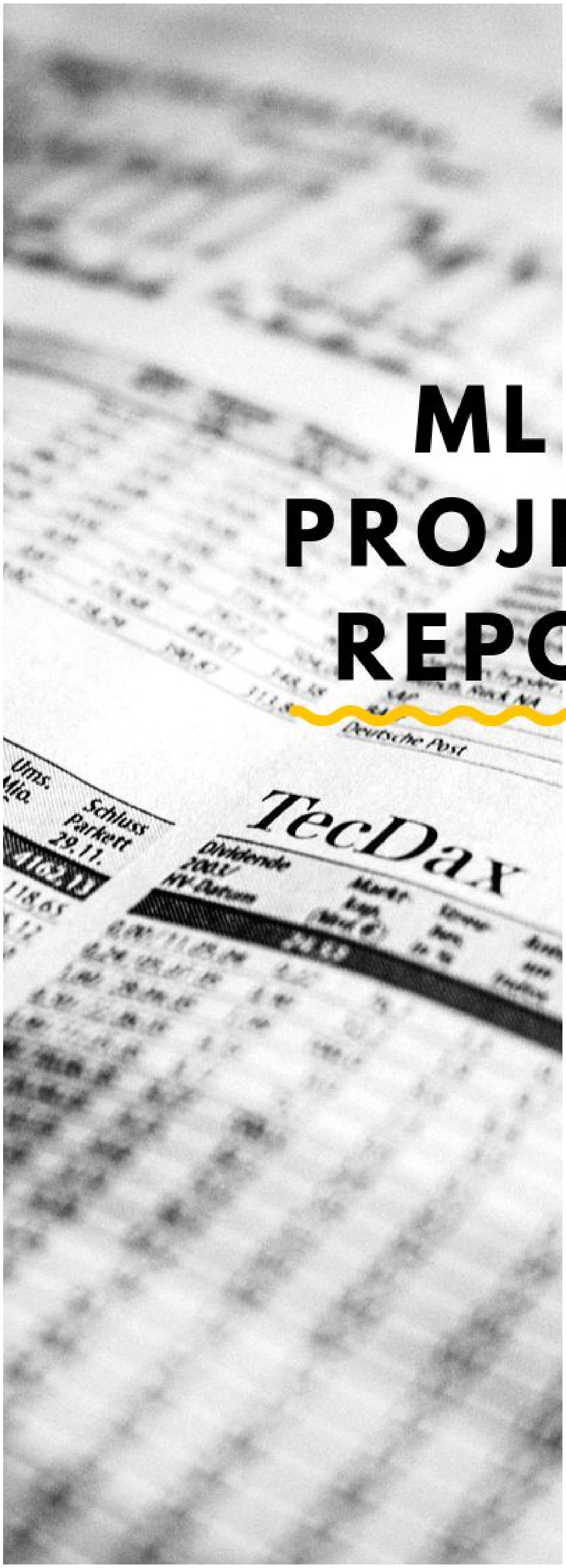
NUS FINTECH SOCIETY



ML G6 PROJECT REPORT

10 APR 2019

Done by:

Darren Lim
Leonard Tan
Qing Lin
Samuel Khoo

1. Introduction

1.1 Abstract

The objective of our project was to come out with a machine-learning model to predict the overall stock market performance using the top headlines from selected news sources as inputs.

1.2 Datasets

Our team will be using 3 csv files obtained from Kaggle:
https://www.kaggle.com/aaron7sun/stocknews

1. **RedditNews**: two columns The first column is the "date", and second column is the "news headlines". All news are ranked from top to bottom based on how hot they are. Hence, there are 25 lines for each date.
2. **DJIA_table**: Downloaded directly from Yahoo Finance: check out the web page for more info.
3. **Combined_News_DJIA**: To make things easier for my students, I provide this combined dataset with 27 columns. The first column is "Date", the second is "Label", and the following ones are news headlines ranging from "Top1" to "Top25".

