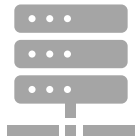
IND5003

Market Trend Prediction

Using Sentiment Analysis on data extracted from *StockTwits*

Outline

Data Collection



Data Pre-processing

Sentiment Analysis

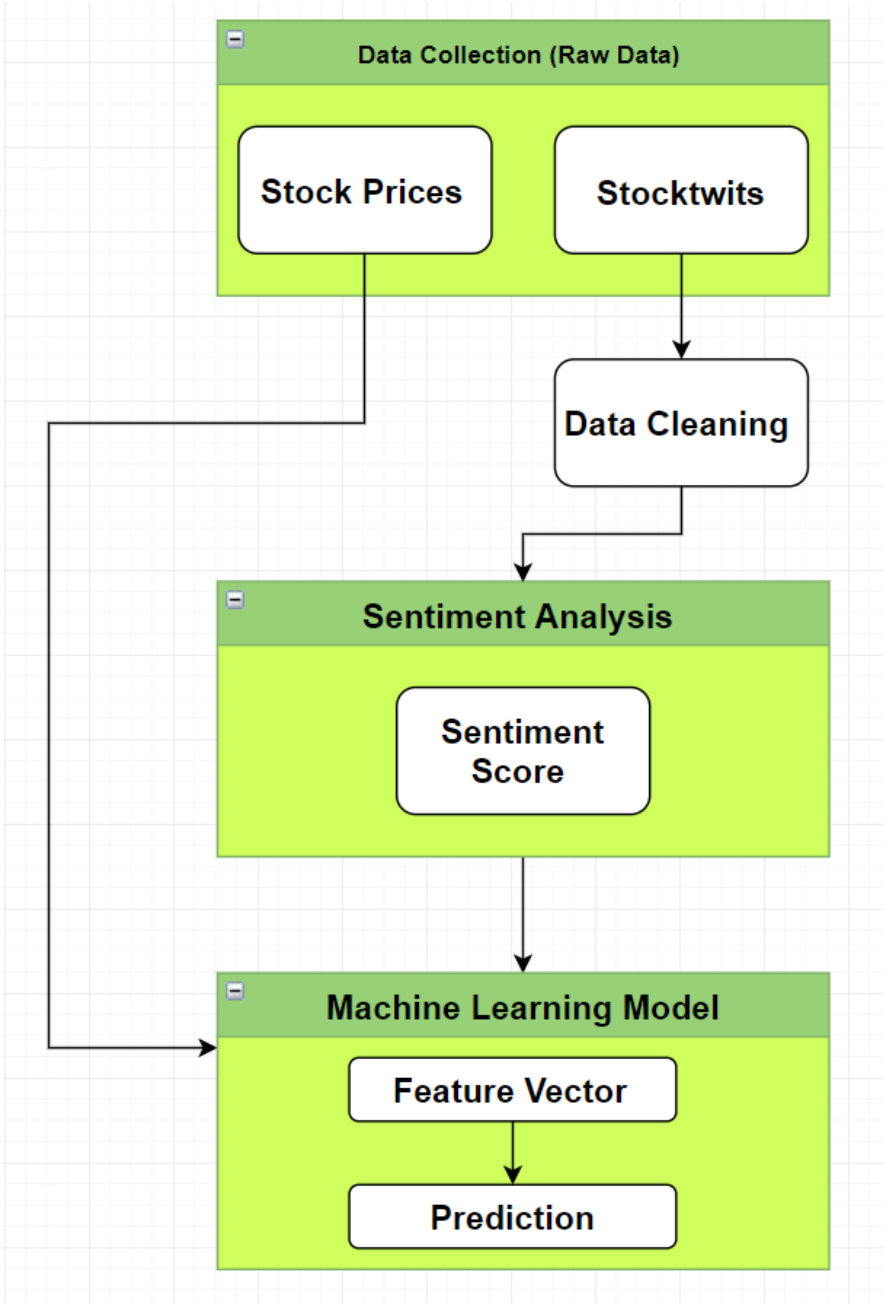
Feature Vector Formation

Prediction

Problem Statement

Predict if Apple's stock (AAPL) would go *up* or *down* by using:

