The National University of Ireland, Galway

MS5115 Business Analytics Major Project –

Final Report

Machine Learning and Graded Sentiment Analysis to Predict DJIA Directional Movement

Supervisor: Murray Scott

Submitted By Group 29

Student ID	Student Name
21230412	Yiyu Wang
21237523	Prashil Wanjari
21239610	Jiawen Zhang
21237514	Cong Wang

Date: 10 June 2022

Declaration

In submitting this work, we confirm that it is entirely our own. We acknowledge that we may be invited to an online interview if there is any concern in relation to the integrity of our submission, and we are aware that any breach will be subject to the University's Procedures for dealing with breaches of Exam Regulations. We are aware of what the NUI Galway plagiarism policy entails.

Machine Learning and Graded Sentiment Analysis to Predict DJIA Directional Movement

Table of Contents

1. Introduction

1.1 Background of DJIA and Reddit

The stock market is a marketplace where investors can buy and sell financial instruments like stocks and bonds. Stock market investments, however, are considered risky owing to their unpredictability and volatility. Stock prices are primarily affected by supply and demand. Stock prices can be boosted or depressed by an imbalance between supply and demand. There are so many factors that affect supply and demand, and stock price, for example, political situations, negotiations between countries, product breakthroughs, mergers and acquisitions, the performance of a company, industry trends, geographical location, and other unforeseen events, such as Covid-19 pandemic. According to (Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, Dan Jurafsky, 2014), daily news titles have a significant impact on stock price movements and the analysis of text sentiment of news titles can make better predictions than without it (JUSTINA DEVEIKYTE, HELYETTE GEMAN, CARLO PICCARI, ALESSANDRO PROVETTI, 2020).

These days, the stock market is heavily influenced by the overall social mood on the Internet, for instance, company news. Investors usually want to buy more stock in a company when there is positive news about the company, or a certain event happens within the company. In contrast, negative news will cause investors to sell their stocks. Individuals, groups of investors, companies and policy makers are increasingly finding it difficult to navigate the vast amount of information available online as the number of news and news sources increases. Markets and economies around the world are interconnected, so the news in one country can affect investors in another almost instantly. In addition to news about a particular company, such as the release of earnings reports, the price of the stock can be affected by company-specific news.

During our analysis, we chose news headlines from Reddit's sub-forum r/worldnews as one of our datasets because Reddit is the world's 9th most visited website, and the U.S.'s 6th most visited website, based on Semrush data as of March 2022. The other initial dataset is the Dow Jones Industrial Average (US stock market) index. The Dow Jones Industrial Average stock index is a price-weighted average of stock prices from 30 of the largest publicly-traded companies in the United States. Stocks in the DJIA are mature and stable relative to other stocks, making it one of the most closely watched indexes. Generally, most stocks do well when the Dow is rising, and it is considered bullish. A fall in the Dow suggests a bear market, and most stocks typically lose money. In this project, the Dow Jones Industrial Average is analyzed from a sentiment analysis perspective with world news headlines.

1.2 Related Work

It was one of the first papers to show a significant improvement in predictions of stock price based on textual information in (Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, Dan Jurafsky, 2014), "On the importance of text analysis for stock price prediction". More importantly, they illustrate the variation in the predictive power of text over time concerning a relevant event. In the short term, however, linguistic features have more predictive power compared to non-linguistic features, even while the time intervals are varied. This indicates that the effect of linguistic features diminishes quickly with time. However, they did not implement methods such as Neural Networks, Support Vector Machines, and Linear Discriminant Analysis for classifying the text data.

And as part of (Marc Velay, and Fabrice Daniel, 2018)'s paper "Using NLP on news headlines to predict index trends," they looked at a variety of techniques for predicting intraday movements in the DJIA, including logistic regression, support vector machines, and LSTM. Using logistic regression, they achieved the greatest accuracy, and with LSTM, they suffered severe overfitting and did not improve on randomness.

In (Adam Atkins, Mahesan Niranjan, Enrico Gerding, 2018)'s paper "Financial News Predicts Stock Market Volatility Better Than Close Price", they showed that volatility patterns are more predictable than asset price movements when financial news is used as machine learning input, thus, quantification of volatility could possibly be used in pricing derivatives contracts. As an alternative, we wonder if top global news headlines could also be used as data input, not just financial news.

1.3 Objective of Analysis

This project uses sentiment analysis and various machine learning algorithms to forecast intraday movements and market volatility of the DJIA, and to present these insights to investors. Our goal is to identify the best model for prediction to reach the highest level of accuracy. And the following research questions need to be answered:

a. Can textual content information, such as news headlines, be used in conjunction with machine learning to predict the Dow Jones Industrial Index?
b. Can sentiment scores derive from textual information help improve the accuracy of machine learning modelling?
c. How can we help stakeholders in equity investment and prove it from a technical and analytical point of view?

It is common for text to be interpreted in multiple ways, including complex words, sarcasm, ambiguities, difficult lexicons or grammar, and other situations that humans can solve subconsciously, but computers cannot, and so these obstacles have to be overcome in order for the computer to understand. A challenge associated with headlines is that they are often presented in a neutral manner, which makes it difficult to discern the underlying message. Additionally, different authors can interpret the same news from different perspectives and write different headlines. In other words, subjectivity and objectivity play an important role in understanding the meaning behind the news (V .V .Ramalingam, A. Pandian, Shivam Dwivedi, Jigar P.Bhatt, 2018).

We will then create the appropriate machine learning model using sentiment features and historical DJIA index price data. We may consider developing a mobile application or a website for investors or companies if the classification model proves promising. It will convert this vast amount of public information into key indicators of assets, including the overall temporality of key news headlines. Word count will be there to illustrate the sentiment of positive, negative and neutral on the stock market for better visualization and easy understanding. Additionally, the page will include an accurate estimate and a confidence rating.

2. Research Methodology

Cross-Industry Standard Process for Data Mining, or CRISP-DM, is a comprehensive data mining methodology and process model widely used in the data mining industry due to its multiple benefits and its accessibility, which allows anyone to conduct data mining projects, regardless of previous experience (Shearer, 2000). CRISP-DM illustrates six major phases of data mining, which will be discussed in further detail below. The most important thing you need to know is that CRISP-DM is not a linear process that starts with one step and moves neatly through each stage one by one. This means the analytical team can frequently move back and forth between the phases.

2.1 Business Understanding

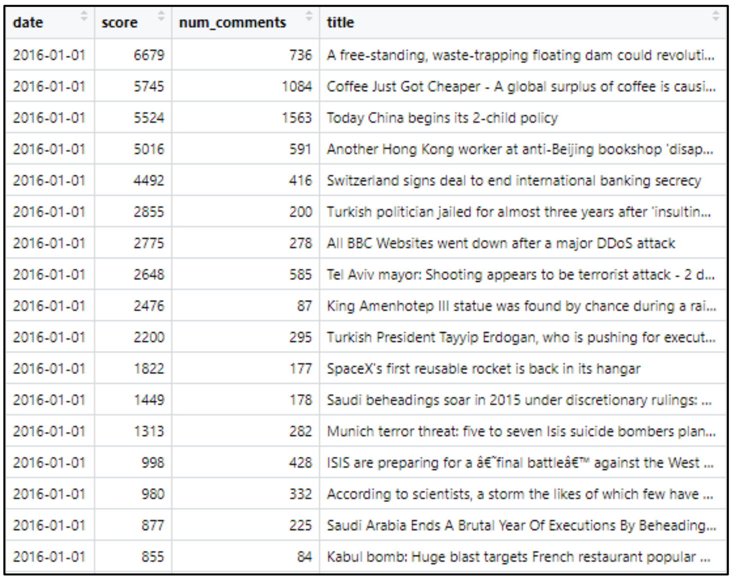
Many factors influence how stock prices are influenced by supply and demand, including political developments, trade agreements, technological breakthroughs, mergers and acquisitions, business performance, industry trends, geographical location, and other unforeseen events. Stock market performance is also influenced by how investors put their money to use. Likewise, stock markets provide useful information about a country's economy and corporate health. Therefore, they can serve as sources of capital for businesses. We would like to create an application that would help stakeholders predict the direction of the stock market and provide investment advice to reduce their investment risk. Textual web information found on news headlines, such as Reddit, could be valuable for predicting stock prices. Using news headlines, the sentiment of the public might be detected, which could greatly affect stock market movement. Stock closing prices and stock price movements are difficult to predict with high accuracy. However, we are interested in doing research and developing a model to see how NLP techniques and sentiment analysis contribute to the prediction of DJIA index volatility movements. Our analysis was based on historical news headlines from the Reddit r/worldnews and Dow Jones Industrial Average data from Yahoo Finance. It is the objective of this project to apply sentiment analysis and machine learning to identify the best model for forecasting intraday movements and market volatility ('up' or 'down') of the Dow Jones Industrial Average and to develop a deeper understanding of the market and present these insights to investors.

2.2 Data Understanding

2.2.1 Collect the Initial Data

News Data – Reddit World News Channel headlines are scraped using PRAW and PSAW to provide the textual data. Since there is a lot of data from 2011 to 2021, it will be easier to collect the data year by year by changing the index each year. Once we have scraped all the data, we use Python with Power BI to combine all the CSV files and get the top 25 headlines ranked by Reddit users' votes (see columns 'score' and 'num_comments' for each transaction date). Later, Excel can be used for easy data double-checking and data processing with DJIA index data and the combination of data. The Reddit scraping code and its related data processing can be found in Appendix A, while the following tables show the top 17 headlines as shown in the dataset on 1st January 2016.

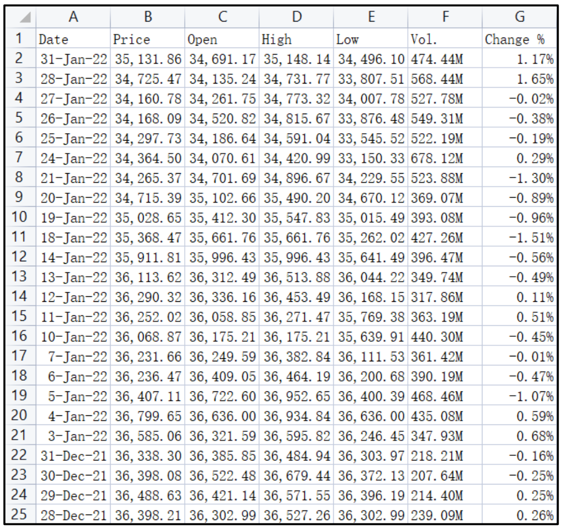
Table 1 Top 17 News Headlines



date	score	num_comments	title
2016-01-01	6679	736	A free-standing, waste-trapping floating dam could revoluti...
2016-01-01	5745	1084	Coffee Just Got Cheaper - A global surplus of coffee is causi...
2016-01-01	5524	1563	Today China begins its 2-child policy
2016-01-01	5016	591	Another Hong Kong worker at anti-Beijing bookshop 'disap...
2016-01-01	4492	416	Switzerland signs deal to end international banking secrecy
2016-01-01	2855	200	Turkish politician jailed for almost three years after 'insultin...
2016-01-01	2775	278	All BBC Websites went down after a major DDoS attack
2016-01-01	2648	585	Tel Aviv mayor: Shooting appears to be terrorist attack - 2 d...
2016-01-01	2476	87	King Amenhotep III statue was found by chance during a rai...
2016-01-01	2200	295	Turkish President Tayyip Erdogan, who is pushing for execut...
2016-01-01	1822	177	SpaceX's first reusable rocket is back in its hangar
2016-01-01	1449	178	Saudi beheadings soar in 2015 under discretionary rulings: ...
2016-01-01	1313	282	Munich terror threat: five to seven Isis suicide bombers plan...
2016-01-01	998	428	ISIS are preparing for a â€˜final battleâ€™ against the West ...
2016-01-01	980	332	According to scientists, a storm the likes of which few have ...
2016-01-01	877	225	Saudi Arabia Ends A Brutal Year Of Executions By Beheading...
2016-01-01	855	84	Kabul bomb: Huge blast targets French restaurant popular ...