The dates of the dataset: 2008/06/08 - 2016/07/01

1.3 Methodology

Our team conducted the following analysis on the data:

1. **Jaccard Similarity Coefficient:** To filter out relevant news from news that are not relevant to the stock market.
2. **Sentiment Analysis:** We used nltk.sentiment.vader library to help us perform our sentiment analysis.
3. **Long Short-Term Memory:** We decided to use LSTM as our machine-learning model in order to be able to take into account of the ‘context’ of the stock market.

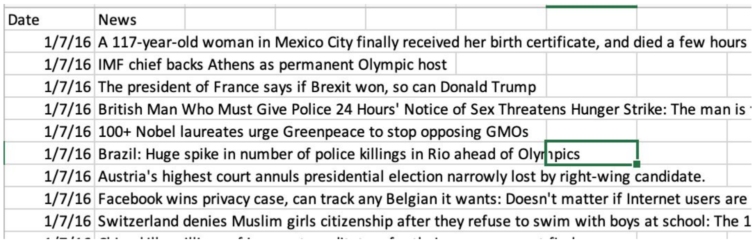
1.4 Content

Our paper will cover the following processes in this order:

1. Introduction
2. Data Cleaning and Preprocessing
3. Sentiment Analysis
4. Data Visualisation
5. Long-Short Term Memory Analysis and Results

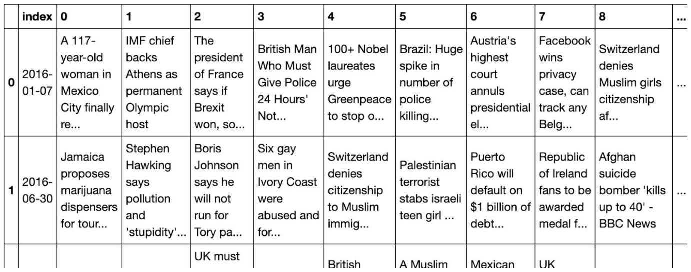
2. Data Cleaning and Preprocessing

The reddit news dataset is in the following format:



Date	News
1/7/16	A 117-year-old woman in Mexico City finally received her birth certificate, and died a few hours
1/7/16	IMF chief backs Athens as permanent Olympic host
1/7/16	The president of France says if Brexit won, so can Donald Trump
1/7/16	British Man Who Must Give Police 24 Hours' Notice of Sex Threatens Hunger Strike: The man is
1/7/16	100+ Nobel laureates urge Greenpeace to stop opposing GMOs
1/7/16	Brazil: Huge spike in number of police killings in Rio ahead of Olympics
1/7/16	Austria's highest court annuls presidential election narrowly lost by right-wing candidate.
1/7/16	Facebook wins privacy case, can track any Belgian it wants: Doesn't matter if Internet users are
1/7/16	Switzerland denies Muslim girls citizenship after they refuse to swim with boys at school: The 1

However, as we wanted to calculate the jaccard score and conduct sentiment analysis on the text dataset, we needed the data in the following format:



	index	0	1	2	3	4	5	6	7	8	...
0	2016-01-07	A 117-year-old woman in Mexico City finally re...	IMF chief backs Athens as permanent Olympic host	The president of France says if Brexit won, so...	British Man Who Must Give Police 24 Hours' Not...	100+ Nobel laureates urge Greenpeace to stop o...	Brazil: Huge spike in number of police killing...	Austria's highest court annuls presidential el...	Facebook wins privacy case, can track any Belg...	Switzerland denies Muslim girls citizenship af...	...
1	2016-06-30	Jamaica proposes marijuana dispensers for tour...	Stephen Hawking says pollution and 'stupidity'...	Boris Johnson says he will not run for Tory pa...	Six gay men in Ivory Coast were abused and for...	Switzerland denies citizenship to Muslim immig...	Palestinian terrorist stabs israeli teen girl ...	Puerto Rico will default on $1 billion of debt...	Republic of Ireland fans to be awarded medal f...	Afghan suicide bomber 'kills up to 40' - BBC News	...
				UK must		British	A Muslim	Mexican	UK		

Therefore, we needed to pivot the dataset into the dataframe we wanted. In order to complete this, we created a dictionary with each dates as keys and each headline as a value for that particular key. This is done by an iteration across the dataset and doing an ifelse to add each headline to a particular date.