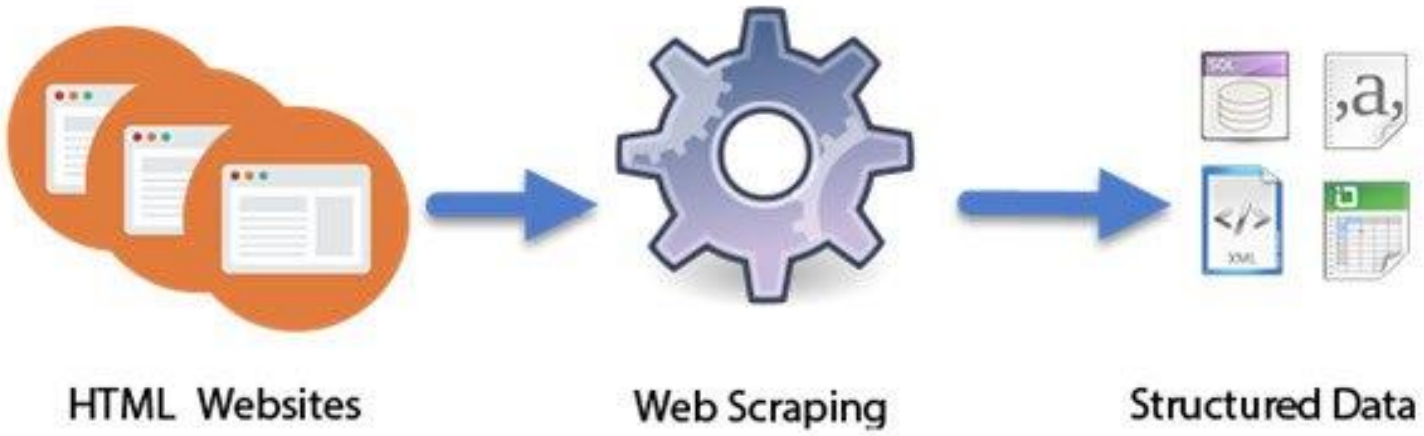
- AAPL stock price data from (since 2016)
- Sentiments from *StockTwits* data of the same time period (37K Messages)



Collection of Data

StockTwits and Stock Prices

Scraping



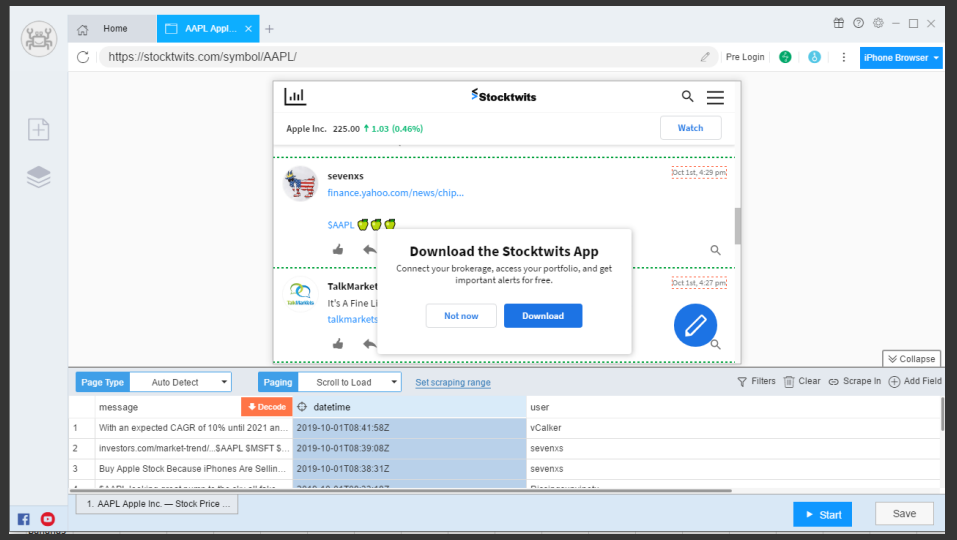
- Computer software technique to extract information from websites
- Storing of unstructured data for Analysis
- Why Scraping Websites?