Stock Data - The data includes the opening, high, low, volume and closing prices of the Dow Jones Industrial Average (DJIA) (see Table 2), which were obtained directly from (https://www.investing.com/indices/us-30-historical-data) due to time constraints. The Dow Jones Industrial Average measures the stock performance of 30 large companies listed on the New York Stock Exchange.

Table 2 DJIA index

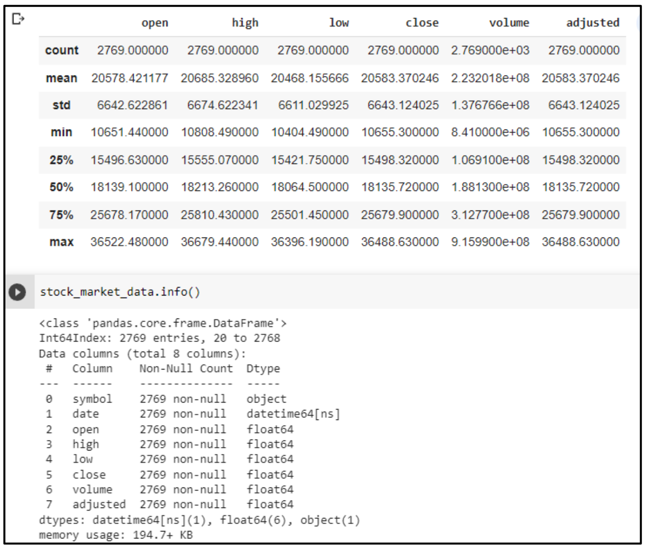


	A	B	C	D	E	F	G
1	Date	Price	Open	High	Low	Vol.	Change %
2	31-Jan-22	35, 131. 86	34, 691. 17	35, 148. 14	34, 496. 10	474. 44M	1. 17%
3	28-Jan-22	34, 725. 47	34, 135. 24	34, 731. 77	33, 807. 51	568. 44M	1. 65%
4	27-Jan-22	34, 160. 78	34, 261. 75	34, 773. 32	34, 007. 78	527. 78M	−0. 02%
5	26-Jan-22	34, 168. 09	34, 520. 82	34, 815. 67	33, 876. 48	549. 31M	−0. 38%
6	25-Jan-22	34, 297. 73	34, 186. 64	34, 591. 04	33, 545. 52	522. 19M	−0. 19%
7	24-Jan-22	34, 364. 50	34, 070. 61	34, 420. 99	33, 150. 33	678. 12M	0. 29%
8	21-Jan-22	34, 265. 37	34, 701. 69	34, 896. 67	34, 229. 55	523. 88M	−1. 30%
9	20-Jan-22	34, 715. 39	35, 102. 66	35, 490. 20	34, 670. 12	369. 07M	−0. 89%
10	19-Jan-22	35, 028. 65	35, 412. 30	35, 547. 83	35, 015. 49	393. 08M	−0. 96%
11	18-Jan-22	35, 368. 47	35, 661. 76	35, 661. 76	35, 262. 02	427. 26M	−1. 51%
12	14-Jan-22	35, 911. 81	35, 996. 43	35, 996. 43	35, 641. 49	396. 47M	−0. 56%
13	13-Jan-22	36, 113. 62	36, 312. 49	36, 513. 88	36, 044. 22	349. 74M	−0. 49%
14	12-Jan-22	36, 290. 32	36, 336. 16	36, 453. 49	36, 168. 15	317. 86M	0. 11%
15	11-Jan-22	36, 252. 02	36, 058. 85	36, 271. 47	35, 769. 38	363. 19M	0. 51%
16	10-Jan-22	36, 068. 87	36, 175. 21	36, 175. 21	35, 639. 91	440. 30M	−0. 45%
17	7-Jan-22	36, 231. 66	36, 249. 59	36, 382. 84	36, 111. 53	361. 42M	−0. 01%
18	6-Jan-22	36, 236. 47	36, 409. 05	36, 464. 19	36, 200. 68	390. 19M	−0. 47%
19	5-Jan-22	36, 407. 11	36, 722. 60	36, 952. 65	36, 400. 39	468. 46M	−1. 07%
20	4-Jan-22	36, 799. 65	36, 636. 00	36, 934. 84	36, 636. 00	435. 08M	0. 59%
21	3-Jan-22	36, 585. 06	36, 321. 59	36, 595. 82	36, 246. 45	347. 93M	0. 68%
22	31-Dec-21	36, 338. 30	36, 385. 85	36, 484. 94	36, 303. 97	218. 21M	−0. 16%
23	30-Dec-21	36, 398. 08	36, 522. 48	36, 679. 44	36, 372. 13	207. 64M	−0. 25%
24	29-Dec-21	36, 488. 63	36, 421. 14	36, 571. 55	36, 396. 19	214. 40M	0. 25%
25	28-Dec-21	36, 398. 21	36, 302. 99	36, 527. 26	36, 302. 99	239. 09M	0. 26%

2.2.2 Describe the Data

As part of this step, we will examine issues such as the format and quantity of the data, as well as the number of records and fields in each table. Additionally, we will consider the identity of each field and other surface characteristics. We can conclude from the tables below that the data acquired satisfies the requirements, such as no null values. As shown in Figure 19, Tableau shows the DJIA's price value over time.

Table 3 DJIA Index Descriptive Statistics

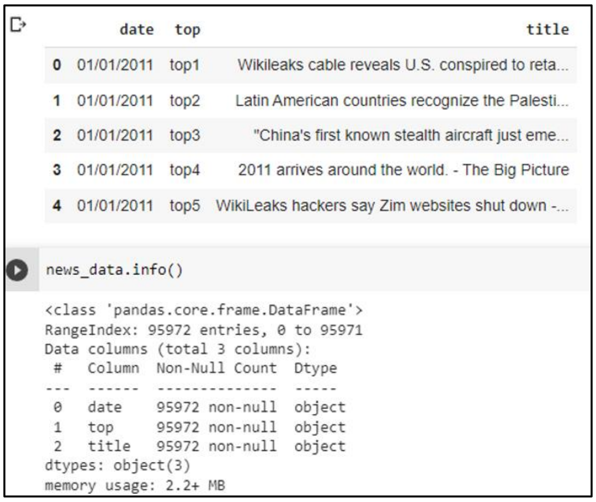


	open	high	low	close	volume	adjusted
count	2769.000000	2769.000000	2769.000000	2769.000000	2.769000e+03	2769.000000
mean	20578.421177	20685.328960	20468.155666	20583.370246	2.232018e+08	20583.370246
std	6642.622861	6674.622341	6611.029925	6643.124025	1.376766e+08	6643.124025
min	10651.440000	10808.490000	10404.490000	10655.300000	8.410000e+06	10655.300000
25%	15496.630000	15555.070000	15421.750000	15498.320000	1.069100e+08	15498.320000
50%	18139.100000	18213.260000	18064.500000	18135.720000	1.881300e+08	18135.720000
75%	25678.170000	25810.430000	25501.450000	25679.900000	3.127700e+08	25679.900000
max	36522.480000	36679.440000	36396.190000	36488.630000	9.159900e+08	36488.630000

```
stock_market_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2769 entries, 20 to 2768
Data columns (total 8 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   symbol    2769 non-null   object
 1   date      2769 non-null   datetime64[ns]
 2   open      2769 non-null   float64
 3   high      2769 non-null   float64
 4   low       2769 non-null   float64
 5   close     2769 non-null   float64
 6   volume    2769 non-null   float64
 7   adjusted  2769 non-null   float64
dtypes: datetime64[ns](1), float64(6), object(1)
memory usage: 194.7+ KB
```

Table 4 News Headlines Data



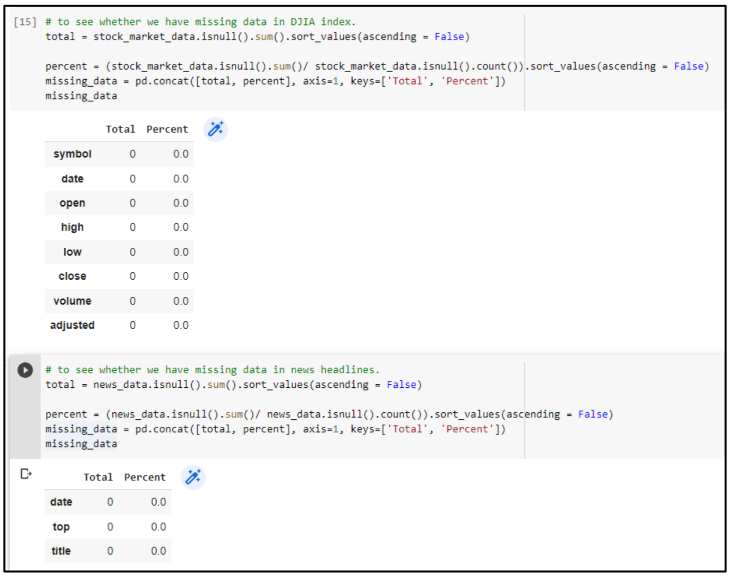
	date	top	title
0	01/01/2011	top1	Wikileaks cable reveals U.S. conspired to reta...
1	01/01/2011	top2	Latin American countries recognize the Palesti...
2	01/01/2011	top3	"China's first known stealth aircraft just eme...
3	01/01/2011	top4	2011 arrives around the world. - The Big Picture
4	01/01/2011	top5	WikiLeaks hackers say Zim websites shut down -...

```
news_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 95972 entries, 0 to 95971
Data columns (total 3 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   date    95972 non-null  object
 1   top     95972 non-null  object
 2   title   95972 non-null  object
dtypes: object(3)
memory usage: 2.2+ MB
```

2.2.3 Verify Data Quality

In our initial datasets, there are no blank fields, no missing attributes, and no missing values (see Figure 1).



```
[15] # to see whether we have missing data in DJIA index.
     total = stock_market_data.isnull().sum().sort_values(ascending = False)

     percent = (stock_market_data.isnull().sum()/ stock_market_data.isnull().count()).sort_values(ascending = False)
     missing_data = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
     missing_data
```

	Total	Percent
symbol	0	0.0
date	0	0.0
open	0	0.0
high	0	0.0
low	0	0.0
close	0	0.0
volume	0	0.0
adjusted	0	0.0

```
# to see whether we have missing data in news headlines.
total = news_data.isnull().sum().sort_values(ascending = False)

percent = (news_data.isnull().sum()/ news_data.isnull().count()).sort_values(ascending = False)
missing_data = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
missing_data
```

