Predicting stock market movement using sentiments

BUSINESS EARNINGS

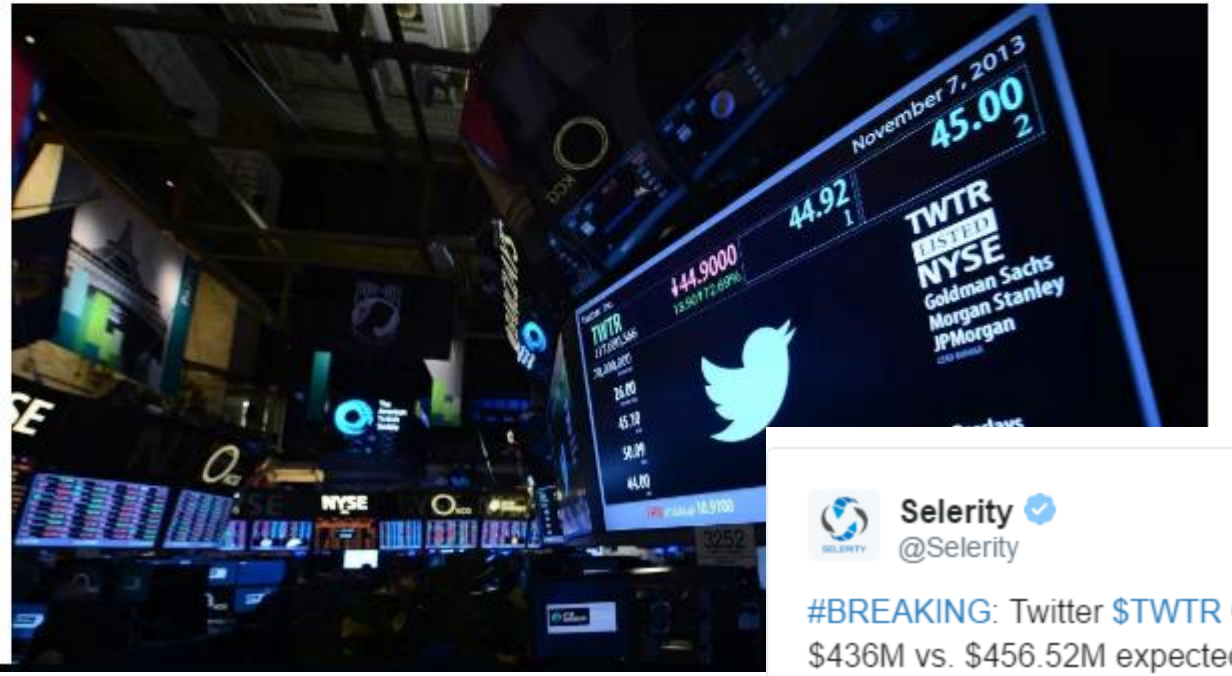
This Tweet Just Made Twitter's Stock Crash Hard

Alex Fitzpatrick @alexjamesfitz | April 28, 2015

The stock was down nearly 20%

Twitter's stock dropped nearly 20% Tuesday afternoon after disappointing quarterly earnings leaked online ahead of their expected release.

Twitter's first quarter revenue was $436 million, up 74% year-over-year but widely missing analysts' estimates of $457 million. The



Selerity
@Selerity

Follow

#BREAKING: Twitter $TWTR Q1 Revenue misses estimates, $436M vs. $456.52M expected

3:07 PM - 28 Apr 2015

154 87

Elon Musk created nearly $1B in value today with a single tweet

Elon Musk
@elonmusk

Follow

Major new Tesla product line -- not a car -- will be unveiled at our Hawthorne Design Studio on Thurs 8pm, April 30