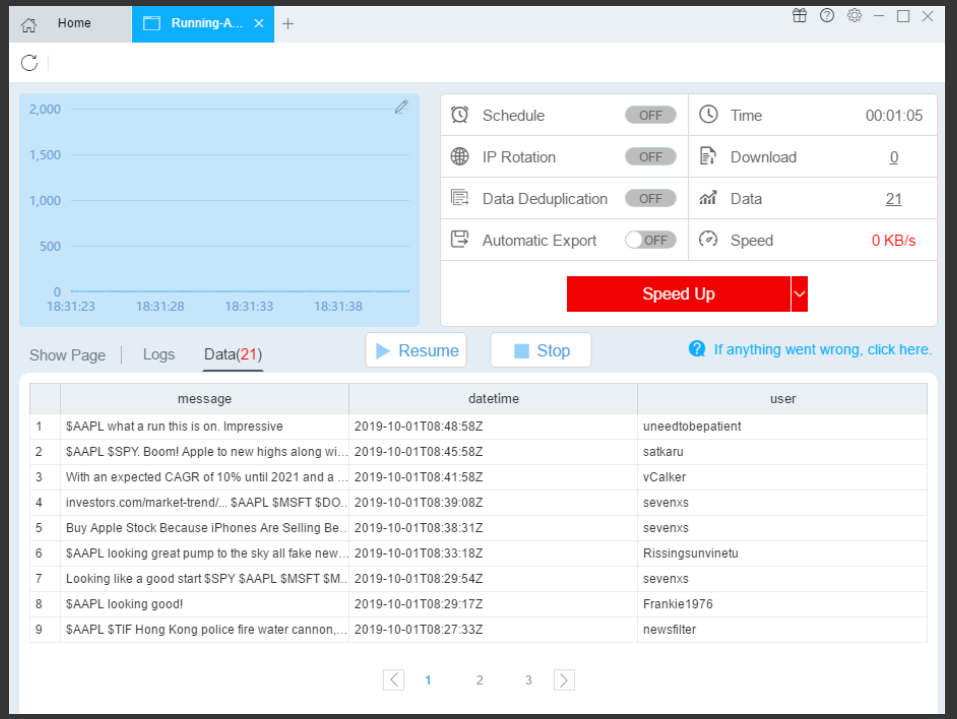
Stocktwit Scraping

Tool	Required Instillation	Subscription	Accuracy %	Operation System
ScrapeStorm	Yes	FREE	100%	All
ParseHub	Yes	Yes	100%	All
webscraper	No	1,000 URL queries per month Free	100%	Cloud

ScrapeStorm

- Intelligent identification of data, no manual operation required
- Point and click
- Fully web-based and hosted scraping





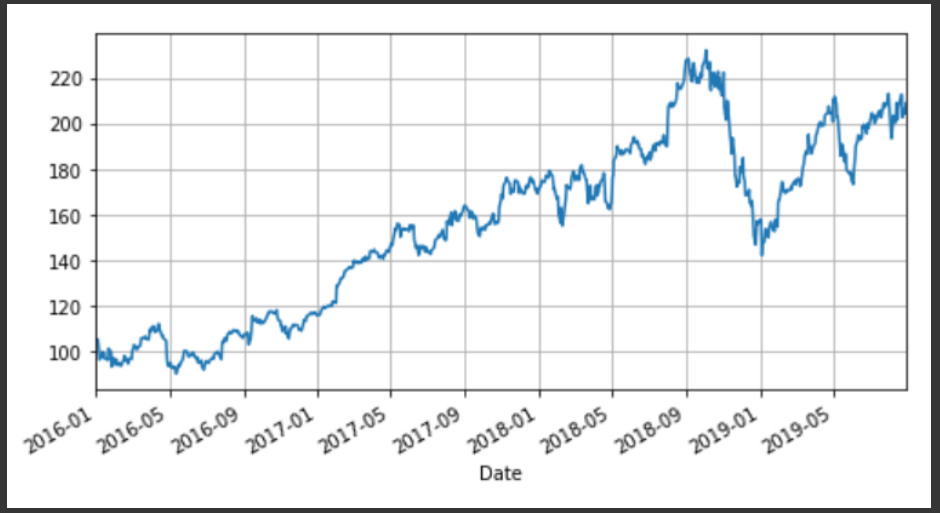
Stock Prices data

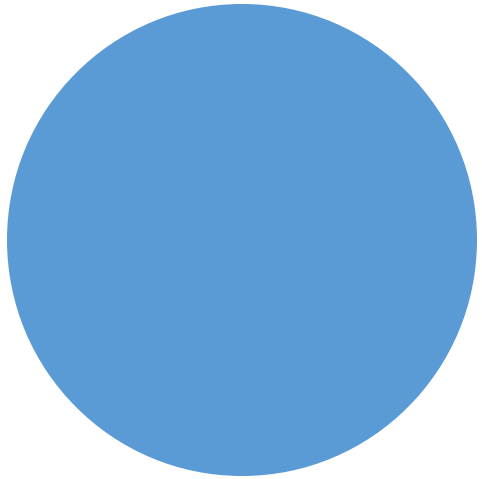
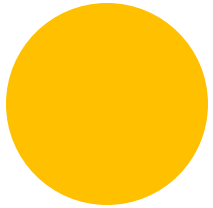
- It is not an API
  - Yahoo! Finance decommissioned their historical data API
  - Python library
- Originally named `fix-yahoo-finance`
- Scrape data from Yahoo! Finance and return them in DataFrame format

```
import yfinance as yf

apple = yf.download('AAPL','2016-01-01','2019-09-01')
apple.head(2)
```

Date	Open	High	Low	Close	Adj Close	Volume
2015-12-31	107.01	107.03	104.82	105.26	98.66	40912300
2016-01-04	102.61	105.37	102.00	105.35	98.74	67649400



Sentiment Analysis

Sentiment Analysis – What?

- Using NLP to identify opinions or emotions conveyed in a piece of text
  - Positive, Negative, Neutral
  - Happiness, Anger

Sentiment Analysis – How?