	Total	Percent
date	0	0.0
top	0	0.0
title	0	0.0

Figure 1 Check Missing Value

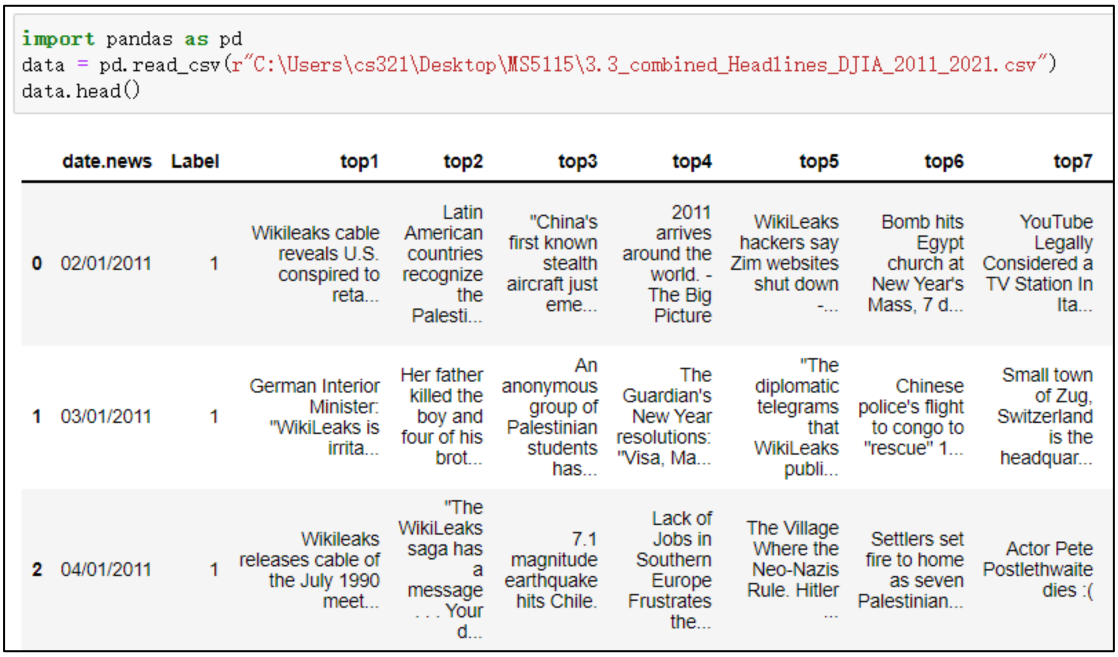
2.3 Data Preparation

2.3.1 Select Data

After collecting the initial news data, we need to use R to transform the data (see Table

5). Another data set we selected was the DJIA index data (see Table 2).

Table 5 Sample of the News Headlines Data after Transformation



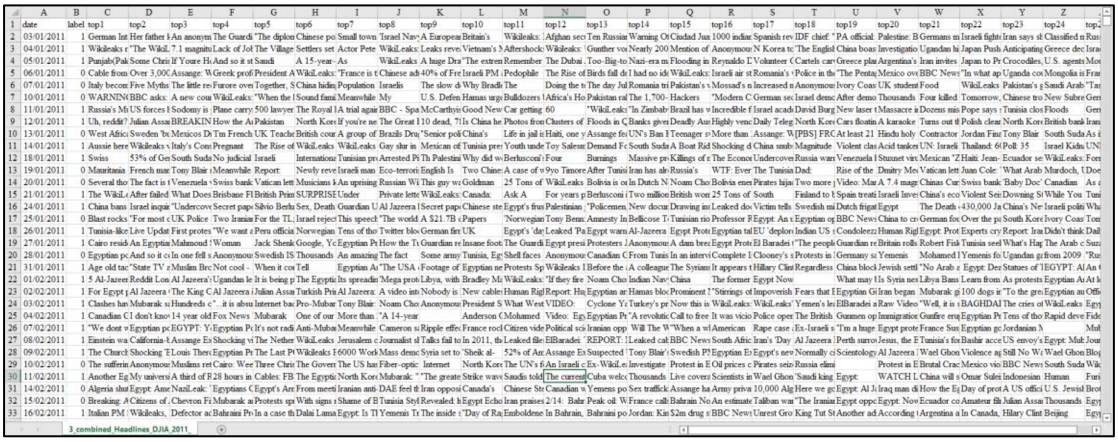
```
import pandas as pd
data = pd.read_csv(r"C:\Users\cs321\Desktop\MS5115\3.3_combined_Headlines_DJIA_2011_2021.csv")
data.head()
```

	date.news	Label	top1	top2	top3	top4	top5	top6	top7
0	02/01/2011	1	Wikileaks cable reveals U.S. conspired to reta...	Latin American countries recognize the Palesti...	"China's first known stealth aircraft just eme...	2011 arrives around the world. - The Big Picture	WikiLeaks hackers say Zim websites shut down -...	Bomb hits Egypt church at New Year's Mass, 7 d...	YouTube Legally Considered a TV Station In Ita...
1	03/01/2011	1	German Interior Minister: "WikiLeaks is irrita...	Her father killed the boy and four of his brot...	An anonymous group of Palestinian students has...	The Guardian's New Year resolutions: "Visa, Ma...	"The diplomatic telegrams that WikiLeaks publi...	Chinese police's flight to congo to "rescue" 1...	Small town of Zug, Switzerland is the headquar...
2	04/01/2011	1	Wikileaks releases cable of the July 1990 meet...	"The WikiLeaks saga has a message . . . Your d...	7.1 magnitude earthquake hits Chile.	Lack of Jobs in Southern Europe Frustrates the...	The Village Where the Neo-Nazis Rule. Hitler ...	Settlers set fire to home as seven Palestinian...	Actor Pete Postlethwaite dies :(

2.3.2 Construct Data

Combined Data Set – A combined dataset with 27 columns was created based on the news headlines and stock datasets. In the first two columns, we have the 'date' and the 'label', which serve as the ground truth label in training the classifier. In this study, we aim to predict the DJIA index volatility movements by categorizing them either as being 'up' or 'down'. Having done some research on the stock market price, we should be careful with the close prices while Adj Close is better as it accounts for corporate actions. In our case, we used financial time series data that could be further enhanced with derived features or trend-following indicators, which are technical tools that measure the direction and strength of trends within a given time frame. Having a value of '1' in the column 'label' signifies that the adjusted close price that day was higher than the day before or stayed the same. A value of '0' indicates that the adjusted close price declined during the day as compared to the previous day. We used yesterday's news headlines, for example, to predict the "up" or "down" for today. Therefore, the date in the combined dataset (see Table 6) will be plus one based on the date of the news headlines posted using Excel's function.