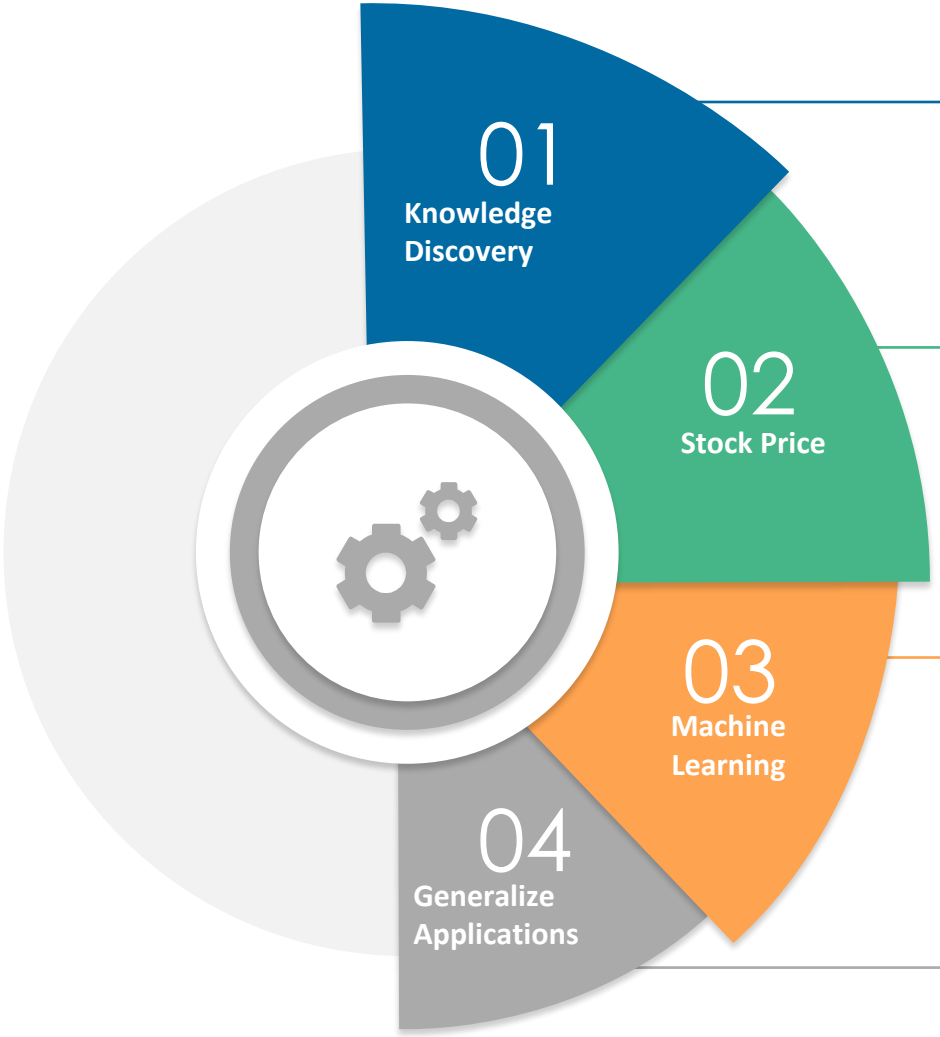
12:35 PM - 30 Mar 2015

7,340 5,874

Elon Musk tweets, Tesla shares leap

Published: Mar 31, 2015 6:13 p.m. ET

MOTIVATION & OBJECTIVE



Social Network is a communication platform containing hidden valuable knowledge. Information of social networks reflects real world events.

Owing to the fluctuations in the opinion, news and sentiment , stock price fluctuation is direct result of these Analyzing Sentiments gives a huge leverage on prediction

Sentiment analysis has been in the forefront of information retrieval using core Natural Language Processing Techniques . Machine Learning hugely makes Information Retrieval process more potent.

Much of the work on this project is generalizable to other modelling problems involving information retrieval techniques.

Business Implication

People use social networking sites, like Facebook, Twitter, etc. to express their opinions and views about a particular topic such as news, movie, event and remarks related to products.

Products

The information available from social network is beneficial for products and considers these opinions as feedback to improve their policies, planning and process for product development.