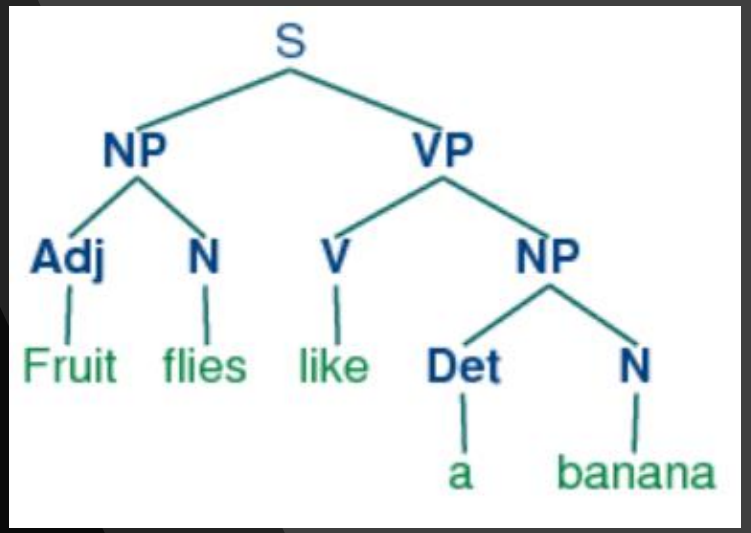
- Pre-processing
- Feature Extraction
- Model Training & Classification

Sentiment Analysis – How?

- Pre-processing
  - Standardize the letter case
  - Change numbers into words
  - Remove punctuation, special characters, stop words
  - Tokenization
  - Stemming and Lemmatization
  - Negation handling

Sentiment Analysis – How?

- Feature Extraction
  - Part-of-speech (POS) tagging
  - Opinion words are identified and weighted
  - Bag-of-Words (BoW)



"Our dog is like family to us"
"We really like our dog"

Positive	Neutral	Negative
Joy	is	disgusting
Happy	and	sad
Sweet	they	unpleasant

Sentiment Analysis – How?

- Model Training & Classification
  - Naive Bayes
  - Logistic Regression
  - Support Vector Machines

Sentiment Analysis – Why?

To analyze sentiments/emotions towards services, products, or events to improve business strategies, policies, etc.

Sentiment Analysis – Project Context

Monitor sentiments from StockTwits posts on Apple's stock to predict stock price changes

Sentiment Analysis Overview

Decided to use pre-trained models

- No labelled StockTwits data to train our own model

Process

- Selection of Model
- Pre-processing
- Implementation

Selection of Sentiment Analysis Model

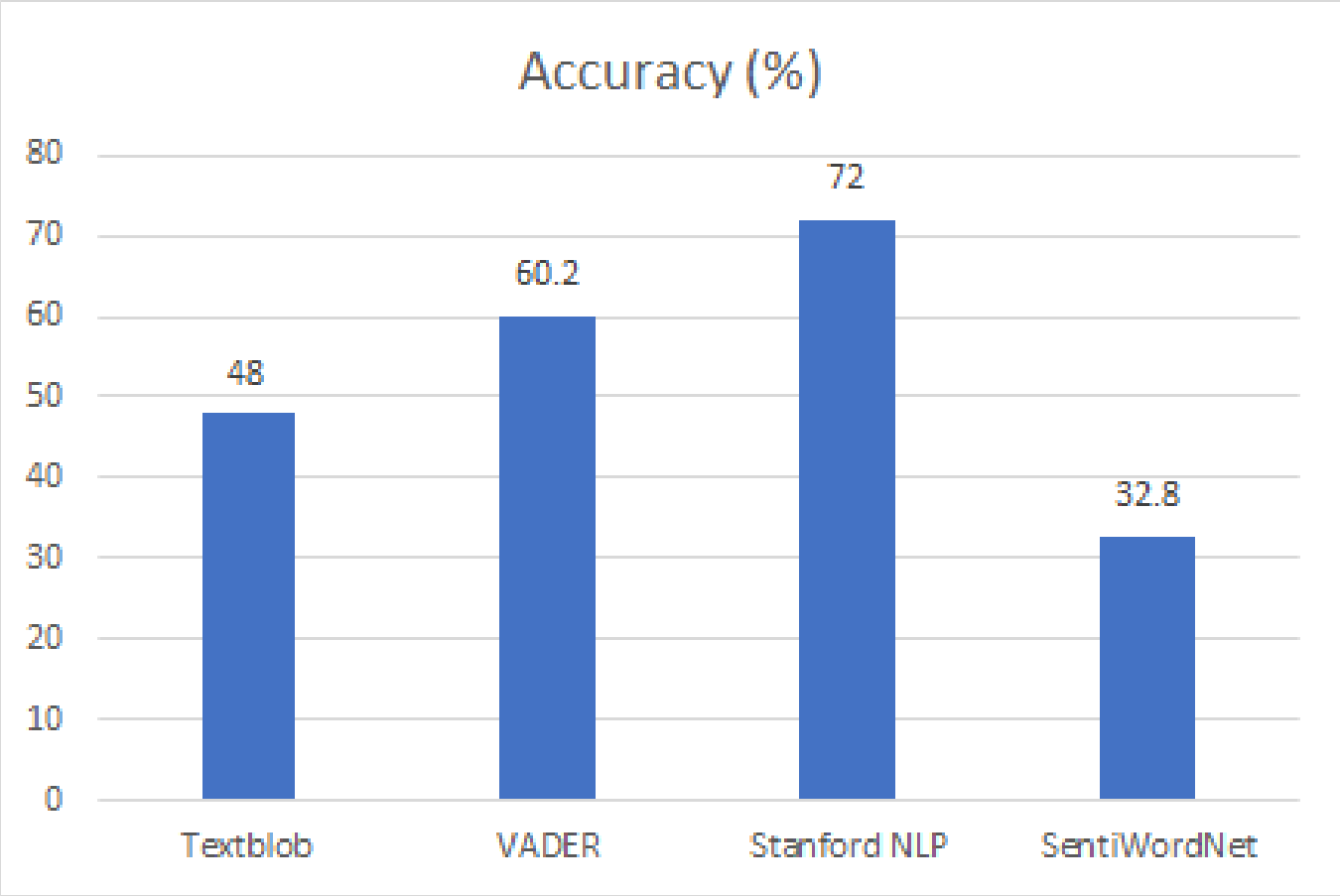
Models we tried:

TextBlob | VADER | Stanford NLP | SentiWordNet

Test accuracy on manually labelled Twitter data (3424 Tweets)

	Topic	Sentiment	TweetId	TweetDate	TweetText
0	apple	positive	126415614616154112	Tue Oct 18 21:53:25 +0000 2011	Now all @Apple has to do is get swype on the i...
1	apple	positive	126404574230740992	Tue Oct 18 21:09:33 +0000 2011	@Apple will be adding more carrier support to ...
2	apple	positive	126402758403305474	Tue Oct 18 21:02:20 +0000 2011	Hilarious @youtube video - guy does a duet wit...
3	apple	positive	126397179614068736	Tue Oct 18 20:40:10 +0000 2011	@RIM you made it too easy for me to switch to ...
4	apple	positive	126395626979196928	Tue Oct 18 20:34:00 +0000 2011	I just realized that the reason I got into twi...

Selection of Sentiment Analysis Model



Model	Avg Time Taken Per Sample (s)
Textblob	10.64
VADER	0.0005476
Stanford NLP	2.920
SentiWordNet	0.009168

VADER (Valence Aware Dictionary for Sentiment Reasoning)

- Specially built for social media posts and short texts
- A pre-built SA model included in the NLTK package.

```
In [12]: import nltk
         from nltk.sentiment.vader import SentimentIntensityAnalyzer

         nltk.download('vader_lexicon')
         sia = SentimentIntensityAnalyzer()
```

- Its lexicon is built by humans employed via a crowd-sourcing e-platform -**Amazon Mechanical Turk**
- Recall: Lexicon is just a dictionary

VADER (5 Heuristics)

1) PUNCTUATION

I’M HUNGRY VS I’M HUNGRY!!!

2) Capitalization

I’m hungry!! vs I’M HUNGRY!!

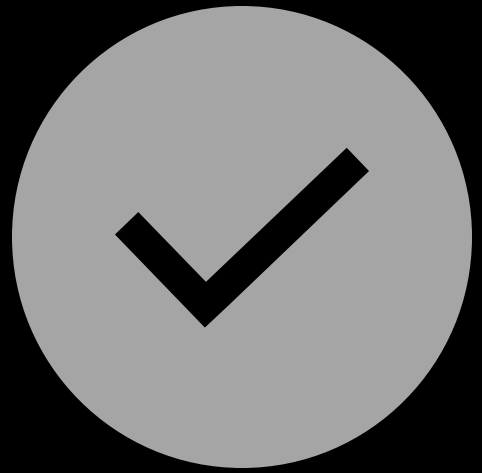
3) Degree modifiers
(use of intensifiers)

I'M HUNGRY!! VS
I'M REALLY HUNGRY!!

VADER (5 Heuristics)

4) Conjunctions
(shift in sentiment polarity, with later dictating polarity)

I love pizza, but I really hate Pizza Hut.
(bad review, instead of good)



5) Preceding Tri-gram
(identifying reverse polarity by taking preceding words into account)

Canadian Pizza is not really all that great.

Vader (Other Acceptances)

Emoji

"I am 😁 today!"

