How Twitter Sentiments Affects Stock Market?

Aishwarya Kale, Sai Saran Akkirajju, Sarang Patil, Shikha C. Patel, Ujjwal

Under the guidance of **Dr. Tony Diana**
DATA 602 – Introduction to Data Analysis and Machine Learning

University of Maryland, Baltimore County

UMBC

Team Members

Sai Saran

Aishwarya

Ujjwal

Shikha

Sarang

What is a stock?

Stock

A form of security indicating owner's proportionate ownership

Raise Funds

Corporation's motives for selling stocks

✓ Historically outperformed other forms of investments over the long run

Stock Exchanges

Platform to sell and buy stocks

What is a stock market?

Stock Market

01 Venues

Buying, selling and issuance of shares

02 Free Market Economy

Democratized access to trading

03 Functions

Efficient price discovery and dealing

04 "Stock exchange"?

Subset of "stock market"

Public Sentiments and Stock Market

Crucial Aspect

Stock Market Stakeholders

Twitter

Rich Source

Real time public opinions

Stock Demand

Sentiments drives demand

Price Impact

Sentiments

Stock markets

Affected by public feelings

Stock Market is a complex environment

Large number of factors having direct or indirect effects

What we did?

Research aspects addressed:

01 Analyse correlation

Sentiment values & stock price attributes

02 Understand Impact

Twitter sentiments on stock prices

03 Duration of Tweets data

Same day, previous day, last 3 days

04 Forecasting Potential

Predicting next day's stock price

05 Tuning predictive models

Experimenting with hyperparameters

Dataset

Dataset Source: Kaggle 'Speculator and Influencer Evaluation in Stock Market by Using Social Media' - Mustafa et al. (2020)		
Company.csv	ticker symbols of the companies	$TSLA
Company_Tweet.csv	unique tweet id and the companies linked with tweet id	
Tweet.csv	different features of tweets. For example, tweet id, author, post date, tweet text, comments number, etc.	
CompanyValues.csv	stock market data of different companies. For example, open value, etc.	

UMBC

Dataset cont.

Stock Price	Description
open_value	Price at which the first stock is traded
close_value	Price at which the last stock is traded
low_value	Lowest price at which the stock is traded
high_value	Highest price at which the stock is traded

Table 1: elements of stock market prices

	ticker_symbol	company_name
0	AAPL	apple
1	GOOG	Google Inc
2	GOOGL	Google Inc
3	AMZN	Amazon.com
4	TSLA	Tesla Inc
5	MSFT	Microsoft

Figure 1: Ticker Symbols

Feature	Description
tweet_id	Unique tweet id of a tweet
writer	Username of the author
post_date	Date on which the tweet was posted (in form of seconds since epoch)
body	Text of the tweet
comment_num	Number of comments
retweet_num	Number of retweets
like_num	Number of thumb-up