Next, we made use of :

```
reddit_df = pd.DataFrame.from_dict(dic, orient ='index')
```

The pandas function to convert the dictionary into the dataframe we needed. Next, we did the following:

1. Conversion of all objects in dataframe into a String
2. Conversion of all strings into lower case.

We then continued on with the stemming process[1], to ensure our data would fit into the Jaccard score calculation that was planned.

[1] https://searchenterpriseai.techtarget.com/definition/stemming

3. Sentiment Analysis

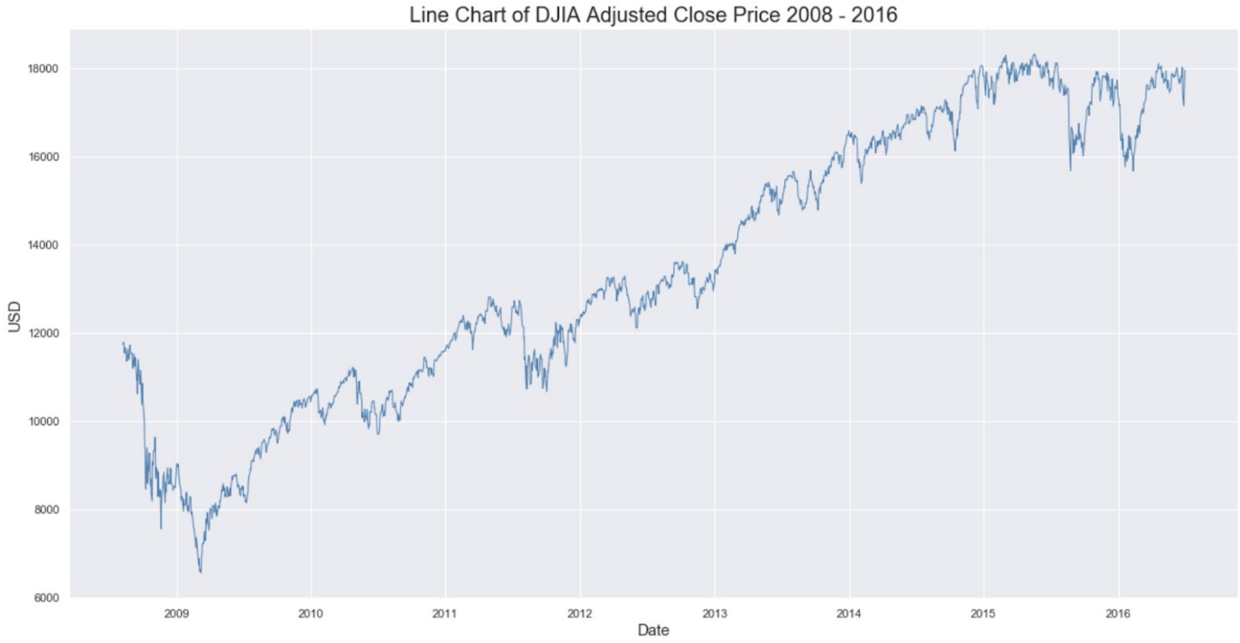
There are three steps required in order to give each day a ‘sentiment score’. Firstly, there is a need to calculate the relevance score across the 25 headlines for each day. Secondly, a weighted sentiment score is calculated for each date. Finally, an overall sentiment score for each day can then be obtained.

1. Calculation of relevance score for each day using Jaccard Similarity
2. Calculation of sentiment score for each headline, for each day
3. Multiplication of the two dataframes obtained from step 1 and 2
4. Calculation of the final weighted sentiment score, giving each day an overall score.