Table 6 Combined Data Set



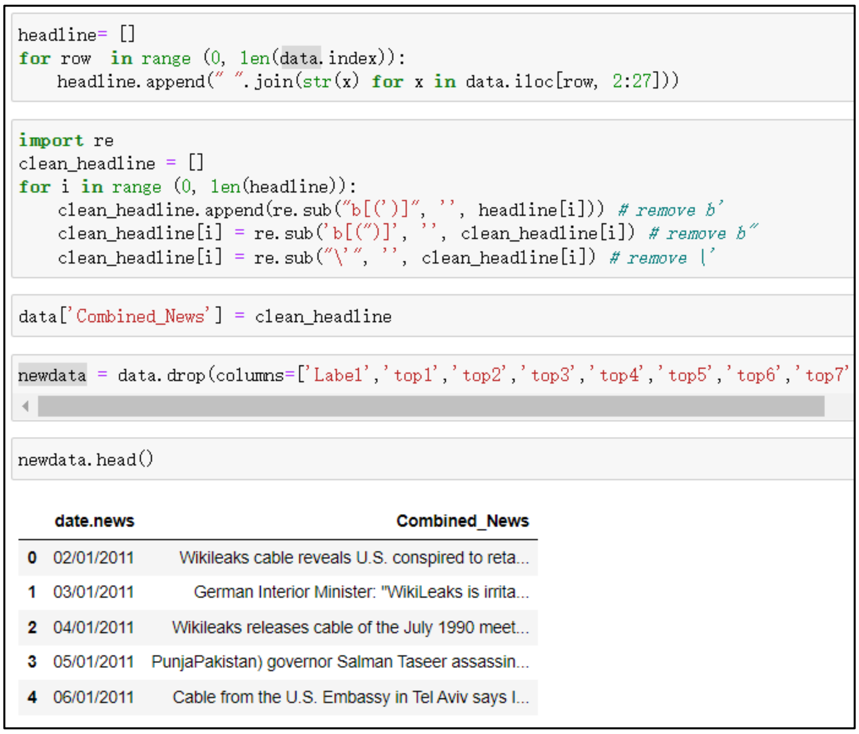
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	
1	date	label	top1	top2	top3	top4	top5	top6	top7	top8	top9	top10	top11	top12	top13	top14	top15	top16	top17	top18	top19	top20	top21	top22	top23	top24	top2
2	03/01/2011	1	German Int	Her father I	An anonym	The Guardi	"The diplon	Chinese po	Small town	'Israel Nav	A Europea	Britain's	Wikileaks:	Afghan sec	Ten Russia	Warning O	Ciudad Jua	1000 india	Spanish rev	IDF chief: "	PA official:	Palestine: B	Germans m	Israeli fight	Iran says sh	Classified n	Rus
3	04/01/2011	1	Wikileaks r	"The WikiL	7.1 magnitu	Lack of Jol	The Village	Settlers set	Actor Pete	WikiLeaks:	Leaks reve	Vietnam's N	Aftershock:	Wikileaks:	Gunther vo	Nearly 200	Mention of	Anonymous	N Korea tc	"The English	China boas	Investigatio	Ugandan hi	Japan Push	Anticipating	Greece dec	Isra
4	05/01/2011	1	Punjab(Pak	Some Chris	If Youre H	And so it st	Saudi	A 15-year-	As	WikiLeaks	A huge Dra	"The extren	Remember	The Dubai	Too-Big-to	Nazi-era m	Flooding in	Reynaldo D	Volunteer (	Cartels car	Greece plar	Argentina's	Iran invites	Japan to Pr	Crocodiles,	U.S. agents	Mo
5	06/01/2011	0	Cable from	Over 3,000	Assange: W	Greek profi	President A	WikiLeaks:	"France is t	Chinese ad	40% of Fre	Israeli PM	Pedophile	The Rise of	Birds fall d	I had no ide	WikiLeaks:	Israeli air st	Romania's	Police in th	"The Penta	Mexico ove	BBC News	"In what ap	Uganda co	Mongolia is	Fra
6	07/01/2011	0	Italy becom	Five Myths	The little re	Furore over	Together, S	China hidin	Population	Israelis	The slow d	Why Bradle	The	Doing the t	The day Jul	Romania tri	Pakistan's v	Mossad's n	Increased n	Anonymous	Ivory Coas	UK student	Food	WikiLeaks	Pakistan's g	Saudi Arab	"Ta
7	10/01/2011	0	WARNING	BBC asks:	A new cou	WikiLeaks:	"When the I	Sound fami	Meanwhile	My	U.S. Defen	Hamas urge	Bulldozers I	Africa's Ho	Pakistan ral	The 1,700-	Hackers	"Modern C	German sec	Israel demo	After demo	Thousands	Four killed	Tomorrow,	Chinese tro	New Subre	Ger
8	11/01/2011	1	Russia's Mi	US forces I	Sodomy is	Plane carry	500 lawyer	The Royal I	A trial agai	BBC - Spa	McCarthyis	Good New	Car getting	60	"WikiLeaks	"In Zimbab	Brazil has w	Incredible f	Israel acad	David Burg	New laser t	Massacre i	Dozens mis	Pope says	Tunisia clos	Floods	Ger
9	12/01/2011	1	Uh, reddit?	Julian Assa	BREAKIN	How the A	Pakistan	North Kore	If you're ne	The Great I	10 dead, 7	Is China he	Photos fron	Clusters of	Floods in Q	Banks give	Deadly Aus	Highly venc	Daily Teleg	North Kore	Cars floatin	A karaoke	Turns out th	Polish clear	North Kore	British bank	Iran
10	13/01/2011	0	West Afric	Sweden 'bu	Mexicos D	Tm French	UK Teache	British cour	A group of	Brazils Dru	"Senior poli	China's	Life in jail is	Haiti, one y	Assange fe	UN's Ban I	Teenager s	More than	Assange: W	[PBS] FRC	At least 21	Hindu holy	Contractor	Jordan Fina	Tony Blair	South Suda	As i
11	14/01/2011	1	Aussie here	Wikileaks v	Italy's Con	Pregnant	The Rise of	WikiLeaks	WikiLeaks	Gay slur in	Mexican of	Tunisia pre	Youth unde	Toy Salesm	Demand Fo	South Suda	A Boat Rid	Shocking d	China snub	Magnitude	Violent clas	Acid tanke	UN: Israeli	Thailand: 6	Poll: 35	Israel Kidn	UN
12	18/01/2011	1	Swiss	53% of Ge	South Suda	No judicial	Israeli	Internation	Tunisian pr	Arrested P	Th Palestini	Why did w	Berlusconi's	Four	Burnings	Massive pr	Killings of r	The Econo	Undercover	Russia war	Venezuela I	Stuxnet vir	Mexican "Z	Haiti: Jean-	Ecuador se	WikiLeaks:	For
13	19/01/2011	0	Mauritania	French mar	Tony Blair	Meanwhile	Report:	Newly reve	Israeli man	Eco-terrori	English Is	Two Chine	A case of w	9yo Timore	After Tunisi	Iran has alr	Russia's	WTF: Ever	The Tunisia	Dad:	Rise of the	Dmitry Me	Vatican lett	Juan Cole:	'What Arab	Murdoch, U	Doe
14	20/01/2011	0	Several tho	The fact is t	Venezuela	Swiss bank	Vatican lett	Musicians k	An uprising	Russian Wi	This guy wa	Goldman	25 Tons of	WikiLeaks	Bolivia is o	In Dutch N	Noam Cho	Bolivia ene	Pirates hija	Two more	Video: Mar	A 7.4 magn	Chinas Cur	Swiss bank	'Baby Doc'	Canadian	As a
15	21/01/2011	1	The WikiL	After failed	What Does	Brisbane Fl	British Prim	SURPRISE	Under	Private lette	WikiLeaks:	Canada:	Ask A	For years p	Berlusconi	Two million	British won	25 Tons of	South	Finland to b	Spain treati	Israeli Inve	China's eco	Violent Sei	Downing St	While You	Tuni
16	24/01/2011	1	China bans	Israel inquir	"Undercove	Secret pape	Silvio Berlu	Sex, Death	Guardian U	Al Jazeera I	Secret pape	Chinese ste	Egypt's frus	Palestinian	"Policemen,	New docur	Drawing ins	Leaked doc	Victim tells	Swedish mi	Dutch frigat	Egypt	The Death	430,000 Ja	China's Ne	Israeli politi	Wha
17	25/01/2011	0	Blast rocks	"For most c	UK Police	Two Irania	For the TL;	Israel reject	This speech	"The world	A $21.7B c	Papers	'Norwegian	Tony Benn:	Amnesty In	Bellicose T	Tunisian rio	Professor R	Egypt: An u	Egyptian op	BBC News	China to cr	German fou	Over the pa	South Kore	Ivory Coas	Ton
18	26/01/2011	1	Tunisia-like	Live Updat	First protes	"We want a	Peru officia	Norwegian	Tens of tho	Twitter blo	German fir	UK	Egypt's 'da	Leaked 'Pa	Egypt warn	Al-Jazeera	Egypt Prot	Egyptian tal	EU 'deplor	Indian US s	Condoleez	Human Rig	Egypt: Prot	Experts cry	Report: Ira	Didn't think	Dail
19	27/01/2011	1	Cairo resid	An Egyptia	Mahmoud	Woman	Jack Shenk	Google, Yo	Egyptian Pr	How the Tu	Guardian re	Insane foot	The Guardi	Egypt presi	Protesters J	Anonymous	A dam brea	Egypt Prot	El Baradei	"The peopl	Guardian re	Britain rolls	Robert Fisk	Tunisia seel	What's Hap	The Arab c	Suz
20	28/01/2011	0	Egyptian pc	And so it c	In one fell s	Anonymous	Swedish IS	Thousands	An amazing	The fact	Some army	Tunisia, Eg	Shell faces	Anonymous	Canadian G	From Tunis	In an intervi	Complete I	Clooney's s	Protests in	Germany sa	Yemenis	Mohamed I	Yemenis fo	Ugandan g	from 2009	"Rus
21	31/01/2011	1	Age old tac	"State TV a	Muslim Bro	Not cool -	When it co	Tell	Egyptian A	"The USA	Footage of	Egyptian ne	Protests Sp	Wikileaks I	Before the	A colleague	The Syrians	It appears t	Hillary Clin	Regardless	China block	Jewish settl	"No Arab a	Egypt: Dea	Statues of T	EGYPT: Al	An (
22	01/02/2011	1	5 Al-Jazeer	Reddit Lon	Al Jazeera's	Ugandan le	It is being p	The Egyptia	Its spreadin	'Mega prot	Libya, with	Bradley Ma	WikiLeaks:	"If they fire	Noam Cho	Indian Nav	China	The former	Egypt Now		What may t	Is Syria nex	Libya Bans	Learn from	As protests	Egyptian A	At l
23	02/02/2011	1	For Egypt	Al Jazeera	The King C	Al Jazeera	Julian Assa	Turkish Pri	Al Jazeera:	A video int	Nobody is	New cable	Human Rig	Report: Hu	Egyptian ar	Hamas blo	Prominent	"Stirrings of	Impoverish	Fears that E	Egyptian Gi	Iran began	Mubarak g	100 dogs in	"To the gre	Egyptian au	Offi
24	03/02/2011	1	Clashes hav	Mubarak sa	Hundreds c	"...it is absu	Internet bac	Pro-Mubar	Tony Blair:	Noam Cho	Anonymous	President S	What West	VIDEO:	Cyclone Ya	Turkey's pr	Now this is	WikiLeaks:	WikiLeaks'	Yemen's le	ElBaradei a	Raw Video	"Well, it is	BAGHDAI	The cries o	WikiLeaks	Egy
25	04/02/2011	1	Canadian C	I don't kno	14 year old	Fox News	Mubarak	One of our	More than	"A 14-year		Anderson C	Mohamed	Video: Eg	Egyptian Pr	"A revolutic	Call to free	It was vicio	Police oper	The British	Gunmen op	Immigratio	Gunfire eru	Egyptian Pr	Tens of tho	Rapid deve	Fide
26	07/02/2011	1	"We dont w	Egyptian pc	EGYPT: Y	Egyptian P	It's not radi	Anti-Mubar	Meanwhile	Cameron sa	Ripple effe	France roc	Citizen vide	Political sci	Iranian opp	Will The W	"When a w	American	Rape case	Ex-Israeli s	"I'm a huge	Egypt prote	France Sus	Egyptian go	Jordanian M		Mub
27	08/02/2011	1	Einstein wa	California-t	Assange Ex	Shocking vi	The Nether	WikiLeaks	Jerusalem c	Journalist sl	Talks fail to	In 2011, th	Leaked file	ElBaradei:	"REPORT:	Leaked cab	BBC News	South Afric	Iran's 'Day	Al Jazeera	Perth surro	Jesus, the E	Tunisia's fo	Bashir acce	US envoy's	Egypt: Mub	Jour
28	09/02/2011	1	The Church	Shocking 'E	Louis Ther	Egyptian Pr	The Last P	Wikileaks I	6000 Work	Mass demo	Syria set to	'Sheik al-	52% of Am	Assange Ex	Suspected	Tony Blair's	Swedish PM	Egyptian Ex	Egypt's nev	Normally c	Scientology	Al Jazeera	Wael Ghon	Violence ag	Still No Wa	Wael Ghon	Blog
29	10/02/2011	0	The sufferin	Anonymous	Muslims ret	Cairo: Wee	Three Chris	The Goverr	The US has	Fiber-optic	Internet	North Kore	The UN's f	An Israeli c	Ex-WikiLe	Investigate	Protest in E	Oil prices c	Pirates seiz	Russia elimi		Protest in E	Brutal Crac	Mexico vio	BBC News	South Suda	Wik
30	11/02/2011	1	Another Eg	My univers	A third of F	28 hours in	Cables: FB	The Egyptia	North Kore	Mubarak: "	"The greate	Strike wave	Saudis told	The current	Cuba welco	Thousands	Live covera	Scientists in	Wael Ghon	'Saudi king	Egypt:	WATCH L	China will s	Omar Sulei	Indonesian	Human	Furi
31	14/02/2011	0	Algeria shut	Egypt: Ame	NaziLeak:	Egyptians C	Egypt's Arr	From meeti	Iranian anti-	DAE feel th	Iran opposi	Canada's	Chinese Str	Canadian w	Yemens po	Sex traffick	Assange ha	Army priva	10,000 Alg	Here we go	Egypt: Al J	Iraq man di	How the Eg	Day of prot	A US offici	U.S. Jewish	Brot
32	15/02/2011	0	Breaking: A	Citizens of	Chevron Fi	Mubarak a	Protests spr	With signs	Shame of B	Tunisia Styl	Revealed: h	Egypt Echo	Iran praises	2/14: Bahr	Peak oil: W	France call	Bahrain No	An estimate	Taliban war	"The Iraniar	Egypt oppo	Egypt: Nov	Ecuador co	Amateur fil	Julian Assa	Thousands	Egy
33	16/02/2011	1	Italian PM	Wikileaks,	Defector ac	Bahraini Pr	In a case th	Dalai Lama	Egypt: Is Th	Yemenis Tr	The inside s	"Day of Ra	Emboldene	In Bahrain,	Bahraini po	Jordan: Kin	$2m drug s	BBC News	Unrest Gro	King Tut St	Another ad	According t	Argentina a	In Canada,	Hilary Clint	Beijing	Egy

3_combined_Headlines_DJIA_2011_

2.3.3 Integrate Data

We need information from two tables and records to construct new records and values before we can do sentiment analysis and build machine learning models. The merging of tables will be performed in a Python environment (see Figure 2).