Table 2: Attributes associated with each tweet

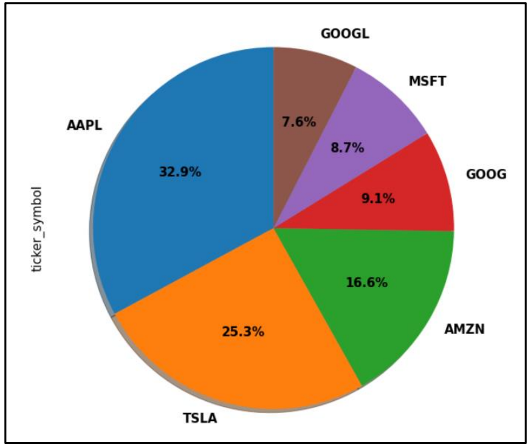
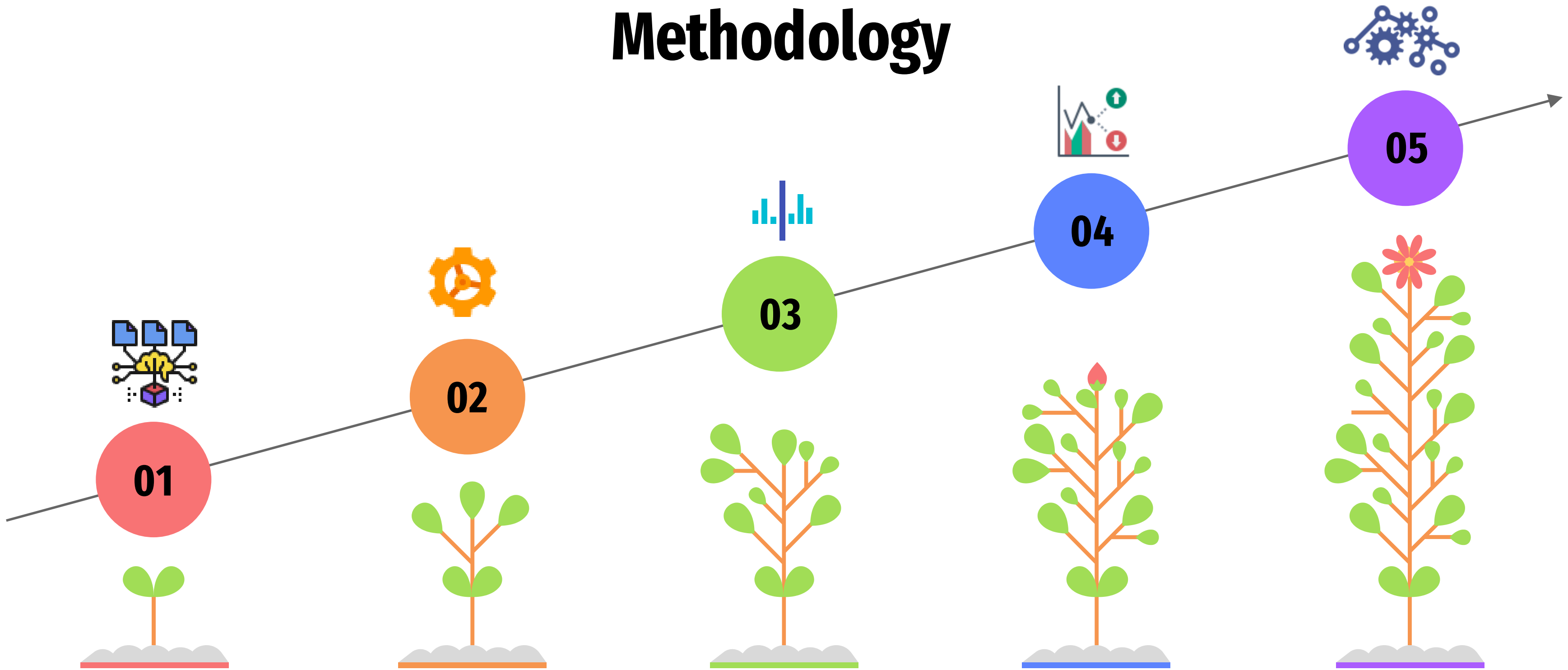