Donald Trump

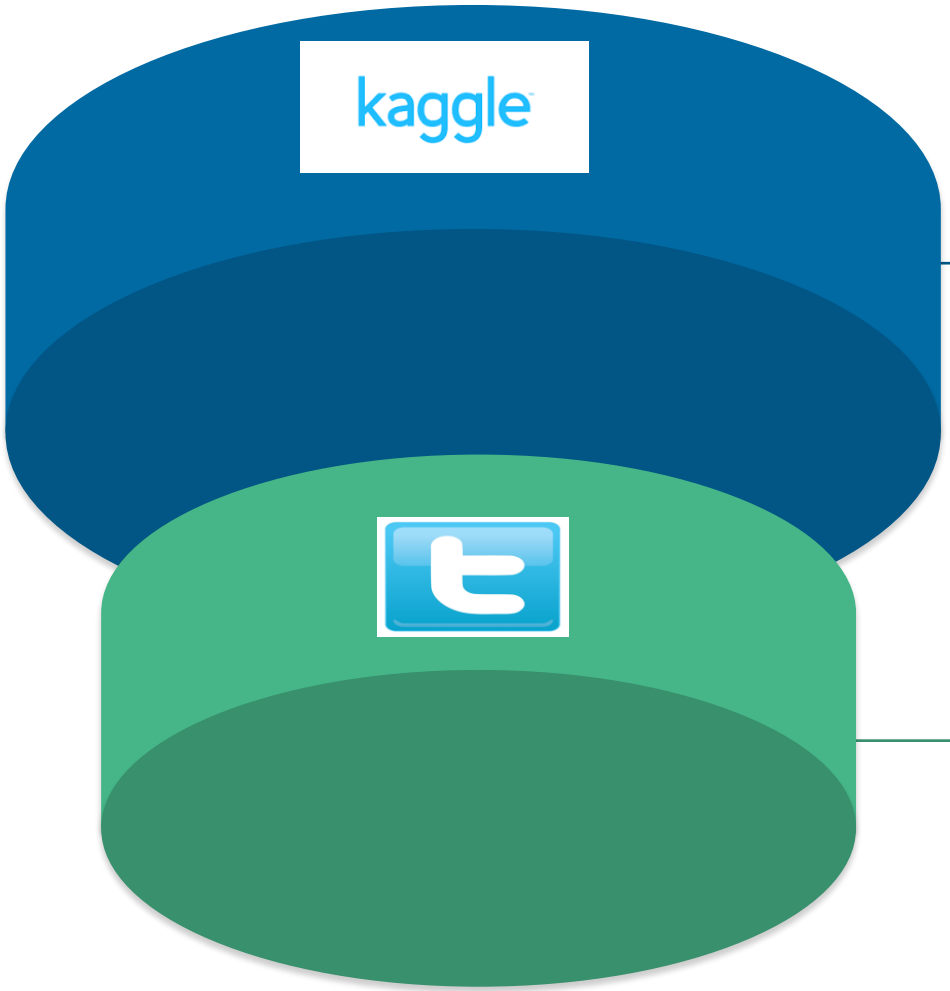
How Donald Trump's Presidency Will Affect The Stock Market .

The knowledge extracted from social media can be applied to different application.

MOVIES

Predicting blockbuster success and the Box-Office revenue of a movie.

Data Source



KAGGLE NEWS DATASET

This dataset comprises of Top 25 news dated from ‘8/8/2008’ to ‘07-01-2016’ with corresponding labels {0/1} where 0 represent drop in SPX and 1 shows an upward movement of SPX.

TWEETS DATASET

This dataset consists of Tweets along with timestamp based on SPX index from ‘1/8/2016’ to ‘3/3/2017’ with features including re-tweets and likes etc . Mapping of the data set with respect to SPX values to get downward and upward movement is done.

Data Source

TRAIN AND TEST DATA SPLIT

The News Data set is almost balanced with 908 cases of downward trend and 1079 cases of upward trend.
In the News data set model was trained on News from 2008 to 2014 and tested on data from 2015.
Train data: 1630 data points
Test Data: 356 data points