Figure 2 Integrate Data



```
headline= []
for row in range (0, len(data.index)):
    headline.append(" ".join(str(x) for x in data.iloc[row, 2:27]))
```

```
import re
clean_headline = []
for i in range (0, len(headline)):
    clean_headline.append(re.sub("b[(')]", '', headline[i])) # remove b'
    clean_headline[i] = re.sub('b[(")]', '', clean_headline[i]) # remove b"
    clean_headline[i] = re.sub("\'", '', clean_headline[i]) # remove \'
```

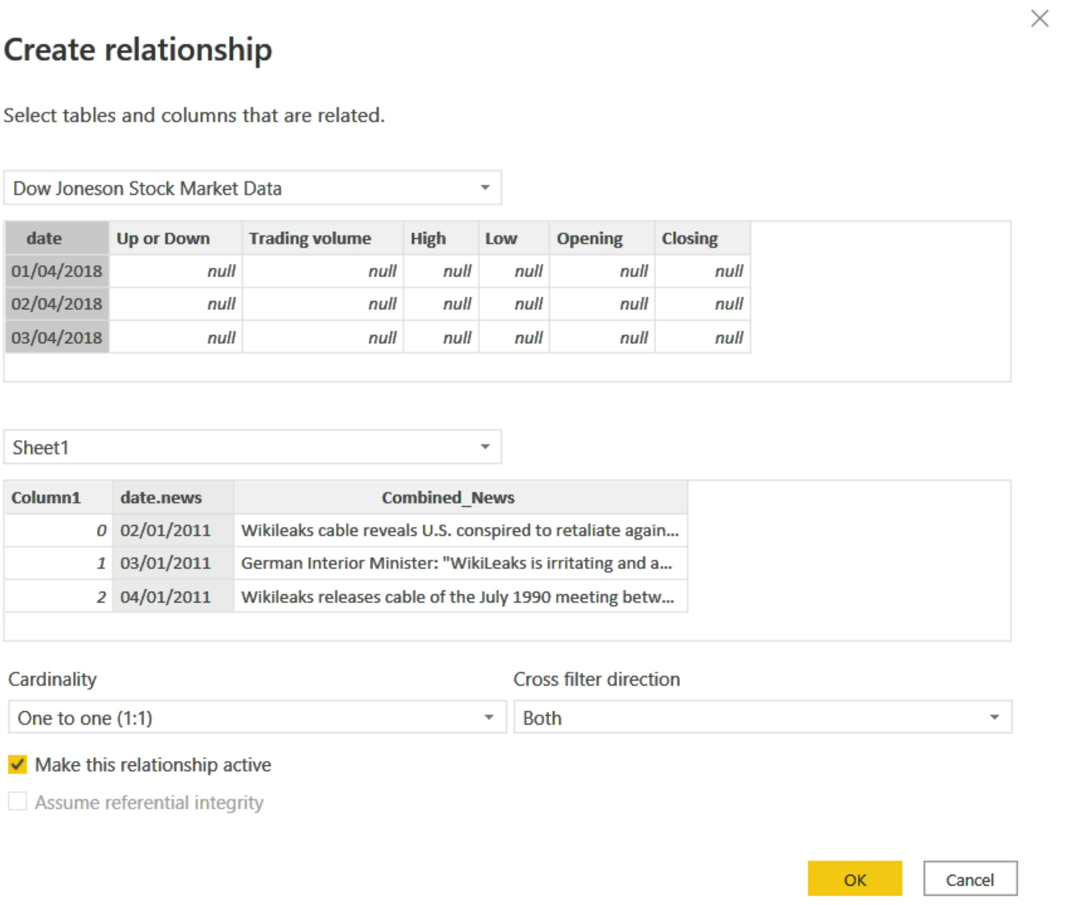
```
data['Combined_News'] = clean_headline
```

```
newdata = data.drop(columns=['Label1','top1','top2','top3','top4','top5','top6','top7'
```

```
newdata.head()
```

	date.news	Combined_News
0	02/01/2011	Wikileaks cable reveals U.S. conspired to reta...
1	03/01/2011	German Interior Minister: "WikiLeaks is irrita...
2	04/01/2011	Wikileaks releases cable of the July 1990 meet...
3	05/01/2011	PunjaPakistan) governor Salman Taseer assassin...
4	06/01/2011	Cable from the U.S. Embassy in Tel Aviv says I...

2.3.4 News text data combined with the Dow Joneson Index



Create relationship

Select tables and columns that are related.

Dow Joneson Stock Market Data

date	Up or Down	Trading volume	High	Low	Opening	Closing
01/04/2018	null	null	null	null	null	null
02/04/2018	null	null	null	null	null	null
03/04/2018	null	null	null	null	null	null

Sheet1

Column1	date.news	Combined_News
0	02/01/2011	Wikileaks cable reveals U.S. conspired to retaliate again...
1	03/01/2011	German Interior Minister: "WikiLeaks is irritating and a...
2	04/01/2011	Wikileaks releases cable of the July 1990 meeting betw...

Cardinality: One to one (1:1)

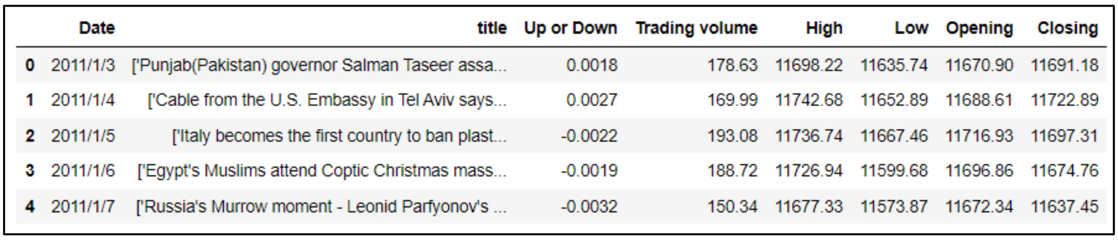
Cross filter direction: Both

☑ Make this relationship active

□ Assume referential integrity

According to relational database principles, One-to-one table join of Table 2 DJIA index and Figure 2 via Power BI (using date columns)

Table 7 Cleaned table Data Set

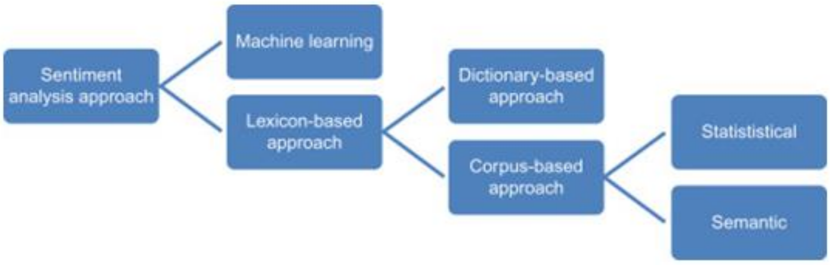


	Date	title	Up or Down	Trading volume	High	Low	Opening	Closing
0	2011/1/3	['Punjab(Pakistan) governor Salman Taseer assa...	0.0018	178.63	11698.22	11635.74	11670.90	11691.18
1	2011/1/4	['Cable from the U.S. Embassy in Tel Aviv says...	0.0027	169.99	11742.68	11652.89	11688.61	11722.89
2	2011/1/5	['Italy becomes the first country to ban plast...	-0.0022	193.08	11736.74	11667.46	11716.93	11697.31
3	2011/1/6	['Egypt's Muslims attend Coptic Christmas mass...	-0.0019	188.72	11726.94	11599.68	11696.86	11674.76
4	2011/1/7	['Russia's Murrow moment - Leonid Parfyonov's ...	-0.0032	150.34	11677.33	11573.87	11672.34	11637.45

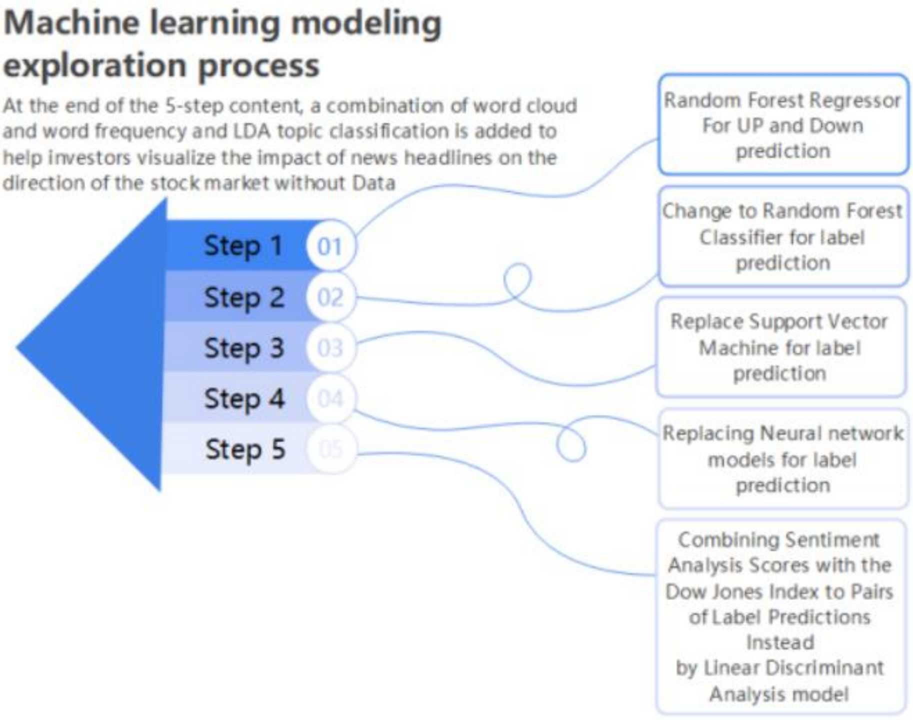
The resulting table is used for machine learning and sentiment analysis for text mining, as well as LDA, word cloud, and word frequency.

2.4 Modelling and Finding of Analysis

Stock Market Forecast There is a line of work that predicts stock markets using text information from daily news. This ground-breaking approach extracts many textual elements from newspapers, including bags of words, noun phrases, named entities, and structured events. Ding et al. (2014) demonstrate that structured events from open information extraction (Yates et al., 2007; Fader et al., 2011) outperform traditional features because they can capture structured interactions. Structured representations of events, on the other hand, have the drawback of increasing sparsity, which may restrict predictive power. Ding et al. (2015) suggest addressing this issue by expressing structured events using dense vectors called event embeddings.



The Lexicon-based technique assesses a document by summing the sentiment scores of all the terms in the document [45–47]. It employs a pre-prepared sentiment lexicon. A term and its related sentiment score should be included in the pre-prepared sentiment lexicon.


Machine learning modeling exploration process

At the end of the 5-step content, a combination of word cloud and word frequency and LDA topic classification is added to help investors visualize the impact of news headlines on the direction of the stock market without Data

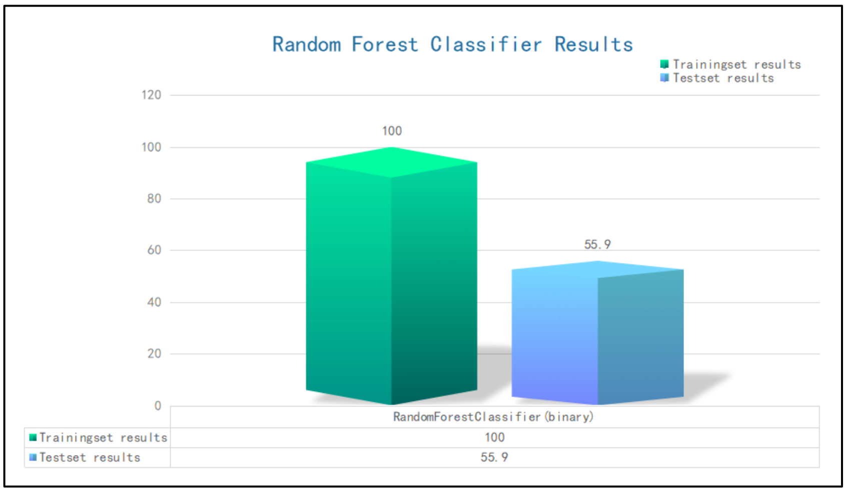
Step 1
01
Step 2
02
Step 3
03
Step 4
04
Step 5
05
Random Forest Regressor For UP and Down prediction
Change to Random Forest Classifier for label prediction
Replace Support Vector Machine for label prediction
Replacing Neural network models for label prediction
Combining Sentiment Analysis Scores with the Dow Jones Index to Pairs of Label Predictions Instead by Linear Discriminant Analysis model

2.4.1 Random Forests Model

Random Forest can work with data in its current form.