4. Data Visualisation

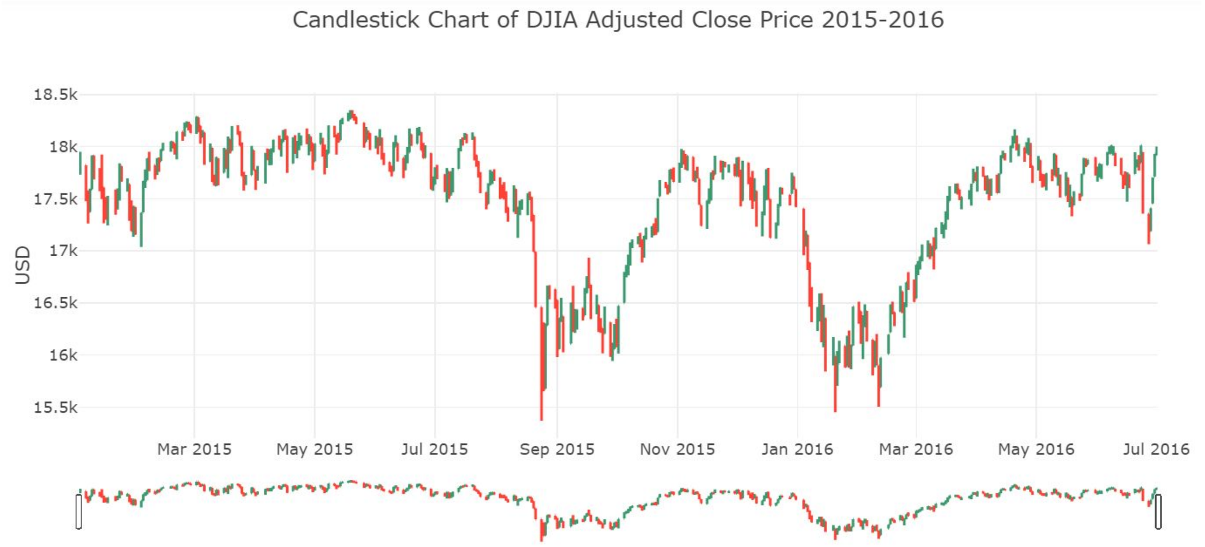
Using the DJIA Table, we have identified and extracted our main test data- the Daily Adjusted Close Values and formed a few visualizations based off our intuition and preliminary hypothesis using sns.

4.1 Adjusted Close Price Line Chart



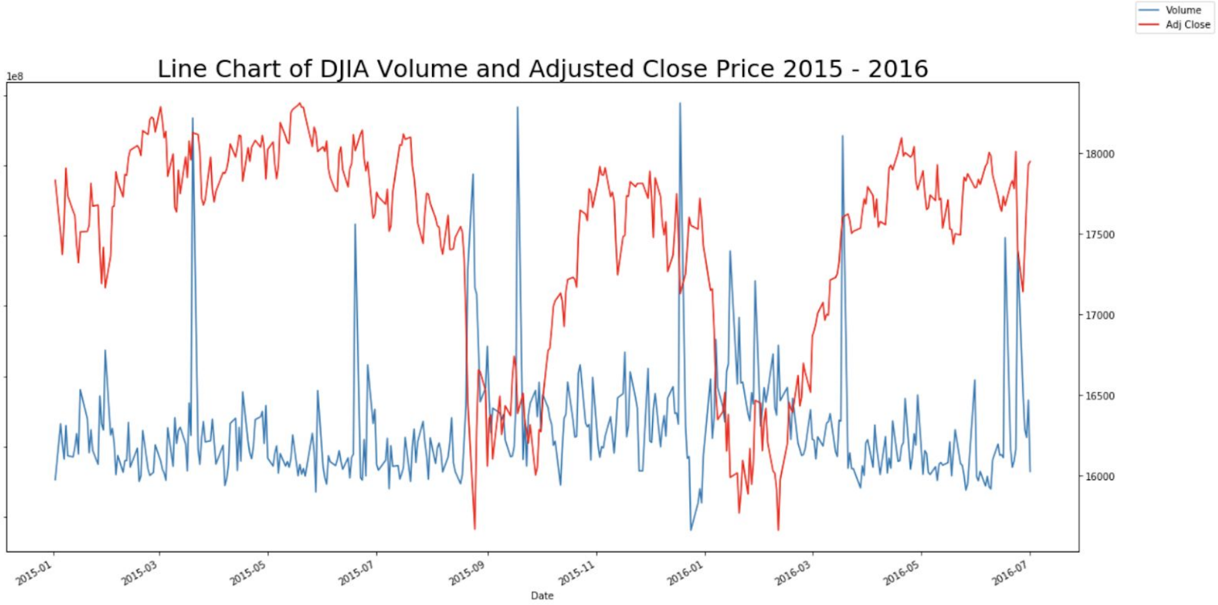
Despite a general upward trend from late 2009 to mid 2016, we were unable to draw any consistent trend.