Figure 2: Tweets Distribution

Methodology



Dataset Preparation

Same day tweets, previous day tweets, last three days tweets

Sentiment Analysis & Feature Engineering

Positivity, negativity, neutrality, polarity, subjectivity, compound

Correlation

Spearman Correlation Tweets sentiments & Stock prices

Stock price prediction

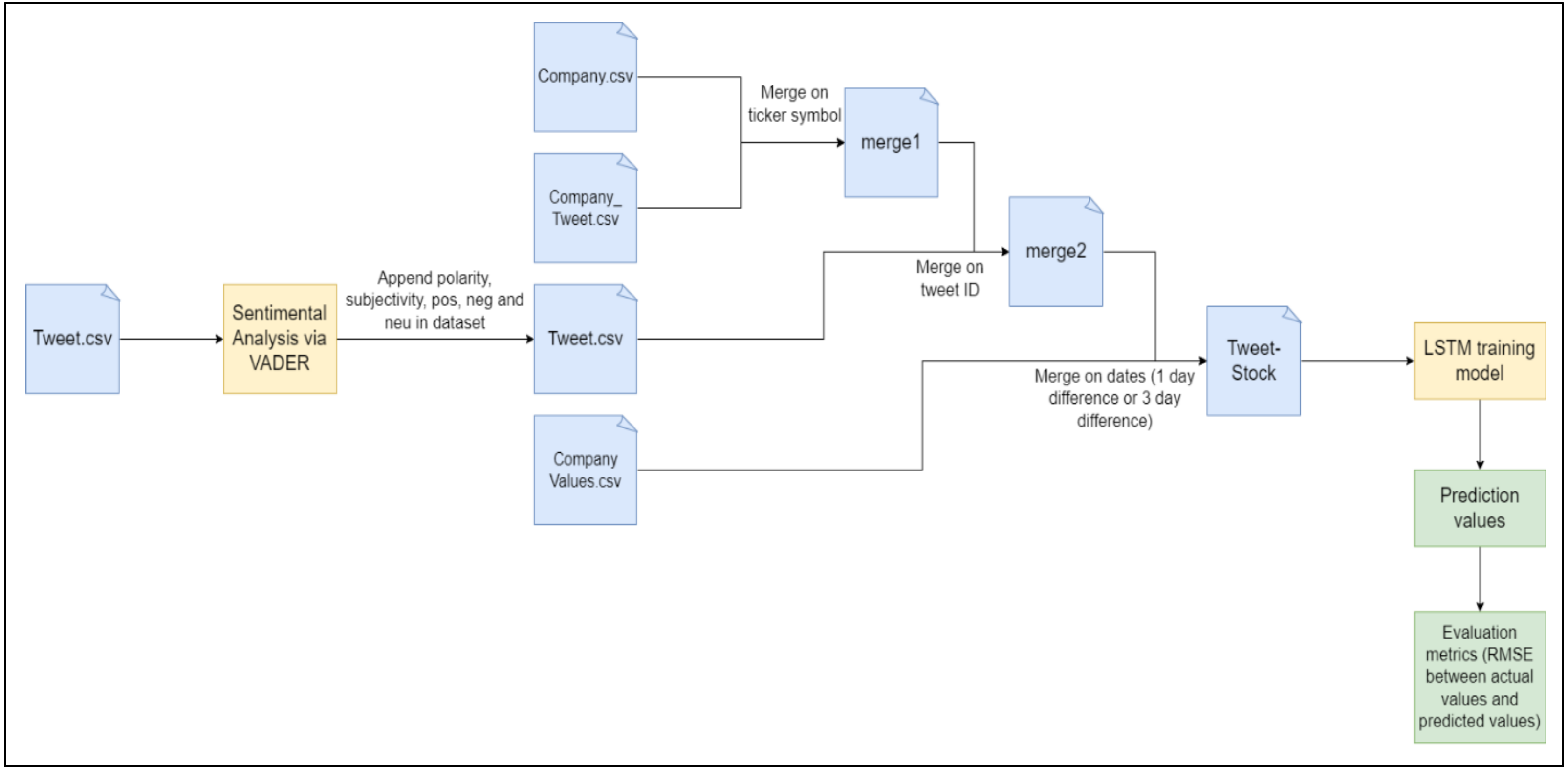
Multi-variate time series forecasting

Model Tuning

Observing effects of varying parameters

UMBC

Process Flow



UMBC

Sentiments

Subjectivity

Quantifies personal opinion & factural information

✓ Higher subjectivity means higher personal opinion based content and lesser factual information based content