First, we modelled the model using Random Forest regression, trying to get the exact (Up and Down) continuous variables, but the results were heavily overfitted(There will be a picture comparison in Part 4 Conclusion), so we downscaled the data using 1 as up and 0 as down, and modelled it using Random Forest classification.

It also reduces overfitting and variance, as it creates as many trees as possible based on a subset of the data and combines the results. With RandomForestClassifier in scikit-learn, we vectorized the title as X and the classification variable label as Y and used a random forest classification model to predict Up or Down. We got the following result.



	RandomForestClassifier(binary)
🟩 Trainingset results	100
🟦 Testset results	55.9

Figure 3 Random Forest Classifier Results

On the training set, the model sometimes reached 100%, but its accuracy on the test set was always around 55%. Random Forest algorithms perform well for training sets, but they fail miserably for test sets due to over-fitting, Although random forests are widely used in machine learning modelling and are suitable for fitting multiple features, they are sometimes prone to overfitting problems due to the excessive branching of random forests.

In this case, we have to solve the overfitting problem by constructing a support vector machine model. Support vector machine models are less prone to overfitting than other models because they consist of a maximum margin hyperplane of the SVM and a margin in which the hyperplane does not seek to fit every sample, but rather finds the hyperplane that best fits the classification by variance. It does not cause overfitting due to too many features.

2.4.2 Support Vector Machines Model

When class separation is clear, Support Vector Machines (SVM) works relatively well. Thus, SVM is intrinsically suited to our two-class label target data. Taking a look at Figure 3, we could easily see that the SVM model performed well on the training set with 99%, but the accuracy on the test set was always somewhere around 56%. As a result of overfitting, SVM does not perform well on new, previously unknown data.

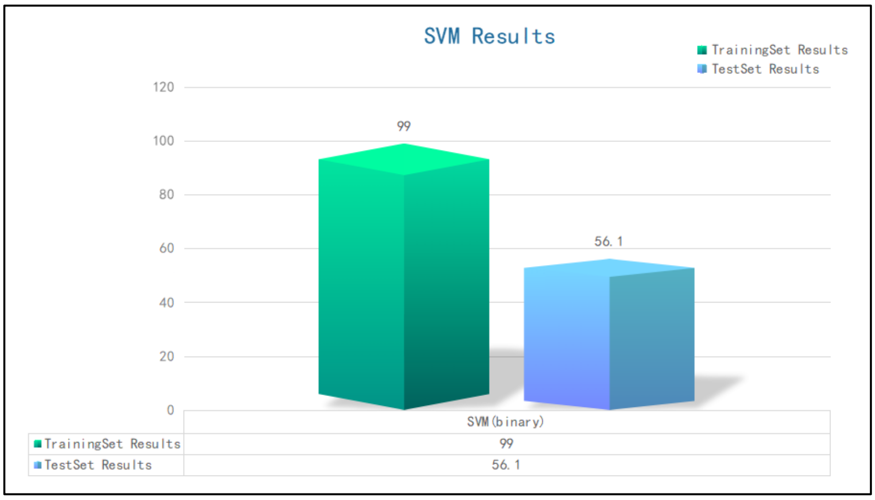


Figure 4 SVM Results

At the same time, many studies nowadays tend to use neural networks for modelling, so we next use neural networks to test the modelling and see if we can reduce the problem of overfitting of the test samples

2.4.3 Neural Network Model

Scikit-learn's MLPClassifier is one of the algorithms that are used in neural network models. MLP uses backpropagation for training, which adjusts the weights of the neurons so that their output approximates the expected output. Because of this, MLPs are considered most suitable for projects involving classification prediction.

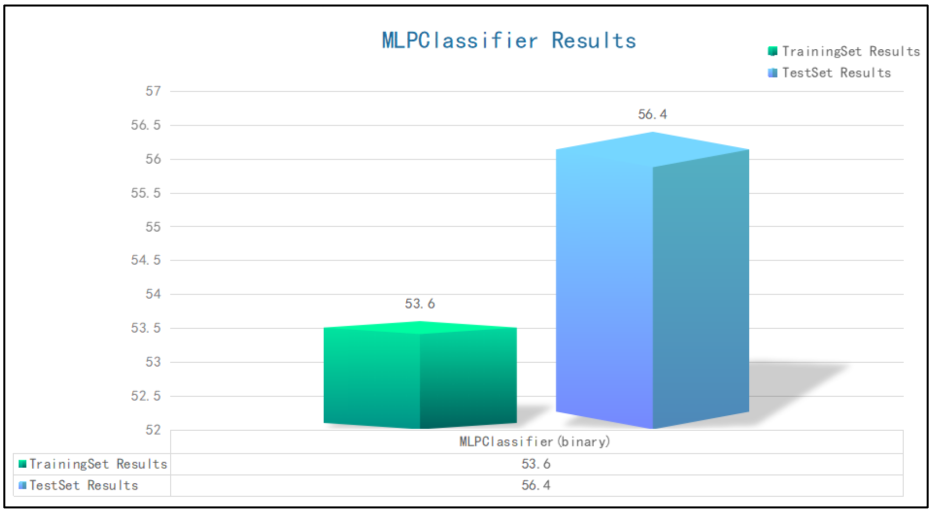
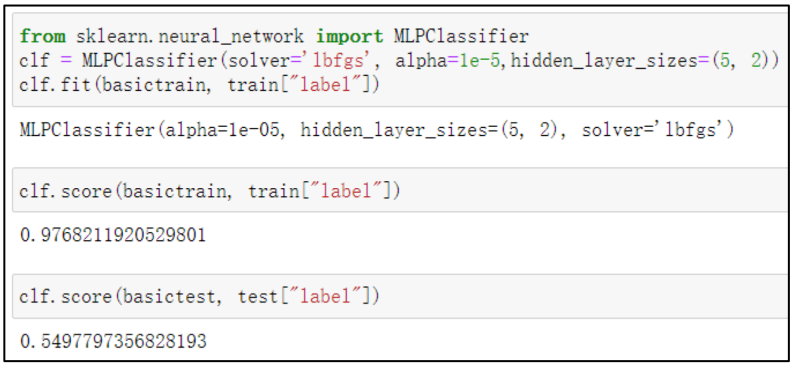


Figure 5 MLP Classifier Results

According to Figure 5, both training and test sets perform equally, but poorly. And our neural network model was retrained with the following results after adjusting the text vectorization of the training set (see Figure 6).



```
from sklearn.neural_network import MLPClassifier
clf = MLPClassifier(solver=’lbfgs’, alpha=1e-5,hidden_layer_sizes=(5, 2))
clf.fit(basictrain, train[″label″])
```

```
MLPClassifier(alpha=1e-05, hidden_layer_sizes=(5, 2), solver=’lbfgs’)
```

```
clf.score(basictrain, train[″label″])
```

```
0.9768211920529801
```

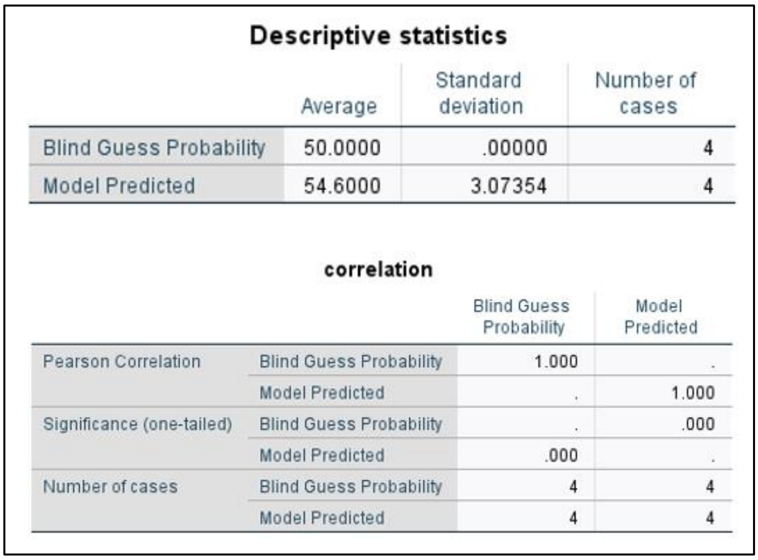
```
clf.score(basictest, test[″label″])
```

```
0.5497797356828193
```

Figure 6 MLP Result after Text Vectorization Adjustment

When we look at the training and test set scores, we can also see that this model suffers from overfitting, as do the SVM and Random Forests models,

so we turned to text mining, using sentiment analysis combined with stock market data in order to train a Linear Discriminant Analysis model. Sentiment Analysis is used in the next step of our process to analyze our headline data and convert it into readable input for Linear Discriminant Analysis Model. Linear Discriminant Analysis Model will act as a classifier and will reduce the dimensionality of the dataset as a result.



Descriptive statistics

	Average	Standard deviation	Number of cases
Blind Guess Probability	50.0000	.00000	4
Model Predicted	54.6000	3.07354	4

correlation

		Blind Guess Probability	Model Predicted
Pearson Correlation	Blind Guess Probability	1.000	.
	Model Predicted	.	1.000
Significance (one-tailed)	Blind Guess Probability	.	.000
	Model Predicted	.000	.
Number of cases	Blind Guess Probability	4	4
	Model Predicted	4	4

The descriptive statistics of the model test results show that there is no significant difference between the model predictions and the blind guesses, probably because of the news information and the high noise of the data due to the influence of too many factors in the reality of the stock market caused our prediction results to be unable to fit the test set well, so we changed our thinking to use Sentiment Analysis to get the sentiment scores of the text content, and then use these scores to assist in the modelling of the stock index to improve the model accuracy.

2.4.5 Sentiment Analysis with Linear Discriminant Analysis Model

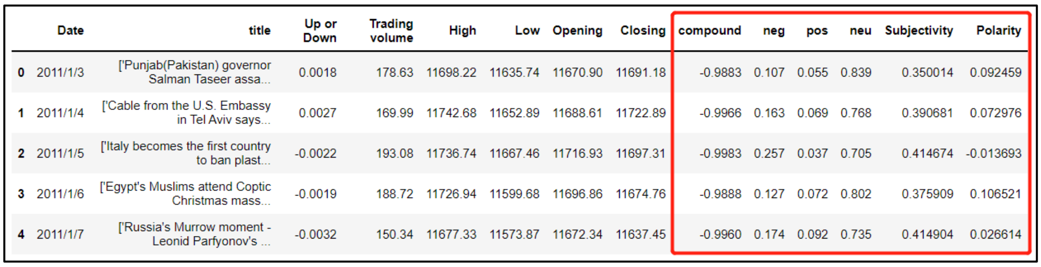
Our project used sentiment analysis through the title column to get sentiment indicators. In order to understand the meaning of news, it is first necessary to measure subjectivity and polarity using the sentiment function of TextBlob in Python. The polarity of a statement is a float that lies between -1 and 1, where +1 means a positive statement and -1 means a negative statement. A subjective sentence refers to a person's opinion, feeling, or judgment, whereas an objective sentence refers to the facts. And subjectivity is a float that ranges between 0 and 1.