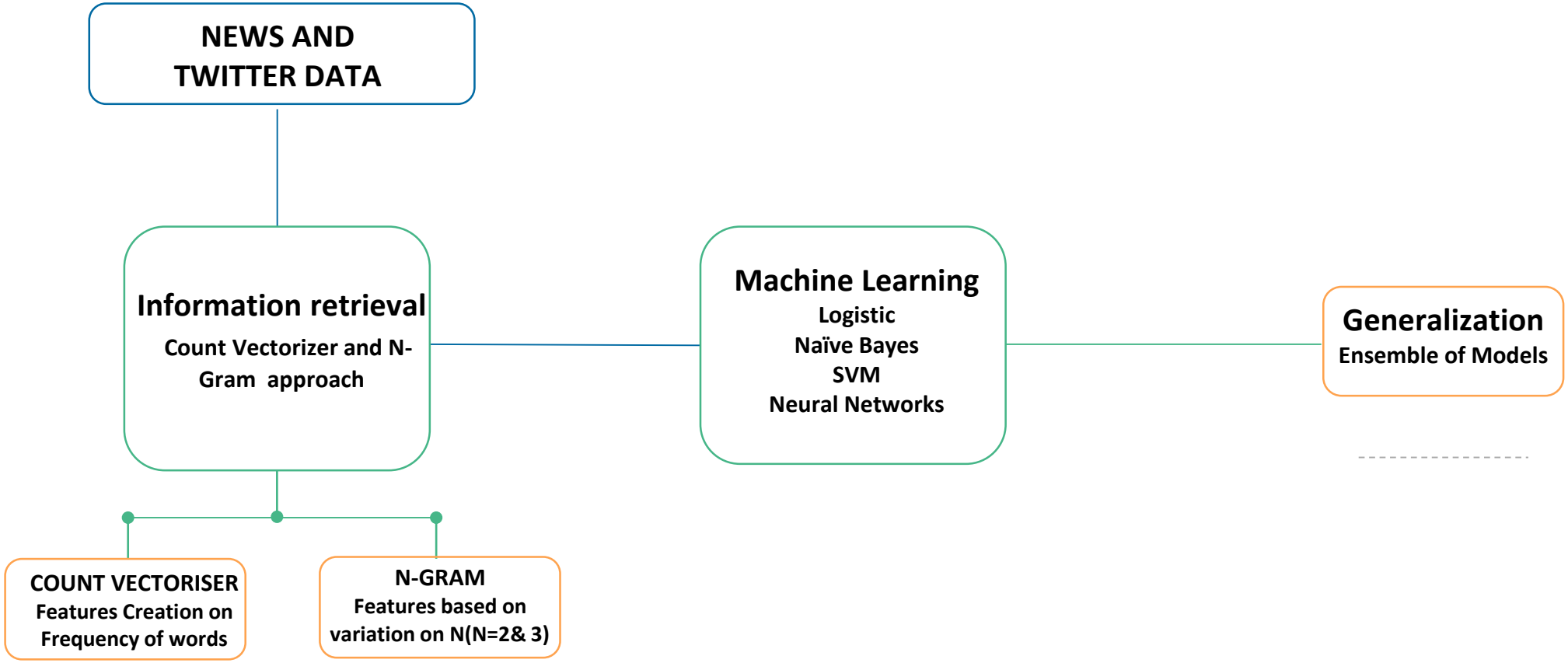
```
In [7]: data.head()
```

Out[7]:

	Date	Label	Top1	Top2	Top3	Top4	Top5	Top6	Top7	Top8	...	Top16	Top17
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 combatent' trials are nothing but...	...	b'Georgia Invades South Ossetia - if Russia ge...	b'Al-Qaeda Faces Islamist Backlash'
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...	...	b'Israel and the US behind the Georgian aggres...	b'"Do not believe TV, neither Russian nor Geor...