Positivity

Percentage of positive content

✓ More the positivity means more positive words with higher positive evaluations

Negativity

Percentage of negative content

✓ More the negativity means more negaive words with higher negative evaluations

Correlations and Significance 

Values	Spearman Correlation Coefficient																	
	Apple						Amazon						Googl					
	same day			last 3 days			same day			last 3 days			same day			last 3 days		
	positivity	negativity	subjectivity	positivity	negativity	subjectivity	positivity	negativit	subjectivity	positivity	negativity	subjectivity	positivity	negativity	subjectivity	positivity	negativity	subjectivity
open_value	0.51	-0.33	0.51	0.54	-0.42	0.53	0.24	0.21	0.36	0.21	0.22	0.47	0.36	-0.03	0.47	0.39	-0.07	0.6
close_value	0.51	-0.34	0.51	0.54	-0.43	0.54	0.24	0.21	0.36	0.21	0.21	0.47	0.36	-0.03	0.47	0.39	-0.08	0.6
low_value	0.5	-0.34	0.51	0.54	-0.43	0.54	0.24	0.21	0.36	0.21	0.21	0.47	0.36	-0.03	0.47	0.39	-0.08	0.6
high_value	0.51	-0.33	0.51	0.54	-0.42	0.53	0.24	0.21	0.36	0.21	0.22	0.46	0.36	-0.02	0.46	0.39	-0.07	0.59

Table 3: Apple (AAPL), Amazon (AMZN) and Google (GOOGL) showing significant Spearman Correlation Coefficient

3 Companies

Significant correlation

Spearman correlation coefficient

95 %

Confidence level

Statistically significant

subjectivity

Highly correlated

Followed by positivity and negativity

Model Training

Stock Price Forecasting

Predict **next day's stock price** based on **previous day's stock** price combined with **Twitter sentiment values**

Model Selection

- ✓ Multi-variate time series forecasting
- ✓ ARIMA and SARIMAX can't be used
- ✓ LSTM is a suitable choice

Baseline LSTM

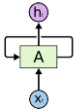
- ✓ Long-Short Term Memory
- ✓ 70% training data, 30% testing data
- ✓ Visible layer with 4 inputs, hidden layer with 50 LSTM blocks, output layer with single prediction value

Changing LSTM Configurations

- ✓ Impact of increasing LSTM blocks in hidden layer
- ✓ Impact of adding a dropout layer
- ✓ Total 6 configurations
- ✓ Total 144 models trained

UMBC

RNNs and LSTMs



RNN:
Recurrent Neural Networks

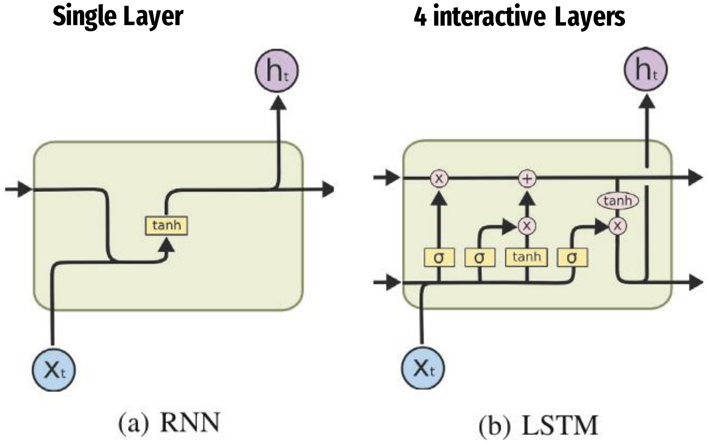
LSTM:
Long-Short Term Memory

A simple recurrent network suffers from a fundamental problem of not being able to capture long-term dependencies in a sequence.

When dealing with a time series, RNN tends to forget old information. Thus, resulting in Vanishing gradient problem.

Hence, when there is a distant relationship of unknown length, we wish to have a "memory" to it.