The SentimentIntensityAnalyzer would then be applied to determine the negative, positive, neutral, and composite sentiment score (see Figure 7). Threshold values are as follows:

positive sentiment: compound score >= 0.05

neutral sentiment: (compound score > -0.05) and (compound score < 0.05)

negative sentiment: compound score <= -0.05



	Date	title	Up or Down	Trading volume	High	Low	Opening	Closing	compound	neg	pos	neu	Subjectivity	Polarity
0	2011/1/3	['Punjab(Pakistan) governor Salman Taseer assa...	0.0018	178.63	11698.22	11635.74	11670.90	11691.18	-0.9883	0.107	0.055	0.839	0.350014	0.092459
1	2011/1/4	['Cable from the U.S. Embassy in Tel Aviv says...	0.0027	169.99	11742.68	11652.89	11688.61	11722.89	-0.9966	0.163	0.069	0.768	0.390681	0.072976
2	2011/1/5	['Italy becomes the first country to ban plast...	-0.0022	193.08	11736.74	11667.46	11716.93	11697.31	-0.9983	0.257	0.037	0.705	0.414674	-0.013693
3	2011/1/6	['Egypt's Muslims attend Coptic Christmas mass...	-0.0019	188.72	11726.94	11599.68	11696.86	11674.76	-0.9888	0.127	0.072	0.802	0.375909	0.106521
4	2011/1/7	['Russia's Murrow moment - Leonid Parfyonov's ...	-0.0032	150.34	11677.33	11573.87	11672.34	11637.45	-0.9960	0.174	0.092	0.735	0.414904	0.026614

Figure 7 Features value of the dataset after sentiment analysis

We combined the sentiment score with stock market indices and built Linear Discriminant Analysis machine learning models using the Python scikit-learn library. Comparing the LDA model to the other three, it had the highest prediction accuracy of 91% (see Figures 8 and 9).

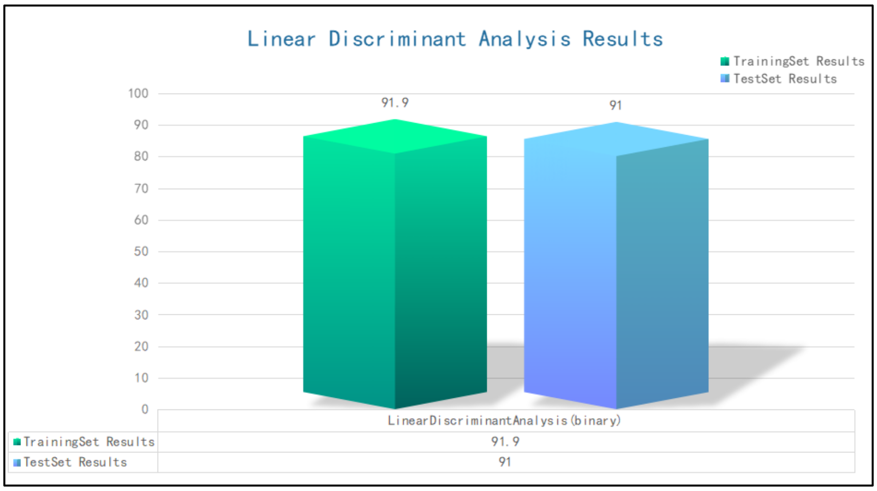
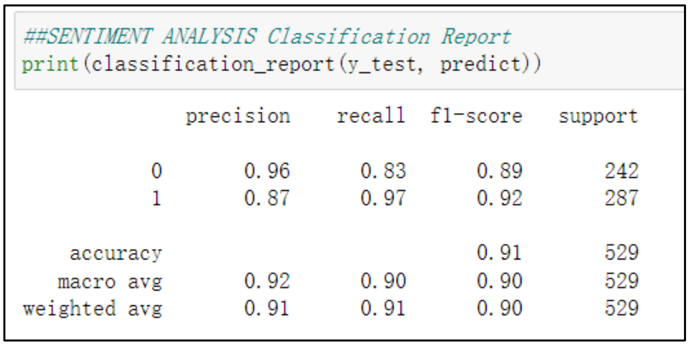


Figure 8 Linear Discriminant Analysis Results



```
##SENTIMENT ANALYSIS Classification Report
print(classification_report(y_test, predict))
```

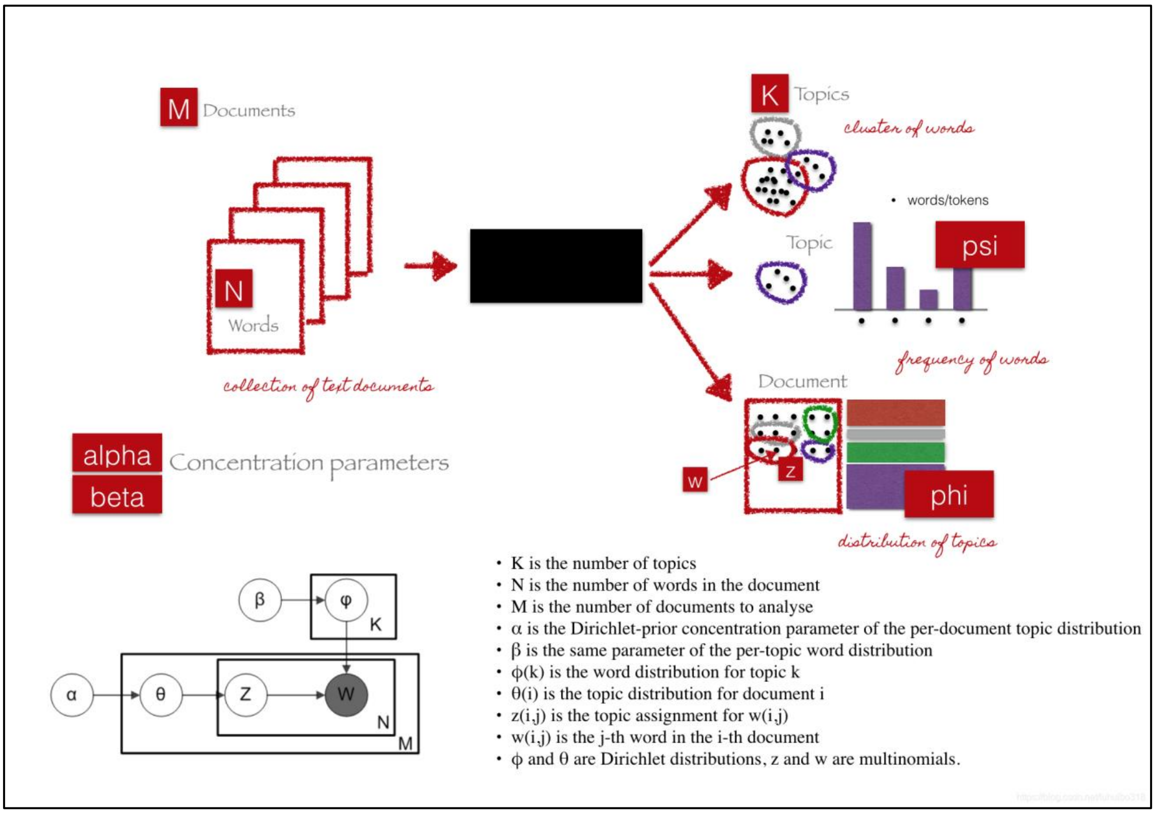
```
              precision    recall  f1-score   support

           0       0.96      0.83      0.89       242
           1       0.87      0.97      0.92       287

    accuracy                           0.91       529
   macro avg       0.92      0.90      0.90       529
weighted avg       0.91      0.91      0.90       529
```

Figure 9 Final Model Report

2.4.5 Word frequency word cloud and Latent Dirichlet Allocation (LDA)

News has a direct impact on short-term stock price movements. In detail, the news dissemination of the media plays a crucial role in the impact of the stock price, causing emotional investors to be easily affected by the news. We try to identify news sentiment keywords that alert investors to upcoming risks, so we consider using word frequency, word cloud, and LDA topic analysis to analyze news headlines that make the Dow Jones stock market up or down, respectively. We hope to not only show the prediction of the stock market tomorrow but also plan to show the word cloud and topics that may affect the stock market soon, so that stock market investors can understand the changes in the stock market from more dimensions

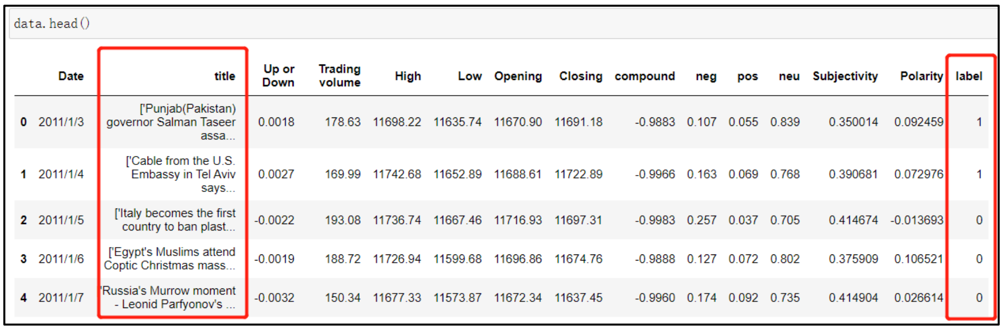
The graph model of LDA is a three-level generation model:



- K is the number of topics
- N is the number of words in the document
- M is the number of documents to analyse
- α is the Dirichlet-prior concentration parameter of the per-document topic distribution
- β is the same parameter of the per-topic word distribution
- φ(k) is the word distribution for topic k
- θ(i) is the topic distribution for document i
- z(i,j) is the topic assignment for w(i,j)
- w(i,j) is the j-th word in the i-th document
- φ and θ are Dirichlet distributions, z and w are multinomials.