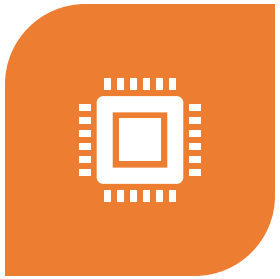
Slang

"Today only kinda sux! But I'll get by, lol"

Emoticons

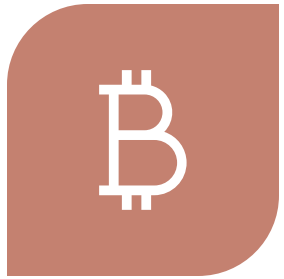
"Make sure you :) or :D today!"

Ie. There is no need to correct every text input into proper sentences.

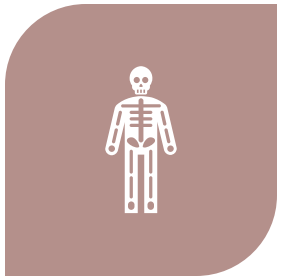


REMOVED LINKS (STARTING WITH 'WWW.' , 'HTTP://', 'HTTPS://')

REMOVED USER TAGS (E.G. '@ST3PHENCURRY)



REMOVED TICKER SYMBOLS (E.G. '$AAPL')



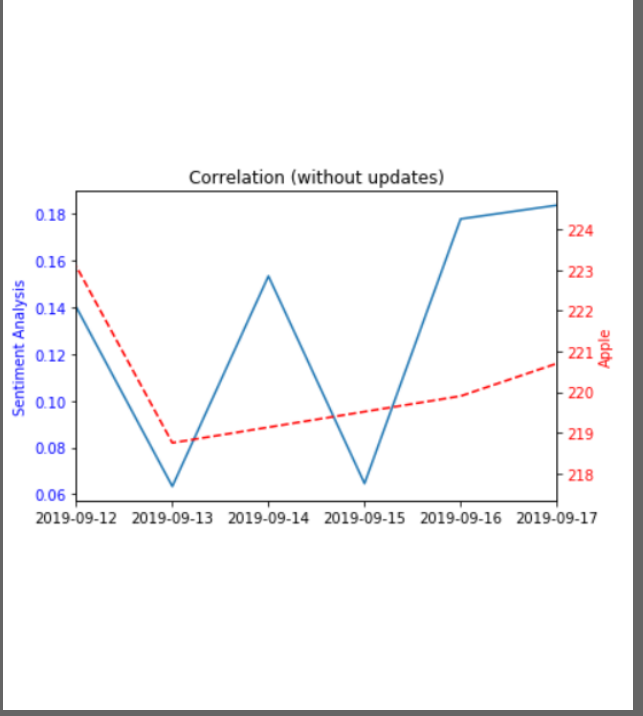
REMOVED SPECIAL CHARACTERS ('@', '$', '%', '/', '\', '_', '-')

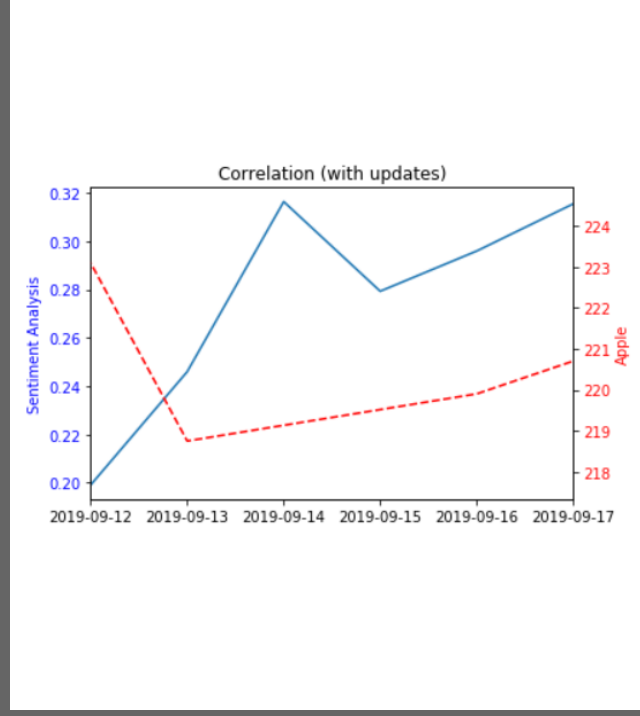
AND SO ON...