Figure 3:
Structure of an RNN Cell and an LSTM cell



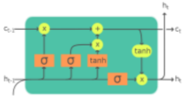
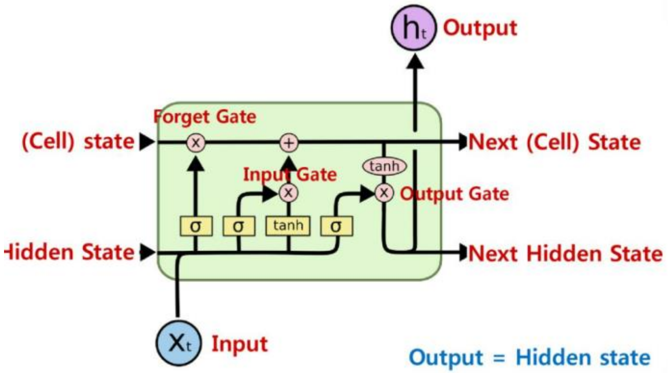
LTSM architecture 

Figure 4:
Components of an LSTM block



Long Short-Term Memory

An RNN architecture which addresses the problem of training over long sequences and retaining memory.

Cell state

A memory which helps LSTMs to selectively remember or forget things.

Gates

LSTM adds 3 gates to control cell states such as Forget Gate, Input gate, Output gate.

Mechanisms:

Forget Information
Add new information
Update cell state
Create output

LSTM configurations

Different LSTM configurations used	
LSTM	**Configuration**
LSTM – 1	50 LSTM blocks (hidden layer)
LSTM – 2	100 LSTM blocks (hidden layer)
LSTM – 3	200 LSTM blocks (hidden layer)
LSTM – 4	50 LSTM blocks (hidden layer), 1 Dropout Layer
LSTM – 5	100 LSTM blocks (hidden layer), 1 Dropout Layer
LSTM – 6	200 LSTM blocks (hidden layer), 1 Dropout Layer

```
Model: "sequential_4"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm_4 (LSTM)               (None, 50)                11000

 dense_4 (Dense)             (None, 1)                 51

=================================================================
Total params: 11,051
Trainable params: 11,051
Non-trainable params: 0
_________________________________________________________________
```

Figure 5:
Example of LSTM – 1 model summary

```
Model: "sequential_49"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 lstm_49 (LSTM)              (None, 50)                11000

 dropout (Dropout)           (None, 50)                0

 dense_48 (Dense)            (None, 1)                 51

=================================================================
Total params: 11,051
Trainable params: 11,051
Non-trainable params: 0
_________________________________________________________________
```

Figure 6:
Example of LSTM – 4 model summary

Results

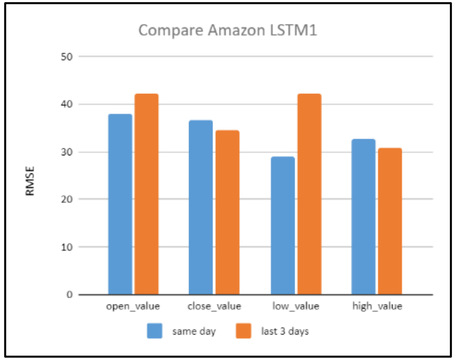


Figure 7: Baseline model comparison for Amazon

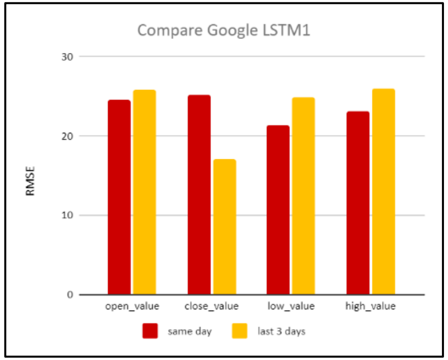


Figure 8: Baseline model comparison for Google

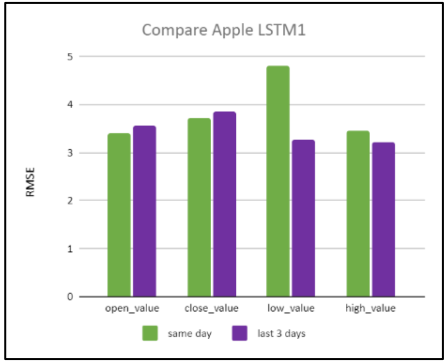


Figure 9: Baseline model comparison for Apple

RMSE below 5.0

Apple

(in the scale of 100 dollars)