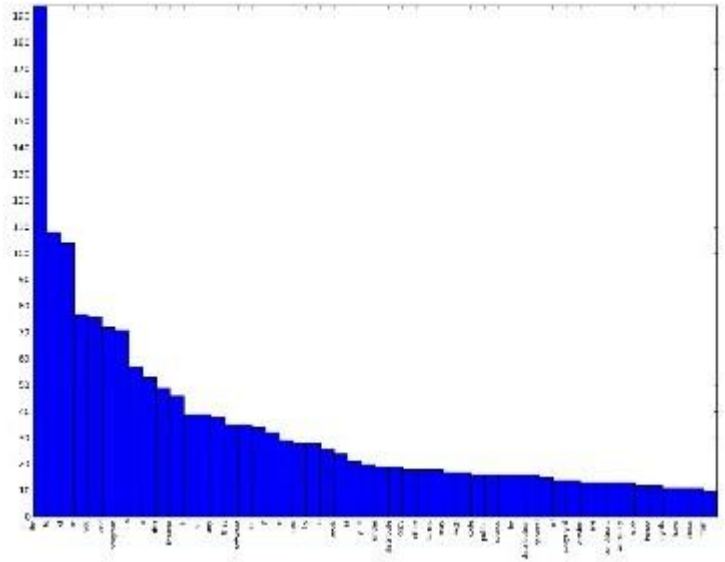
date	time	language	liked	retweet	id	text
01-08-2016	Mon Aug 01 14:27:05	en	9	23	7.6012E+17	Here's how the S&P 500 is trading and the range everyone will be watching this week: https://t.co/BzDxd4Y47q
01-08-2016	Sat Jul 30 11:00:40	en	21	35	7.59343E+17	The blended (actual + estimated) earnings decline for $SPX for Q2 is -3.8% as of Friday. https://t.co/bgq5zwelbk http
01-08-2016	Mon Aug 01 01:01:10	en	15	20	7.59917E+17	The trailing 12-month P/E ratio for $SPX is 19.4. https://t.co/bgq5zvWJMK https://t.co/yukb6wCGx8
01-08-2016	Mon Aug 01 23:59:11	en	0	8	7.60264E+17	RT @MktOutperform: The Call Heard Round the World https://t.co/IOYZl6ETP3 #stocks $SPX
01-08-2016	Mon Aug 01 23:58:06	en	0	47	7.60263E+17	RT @ElixiumCapital: Trade with #Bitcoin
01-08-2016	Mon Aug 01 23:56:49	en	0	0	7.60263E+17	Deep-water mega- $Oil-projects that can take 7- 10 years. In that amount of time I sure hope we have found a sustai
01-08-2016	Mon Aug 01 23:52:19	in	0	29	7.60262E+17	RT @ElixiumCapital: $AAPL $FB $TSLA $TWTR $NFLX $GOOGL $SPY $SPX $ES $USDJPY #GOLD #OIL #BITCOIN #TRADIN
01-08-2016	Mon Aug 01 23:52:01	en	0	7	7.60262E+17	RT @TheDayTradr: Trendline to watch on $SPX for a swing short, 2185-2200 short zone https://t.co/tGJfY77p91
01-08-2016	Mon Aug 01 23:50:37	en	0	0	7.60261E+17	The $SPX / $JNK spread is back to March levels & Correlation is breaking which usually does not end well https:

METHODOLOGY



Term-document counts

- The idea behind this is a document(news) is best describe by the most frequent terms.



- Top 20,0000 words with most no. of counts are taken as features.
- We trained our model with these features and got an accuracy of 50.36 %.
- How can we improve on this ? TF-IDF

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	157	73	0	0	0	0
Brutus	4	157	0	1	0	0
Caesar	232	227	0	2	1	1
Calpurnia	0	10	0	0	0	0
Cleopatra	57	0	0	0	0	0
mercy	2	0	3	5	5	1
worser	2	0	1	1	1	0

TF-IDF

Term frequency (tf)

- The term frequency $\mathrm{tf}_{t,d}$ of term t in document d is defined as the number of times that t occurs in d.

Document frequency (df)

- The document frequency $\mathrm{df}_{t,N}$ of term t is defined as the number of documents that t occurs in .

Inverse Document frequency (idf)

- The inverse document is defined by total number of documents in collection by the total number of documents t occurs in. It is the measure of the general importance of the term.

Term frequency-Inverse Document frequency (tf-idf)