Pre-processing

Update VADER Lexicon

- Update with external sources/lexicons that are often used in financial applications
  - ➢ E.g. Loughran-McDonald Financial Sentiment Word Lists
- The 'weights/strength' of each word is updated in the lexicon.
  - ➢ Make the sentiment score more accurate to Stocktwits' posts






```
final_lex = {}
final_lex.update({word:2.0 for word in positive})
final_lex.update({word:-2.0 for word in negative})
sia.lexicon = final_lex
```

Generating the Sentiment Score

```
text = st.message[2]
score = sia.polarity_scores(text)['compound']

print('The text:\"{0}\" has a score of {1}.'.format(text, score))
```

```
The text:"Possible low risk trading setup on buy stop entry " has a score of 0.3119.
```

- Return the **polarity (+ve or –ve)** as well as the **intensity** strength of the emotion of a text.
  - Range from –1 to 1
  - E.g. – 0.6875
- Repeat for all the texts scrapped

datetime	message	polarity_score
2019-08-02 17:09:07	is a beast! Along with and !!! And hopefully ...	0.7188
2019-07-30 20:48:27	I Tood you sooooo learn next time i wont live...	0.4682
2019-08-05 19:57:50	lets buy the dip to bounce	0.2870

Processed Sentiment analysis data

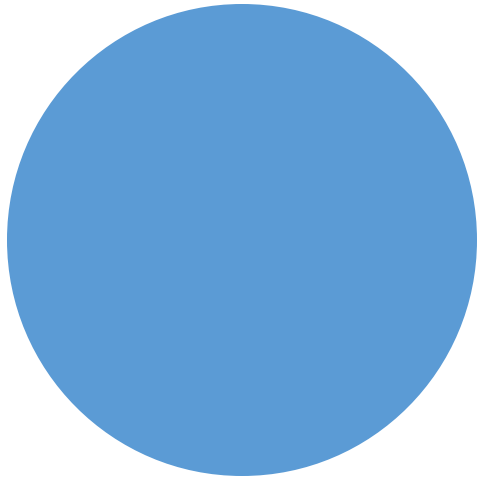
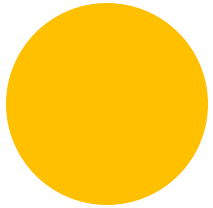
- Score of exactly 0.0000 usually is just meaningless text, and was removed

message	polarity_score
SHORTATHON!	0.0
ð□□□ð□□□ð□□□	0.0
Trending	0.0

- Aggregate the scores into averages of every hour for each day
  - Passed into the prediction model

	(polarity_score, 0)	(polarity_score, 1)	(polarity_score, 2)	. . . 24
2019-07-30	0.424039	0.533864	0.273753	
2019-07-31	0.284620	0.412420	0.257259	

64 rows × 24 columns

Prediction

In depth overview

Prediction: How is it done?

- Preparing Data

Market Data

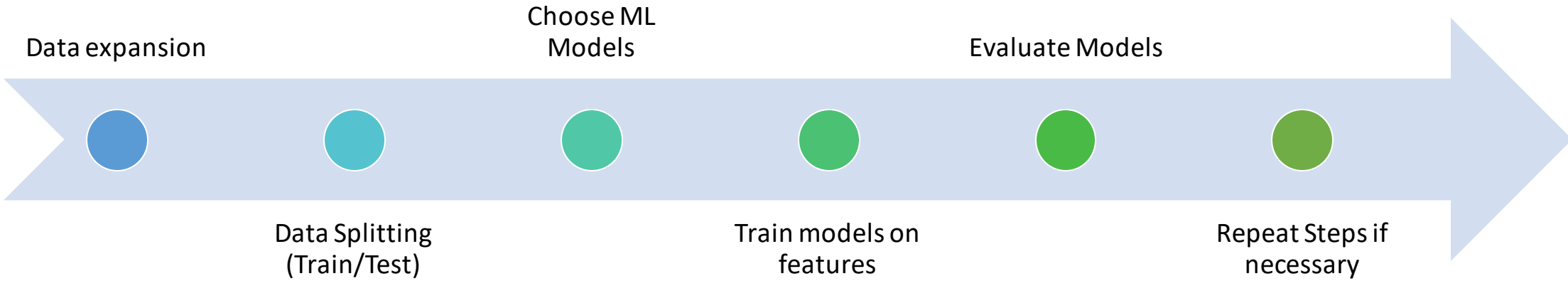
	Date	Open	High	Low	Close	Adj Close	Volume
0	2015-12-31	107.01	107.03	104.82	105.26	98.66	40912300.0
1	2016-01-04	102.61	105.37	102.00	105.35	98.74	67649400.0