Apple's current stock price:
$ 174.86 (9/Dec/2021)

RMSE below 40.0 & 30.0

Amazon & Google

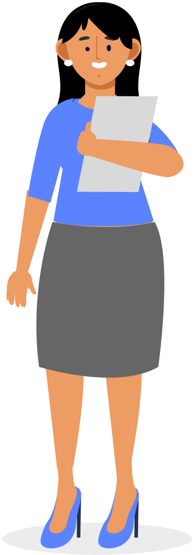
(in the scale of 1000 dollars)

Amazon's current stock price:
$ 3511.32 (9/Dec/2021)

Google's current stock price:
$ 2971.85 (9/Dec/2021)

UMBC

LSTM model comparison for Amazon



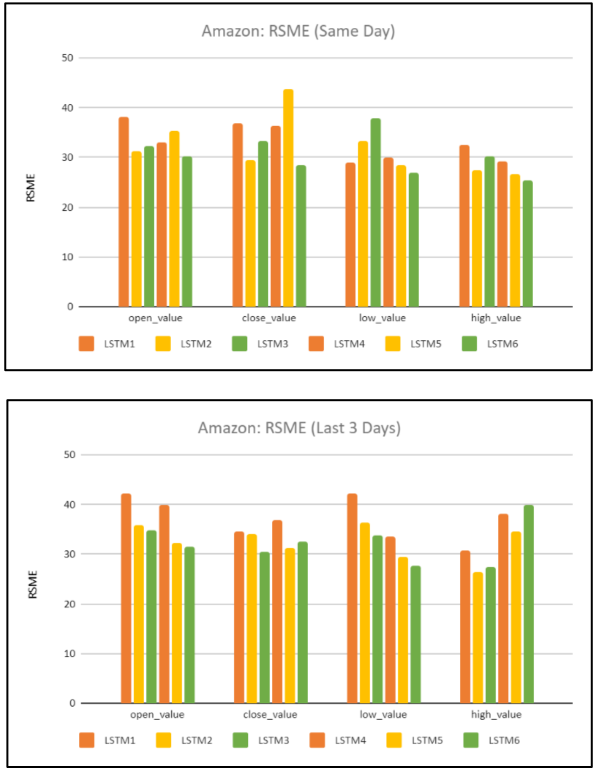


Figure 10: LSTM model results for Amazon

LSTM 6

Best performing

In most of the cases

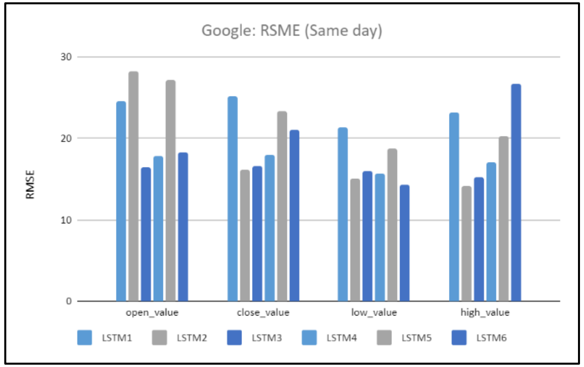
RMSE

Varies from 26.0 to 43.0

(in a scale of 1000 dollars)

UMBC

LSTM model comparison for Google



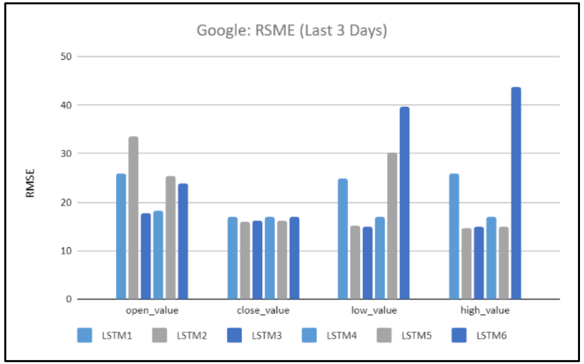


Figure 11: LSTM model results for Google

LSTM 3

Best performing

In most of the cases

RMSE

Varies from 14.0 to 43.0

(in a scale of 1000 dollars)