4.2 Adjusted Close Price Candlestick Chart



We plotted a candlestick plot to take a closer look at the stock performance between May 2015 - Jul 2016. From this visualisation, it is evident that there were 2 significant economic downtain within this period. We can use these two fluctuations to help us see if there are other indicators that can help us predict these sharp falls.

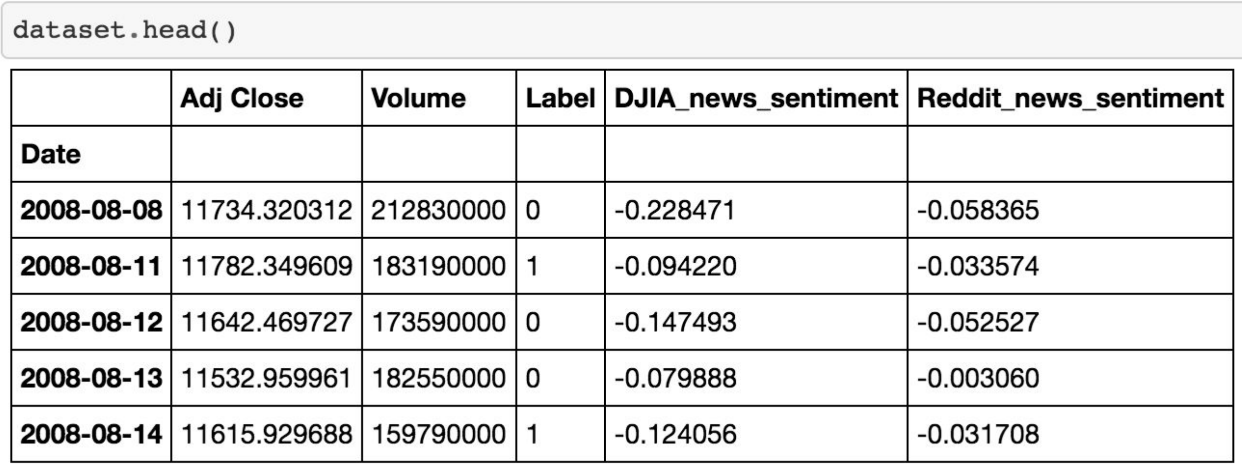
4.3 Volume vs Adjusted Close Price Line Chart



We noted a interesting correlation between volume of trades and great fluctuations in Adjusted Close price. This makes logical sense given widespread volume trades could be due to investor panic or immense bullish sentiment which leads to great daily volatility in open and close price changes.