Previous 5 Days Closure

Sentiment

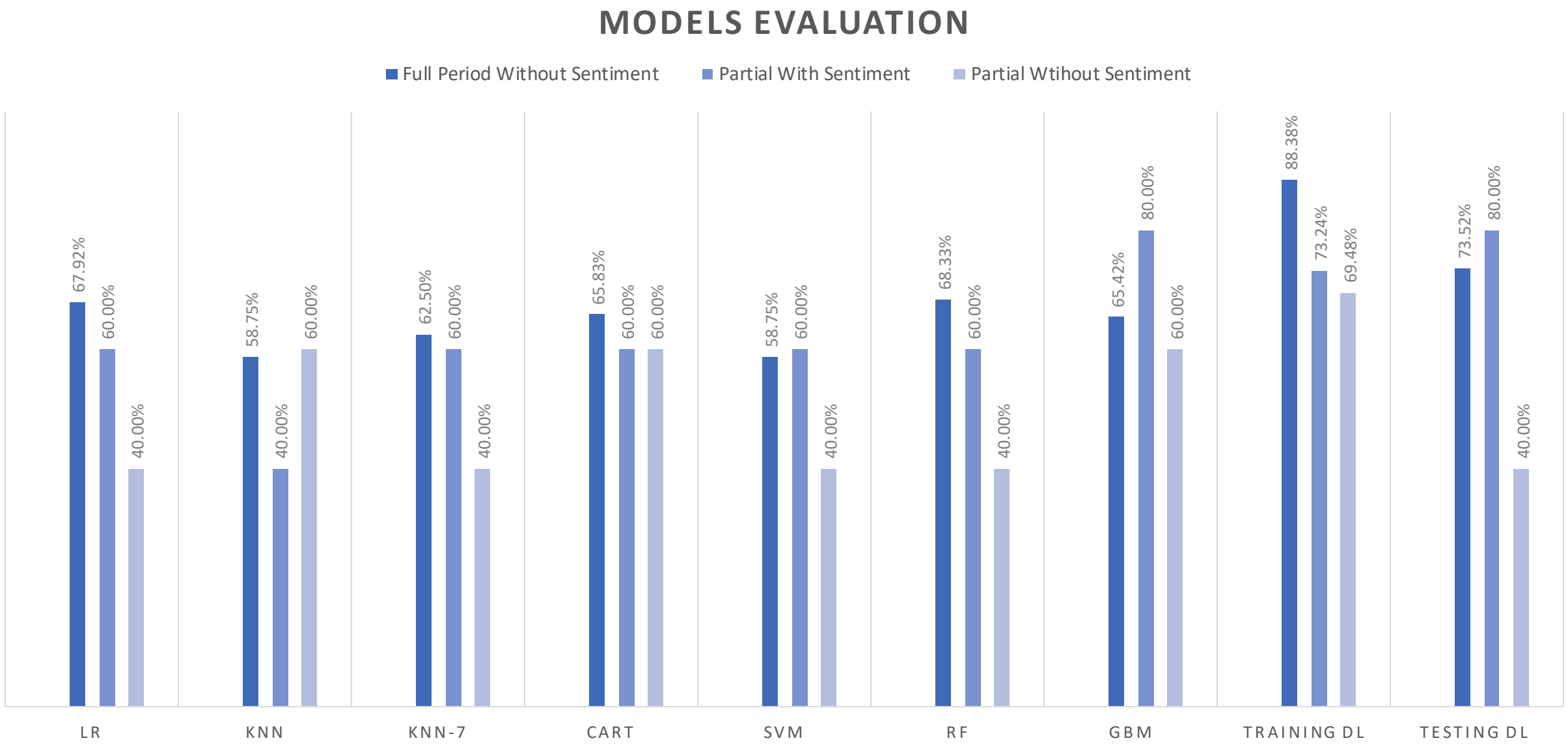
Used for training

Predicting Next day market movement (↑↓)



Prediction: Process

Evaluation


MODELS EVALUATION
Full Period Without Sentiment
Partial With Sentiment
Partial Wtihout Sentiment
67.92%
60.00%
40.00%
58.75%
40.00%
60.00%
62.50%
60.00%
40.00%
65.83%
60.00%
60.00%
58.75%
60.00%
40.00%
68.33%
60.00%
40.00%
65.42%
80.00%
60.00%
88.38%
73.24%
69.48%
73.52%
80.00%
40.00%
LR
KNN
KNN-7
CART
SVM
RF
GBM
TRAINING DL
TESTING DL

Conclusion

KNN

+20% (GBM)

+40% (DL)

Sentiment Wins!

Future Contribution

Data Acquisition

Automating some processes in the workflow

Timeseries Forecasting

Deep Learning Specialized methodologies


#TheRundownBET
Questions?


THAT IS AN EXCELLENT QUESTION

References

- https://www.google.com/url?sa=i&source=images&cd=&ved=2ahUKEwi_0pKZg_3kAhWUfisKHc7MC5IQjRx6BAgBEAQ&url=https%3A%2F%2Fwww.thewindowsclub.com%2Fwhat-is-web-scraping&psig=AOvVaw1d15dZwJLXQ1LdXImz8WBr&ust=1570086749503092
- https://medium.com/swlh/exploring-sentiment-analysis-a6b53b026131
- http://nltk.sourceforge.net/doc/en/ch03.html