UMBC

LSTM model comparison for Apple

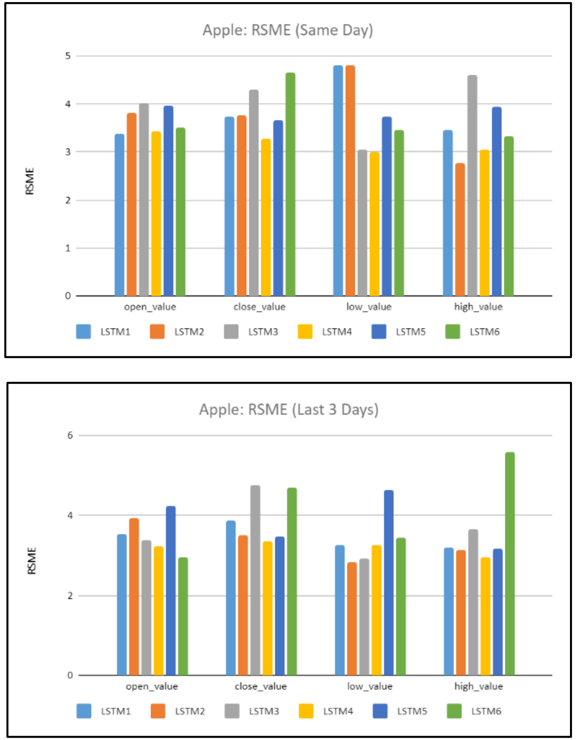


Figure 12: LSTM model comparison for Apple

LSTM 4

Best performing

In most cases

RMSE

Varies from 2.5 to 5.8

(in a scale of 100 dollars)

UMBC

Voila!!!

Conclusion & Future Scope

Conclusion 1
'positivity', 'negativity' and 'subjectivity' correlated with stock price movements

Conclusion 2
Higher the correlation between twitter sentiments and stock price, higher the predictive capability

Conclusion 3
Significant for investors, portfolio management companies, entrepreneurs and other stock market stakeholders

Opportunities in Future
Many other aspects can be researched and explored

Future Scope 1
The inverse relation can be investigated, i.e., stock price movements's impact on twitter sentiments

Future Scope 2
Different social media platforms and other sources like newspapers, Reddit can also be included

Future Scope 3
Number of companies and industries can be increased for exhaustive analysis

References

These are the research papers we referred for our project:

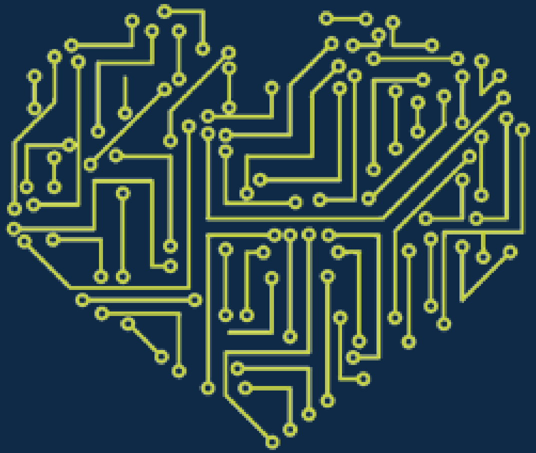
- Bollen, J., Mao, H., & Zeng, X. (2011, March). Twitter mood predicts the stock market. Journal of Computational Science, 2(1). doi:https://doi.org/10.1016/j.jocs.2010.12.007
- Dickinson, B., & Hu, W. (2015, January). Sentiment Analysis of Investor Opinions on Twitter. Social Networking, 04(03), 62-71. doi:10.4236/sn.2015.43008
- Doğan, M., Metin, Ö., Tek, E., Yumuşak, S., & Öztoprak, K. (2020). Speculator and Influencer Evaluation in Stock Market by Using Social Media. 2020 IEEE International Conference on Big Data (Big Data), (pp. 4559-4566). doi:10.1109/BigData50022.2020.9378170
- Hu, D., Jones, C., Zhang, V., & Zhang, X. (2021). The rise of reddit: How social media affects retail investors and short-sellers' roles in price discovery. SSRN Electronic Journal. doi:https://doi.org/10.2139/ssrn.3807655
- Ko, C.-R., & Chang, H.-T. (2021, March 11). LSTM-based sentiment analysis for stock price forecast. PeerJ Computer Science. doi:10.7717/peerj-cs.408
- Kordonis, J., Symeonidis, S., & Arampatzis, A. (2016). Stock Price Forecasting via Sentiment Analysis on Twitter. Proceedings of the 20th Pan-Hellenic Conference on Informatics.
- Mao, H., Counts, S., & Bollen, J. (2011). Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data. Papers 1112.1051, arXiv.org. Retrieved from https://arxiv.org/abs/1112.1051v1
- Mao, Y., Wei, W., Wang, B., & Benyuan, L. (2012). Correlating S&P 500 Stocks with Twitter Data. HotSocial '12: Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research. Association for Computing Machinery. doi:10.1145/2392622.2392634
- Nisar, T. M., & Yeung, M. (2018, June). Twitter as a tool for forecasting stock market movements: A short-window event study. The Journal of Finance and Data Science, 4(2), 101-119. doi:https://doi.org/10.1016/j.jfds.2017.11.002
- Pagolu, V. S., Challa, K., Panda, G., & Majhi, B. (2016). Sentiment Analysis of Twitter Data for Predicting Stock Market Movements. 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), (pp. 1345-1350). doi:DOI:10.1109/SCOPES.2016.7955659

UMBC

References

- Ruiz, E., Hristidis, V., Castillo, C., & Gionis, A. (2012). Correlating Financial Time Series with Micro-Blogging Activity. Proceedings of the Fifth International Conference on Web Search and Web Data Mining, WSDM 2012. Seattle, WA, USA. doi:10.1145/2124295.2124358
- Selvin, S., Ravi, V., Gopalakrishnan, E., & Menon, V. K. (2017). Stock price prediction using LSTM, RNN and CNN-sliding window model. 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI). doi:0.1109/ICACCI.2017.8126078
- Shah, D., Isah, H., & Zulkernine, F. (2019, May 27). Stock Market Analysis: A Review and Taxonomy of Prediction Techniques. International Journal of Financial Studies, 7(2). doi:https://doi.org/10.3390/ijfs7020026
- Siami-Namini, S., Tavakoli, N., & Siami Namin, A. (2019). A Comparison of ARIMA and LSTM in Forecasting Time Series. In M. A. Wani, M. Sayed-Mouchaweh, E. Lughofer, J. Gama, & M. Kantardzic (Eds.), Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018 (pp. 1394-1401). [8614252] Institute of Electrical and Electronics Engineers. doi:https://doi.org/10.1109/ICMLA.2018.00227
- Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. Journal of Finance, Forthcoming, 62(3). doi:http://dx.doi.org/10.2139/ssrn.685145
- Xia, Y., Liu, Y., & Chen, Z. (2013). Support Vector Regression for prediction of stock trend. 2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering (ICIII). doi:10.1109/ICIII.2013.6703098
- Yadav, A., Jha, C., & Sharan, A. (2020). Optimizing LSTM for time series prediction in Indian stock market. Procedia Computer Science, 167, 2091-2100. doi:https://doi.org/10.1016/j.procs.2020.03.257

Thank You



Sai Saran Akkiraju – s157@umbc.edu | Aishwarya Kale – uw76577@umbc.edu | Ujjwal – pf10610@umbc.edu | Shikha Patel – jf14891@umbc.edu | Sarang Patil – spatil4@umbc.edu

UMBC