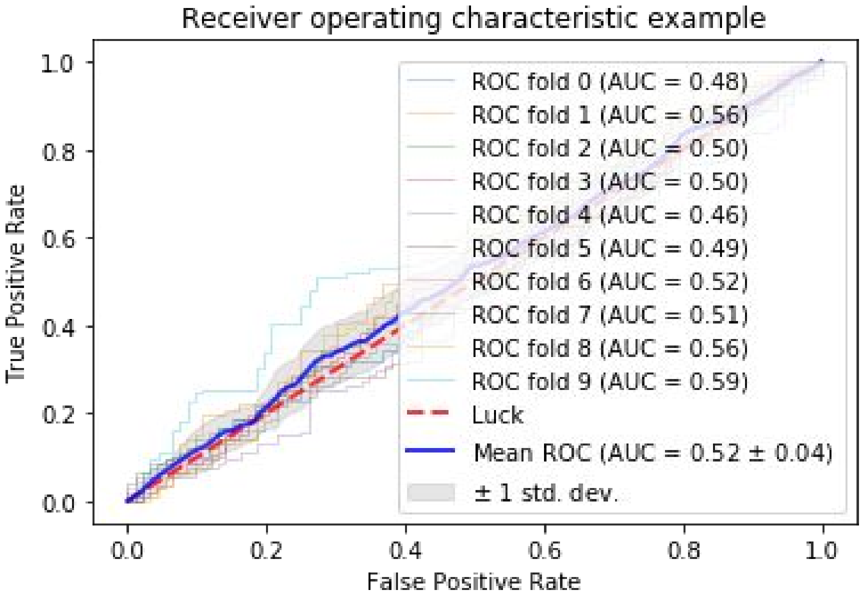


Multinomial Naive Bayesian with two connected words

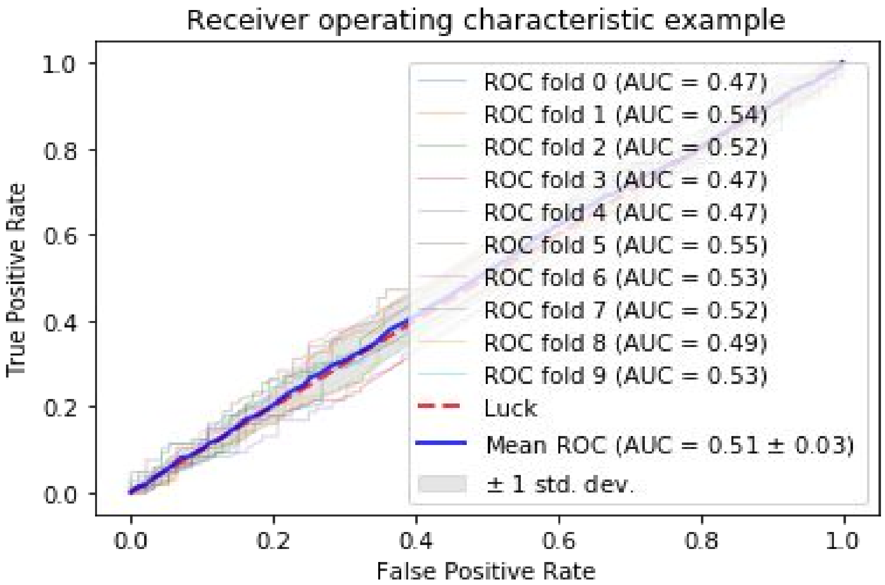


Models

Bernoulli   Naive Bayesian with two connected words

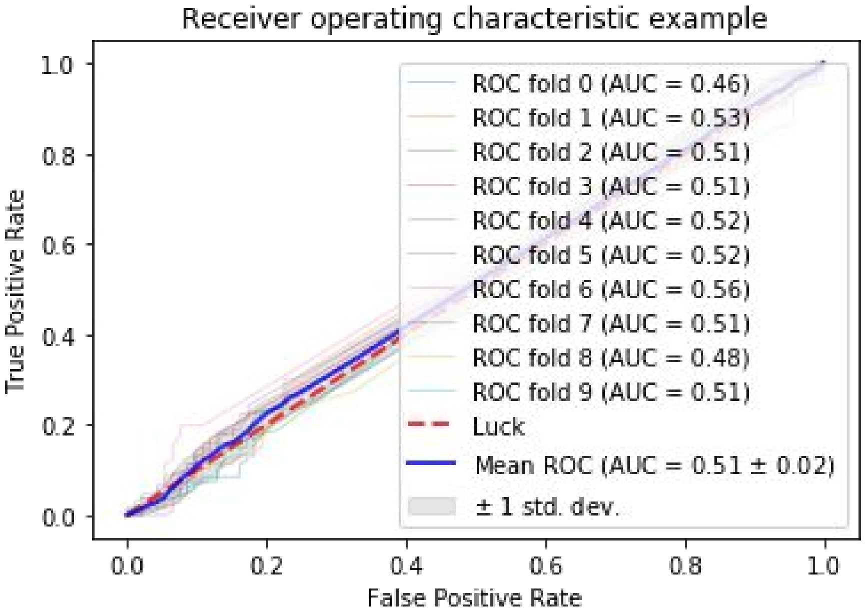


Gradient Boosting with single word

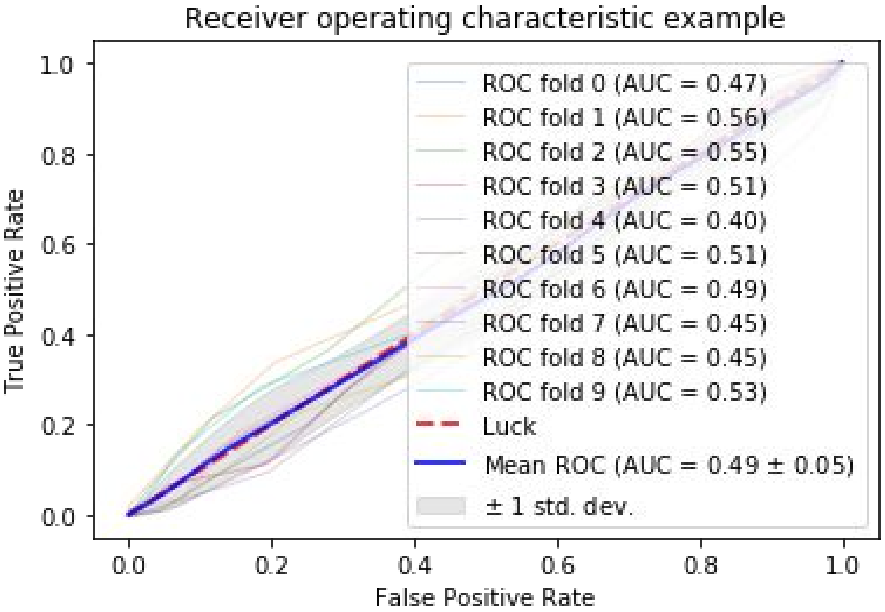


Models

AdaBoost with two connected words



Bagging with two connected word



Top 3 Models so far

Model	Best AUC	K Fold
Random Forest with single word	0.59	9
Bernoulli Naive Bayesian with two connected word	0.59	9
Logistic Regressions with two connected words	0.58	9

Limitation and Next Step:

-Only time series model and classification models

-need more data

-generate more models

-compare the models and pick the best one