4.4 Post-Processed Data Merging

Merging our post-processed data of Sentiment Analysis for Reddit news and DJIA news with our DJIA table, we investigate if there is indeed a correlation between the generic news sentiment with the stock price movement.



```
dataset.head()
```

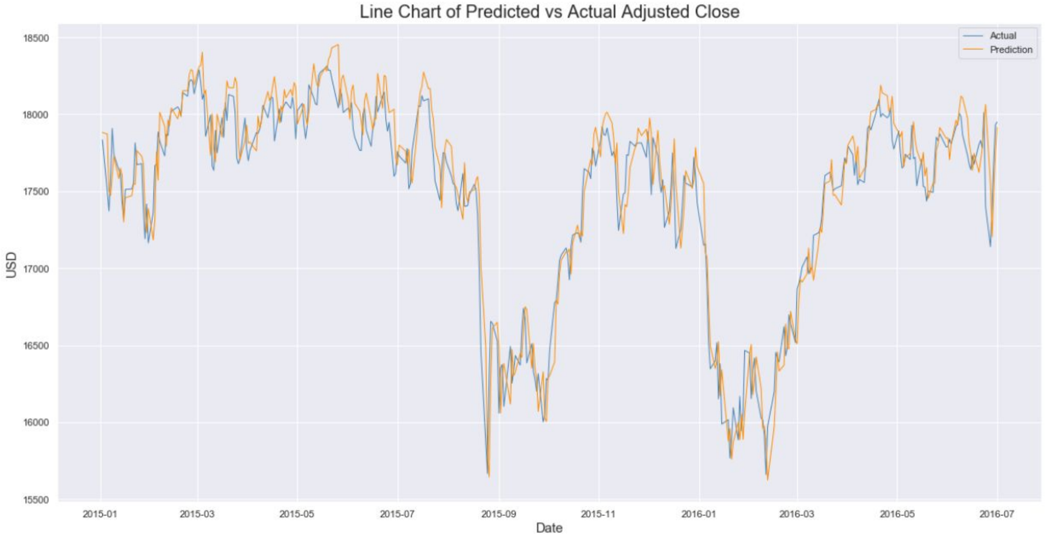
	Adj Close	Volume	Label	DJIA_news_sentiment	Reddit_news_sentiment
Date					
2008-08-08	11734.320312	212830000	0	-0.228471	-0.058365
2008-08-11	11782.349609	183190000	1	-0.094220	-0.033574
2008-08-12	11642.469727	173590000	0	-0.147493	-0.052527
2008-08-13	11532.959961	182550000	0	-0.079888	-0.003060
2008-08-14	11615.929688	159790000	1	-0.124056	-0.031708

5. Long-Short Term Memory Analysis and Results

5.1 Choice of Machine Learning Model - LSTM

For our machine learning model, we decided to use the Long-Short Term Memory network, which is a variant of the Recurrent Neural Network, which allows the model to retain an internal memory and to use it to make more precise decisions. LSTM is the preferred model for time series type data as it allows the computer to form a deeper understanding of a sequence and its context, compared to other algorithms. Our team hypothesized that as real life stock predictions are usually done in the context of the market, using LSTM to emulate that use of context should help us to come out with better predictions.

5.2 Results



After applying the machine learning model into our data, we managed to achieve the above prediction outcome. At first, it appears that our model has achieved a remarkably high accuracy in terms of prediction. However, we also noticed that quite a significant portion of the prediction happens only after the actual price has changed, which makes it less effective for predicting where the stock prices is going to go next.

This could be due to the inclusion of volume of trade as one of our predictors. On hindsight, while volume of trades is strongly correlated with changes in stock prices, it is more likely that the increase in trade volume typically happens in tandem with the rapid changes in stock prices, which prevents us from making advance predictions. However, the close resemblance of the two graphs could be an indication that LSTM was indeed a suitable model to use, and our sentiment analysis was not far off.

In conclusion, our team felt that there was a lack of leading indicators in this model, and a possible improvement would be to come out with better ways to interpret the economic context through the existing financial information, or to provide the algorithm with more leading indicators such as analyst reports to allow it to make more predictions before the actual market changes.

Annex

Important Links

Kaggle Page: https://www.kaggle.com/aaron7sun/stocknews/kernels

Link to our Code: https://github.com/samuelkhoo/NUS-Fintech-Stock-Prediction-Project