- It is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

$$w_{x,y} = \mathrm{tf}_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

term x within document y

$\mathrm{tf}_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

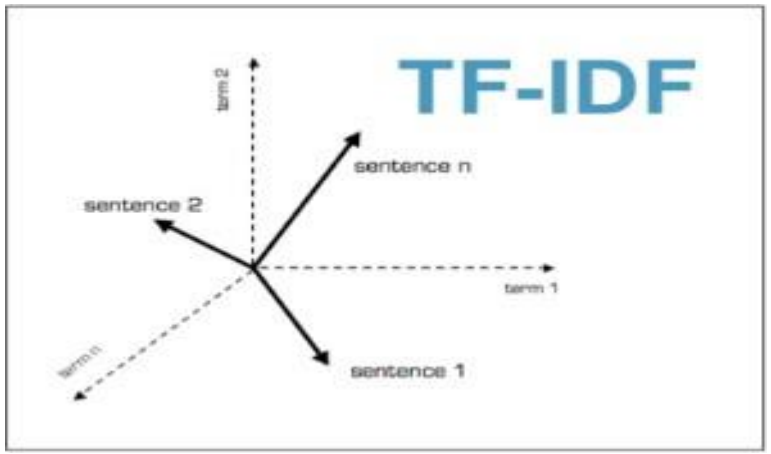
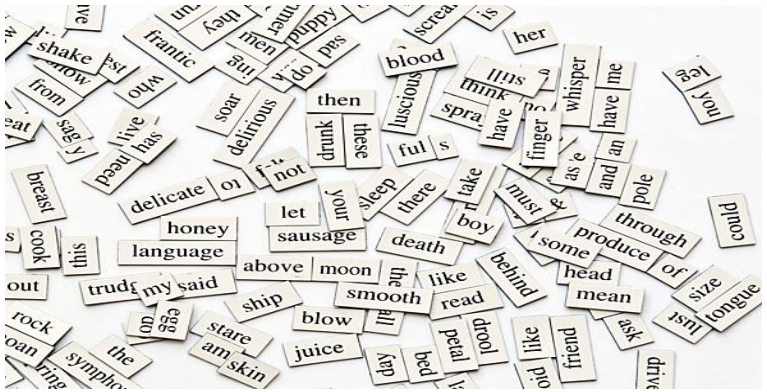


Fig. Representation of document in TF-IDF space

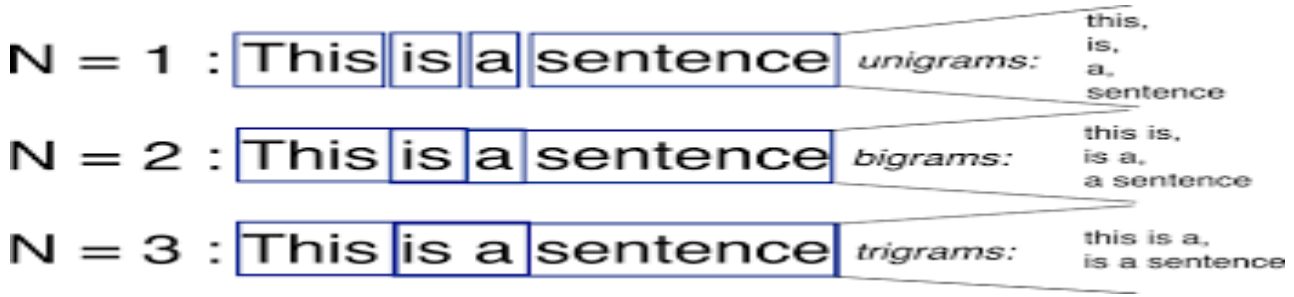
Bag of words ?



N-Gram Features

What are N-Grams ?

They are basically a set of co-occurring words within a given window and when computing the n-grams you typically move one word forward.



01 To *featurize* long text strings, extracting only the most important pieces of information.

02 For scoring N-grams we have used tf-idf weights .

Why N-grams performs better than Bag of words ?

- Dealing with negation
    - Movie not good.

MODEL ANALYSIS

<50%

Preprocessing: Count Vector and TfIdf
Model : Naive Bayes, Logistic Regression, SVM, Neural Network

53.96%

Preprocessing: Bigram TfIdf
Model : Naive Bayes

57.14%

Preprocessing: Bigram TfIdf
Model : Logistic Regression

51.58%

Preprocessing: Trigram TfIdf
Model : Neural Network

54.49%

Preprocessing: Trigram TfIdf
Model : SVM

58.20%

Preprocessing: Bigram TfIdf
Model : Neural Network

59.52%

Preprocessing: Bigram TfIdf
Model : SVM

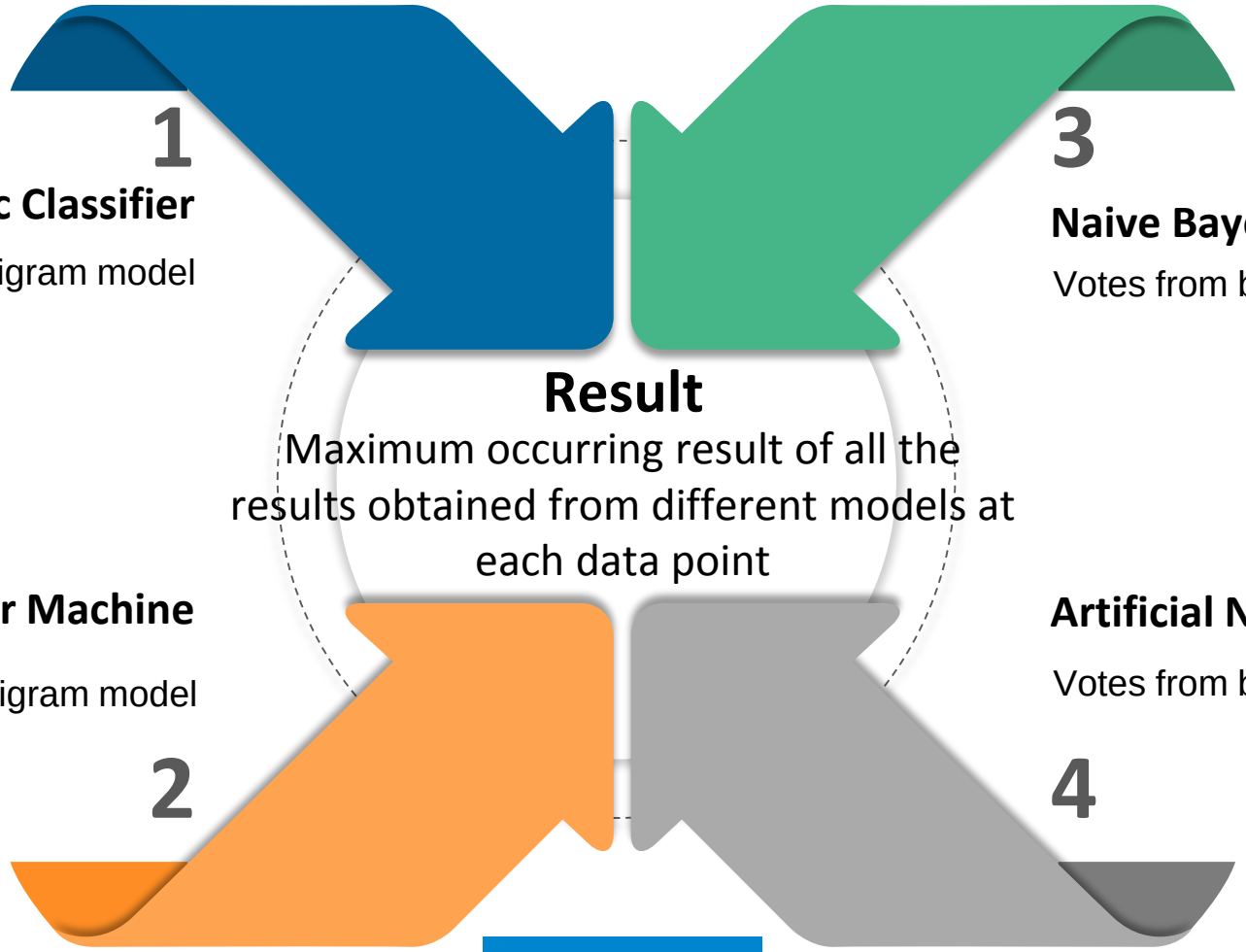
55.82%

Preprocessing: Trigram TfIdf
Model : Logistic Regression

57.61%

Preprocessing: Trigram TfIdf
Model : Naive Bayes

Ensemble: Voted Classifier



1

Logistic Classifier

Votes from bigram and trigram model

3

Naive Bayes

Votes from bigram and trigram model

Result

Maximum occurring result of all the results obtained from different models at each data point

Support Vector Machine

Votes from bigram and trigram model

2

Artificial Neural Network

Votes from bigram model

4

57.42 %

TABLEAU VISUALIZATION

POST MIDSEM PLANNING

Twitter Data Usage for prediction of Stock Movements

Integration of Twitter and News based Stock Prediction Models

Strategy formulation and validation based on short term P&L and drawdown Approach


Thank you!