Figure 10 Latent Dirichlet Allocation Theory Diagram

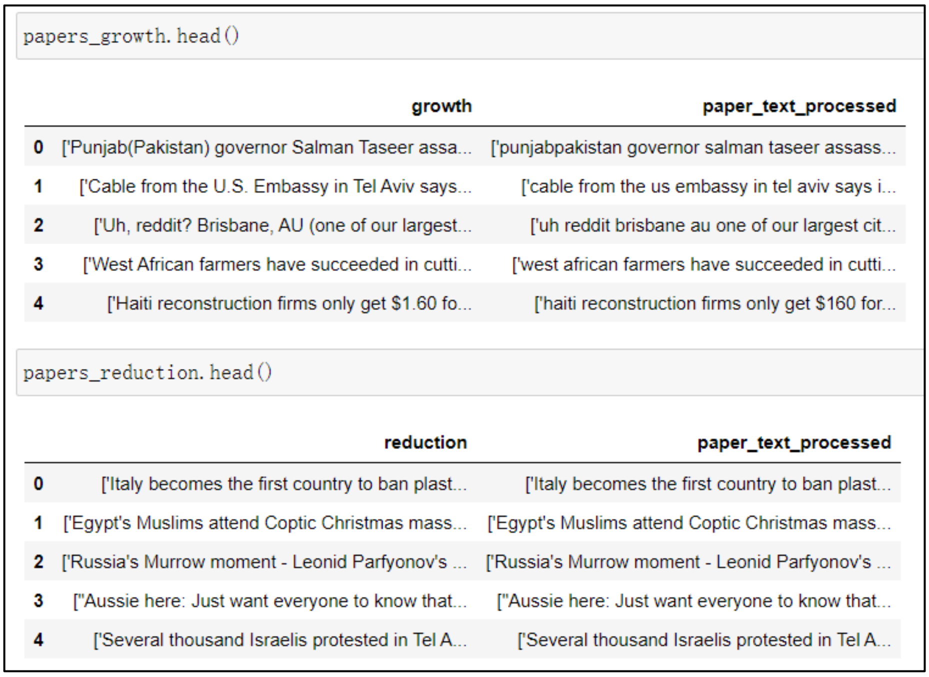
We use the LDA algorithm to model topics and produce clusters of words. They are a mixture of all topics, where each topic has a specific weight label, which is an abstract "topic" that best represents the information in it. For example,........



```
data.head()
```

	Date	title	Up or Down	Trading volume	High	Low	Opening	Closing	compound	neg	pos	neu	Subjectivity	Polarity	label
0	2011/1/3	['Punjab(Pakistan) governor Salman Taseer assa...	0.0018	178.63	11698.22	11635.74	11670.90	11691.18	-0.9883	0.107	0.055	0.839	0.350014	0.092459	1
1	2011/1/4	['Cable from the U.S. Embassy in Tel Aviv says...	0.0027	169.99	11742.68	11652.89	11688.61	11722.89	-0.9966	0.163	0.069	0.768	0.390681	0.072976	1
2	2011/1/5	['Italy becomes the first country to ban plast...	-0.0022	193.08	11736.74	11667.46	11716.93	11697.31	-0.9983	0.257	0.037	0.705	0.414674	-0.013693	0
3	2011/1/6	['Egypt's Muslims attend Coptic Christmas mass...	-0.0019	188.72	11726.94	11599.68	11696.86	11674.76	-0.9888	0.127	0.072	0.802	0.375909	0.106521	0
4	2011/1/7	'Russia's Murrow moment - Leonid Parfyonov's ...	-0.0032	150.34	11677.33	11573.87	11672.34	11637.45	-0.9960	0.174	0.092	0.735	0.414904	0.026614	0

Figure 11 The dataset obtained after sentiment analysis for the group by up as "1" down as "0"



```
papers_growth.head()
```

	growth	paper_text_processed
0	['Punjab(Pakistan) governor Salman Taseer assa...	['punjabpakistan governor salman taseer assass...
1	['Cable from the U.S. Embassy in Tel Aviv says...	['cable from the us embassy in tel aviv says i...
2	['Uh, reddit? Brisbane, AU (one of our largest...	['uh reddit brisbane au one of our largest cit...
3	['West African farmers have succeeded in cutti...	['west african farmers have succeeded in cutti...
4	['Haiti reconstruction firms only get $1.60 fo...	['haiti reconstruction firms only get $160 for...

```
papers_reduction.head()
```

	reduction	paper_text_processed
0	['Italy becomes the first country to ban plast...	['Italy becomes the first country to ban plast...
1	['Egypt's Muslims attend Coptic Christmas mass...	['Egypt's Muslims attend Coptic Christmas mass...
2	['Russia's Murrow moment - Leonid Parfyonov's ...	['Russia's Murrow moment - Leonid Parfyonov's ...
3	["Aussie here: Just want everyone to know that...	["Aussie here: Just want everyone to know that...
4	['Several thousand Israelis protested in Tel A...	['Several thousand Israelis protested in Tel A...

Figure 12 Text dataset with growth and decline as classification divisions

We kept the data after sentiment analysis for LDA and word frequency and word cloud analysis and divided the titles into two categories by labels (the above is the text data after initial de-punctuation and de-capitalization and grouped by labels).

A word cloud is a visual representation of frequently occurring "keywords" in the text. The word cloud can filter out a lot of low-frequency and low-quality text information so that the viewer can grasp the main idea of the text briefly. We plan to show readers the gist of the news that will affect the stock market soon through our webpage, and readers can understand the gist of the stock market news text by simply scanning the word cloud on our webpage.

Cross the following word frequency with the upper and lower classification labels to get the upper- and lower-word frequency and cross the word frequency with the date to get the date word frequency. More information can be obtained in this way. As shown in Figure 11 and Figure 12, we use the word cloud cross-analysis method to analyze the tag identification, and the obtained results are shown in Figure 16 and Figure 17

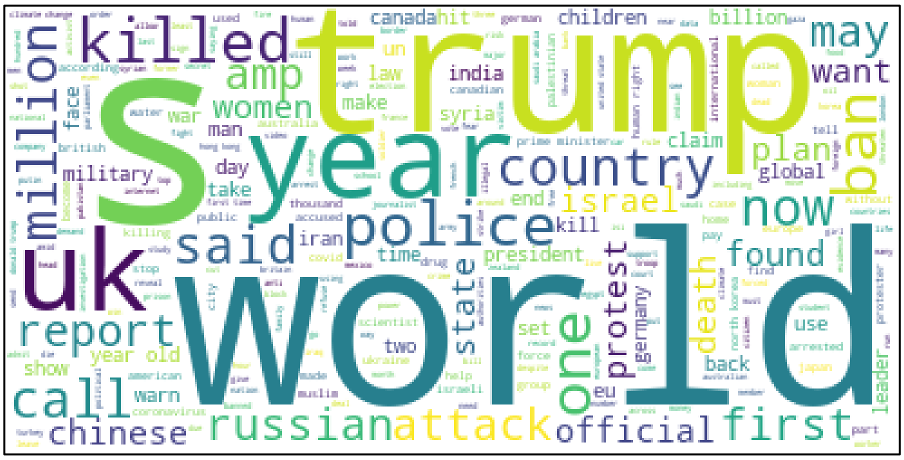


Figure 13 The full text of this word cloud

In our website prototype, we will display the word cloud directly by date, and stock investors can see what news topics related to the stock market have happened this month.

After obtaining the word cloud results, we went on to do a statistical analysis of the word frequency based on the word cloud in order to quantify the direct differences between them and to verify our conjectures from the data.

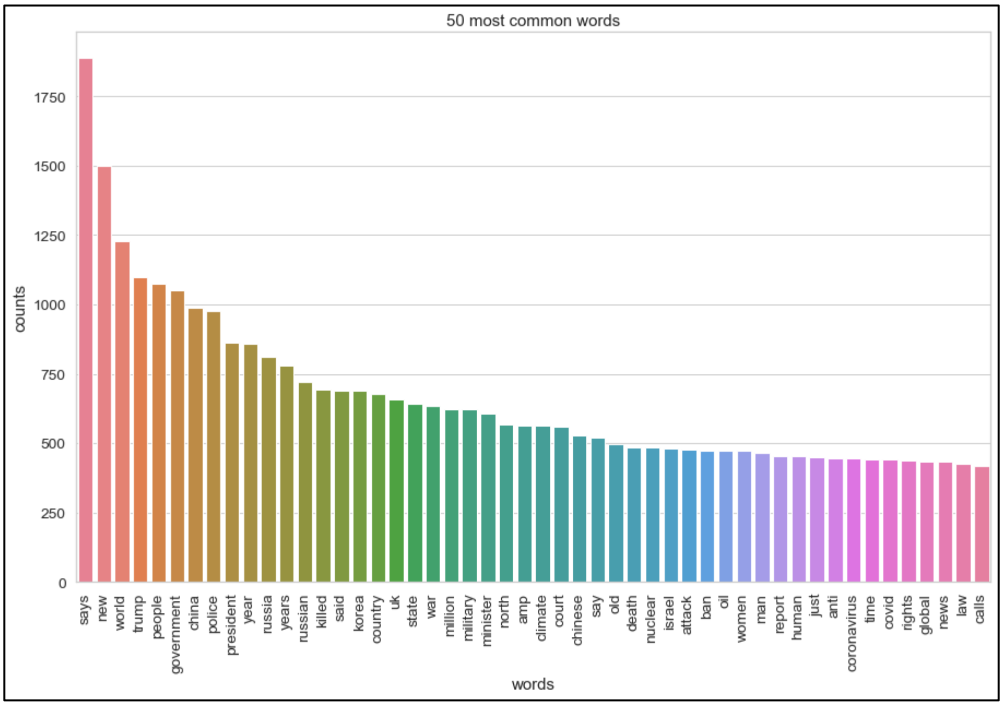


Figure 14 Chart of the top 50 words that made the stock market down

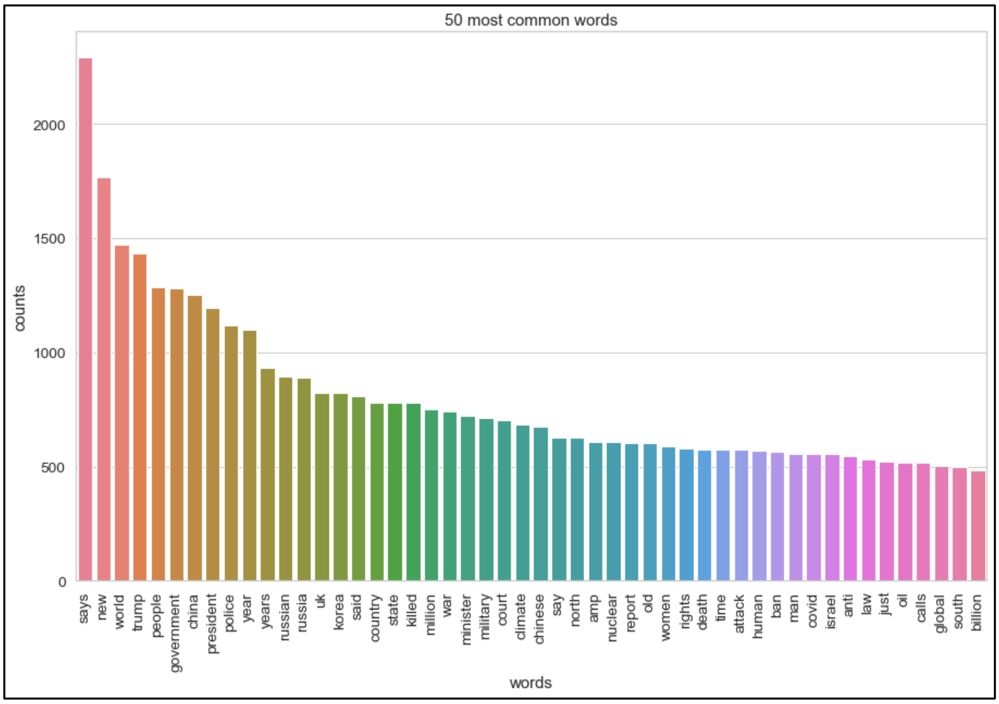
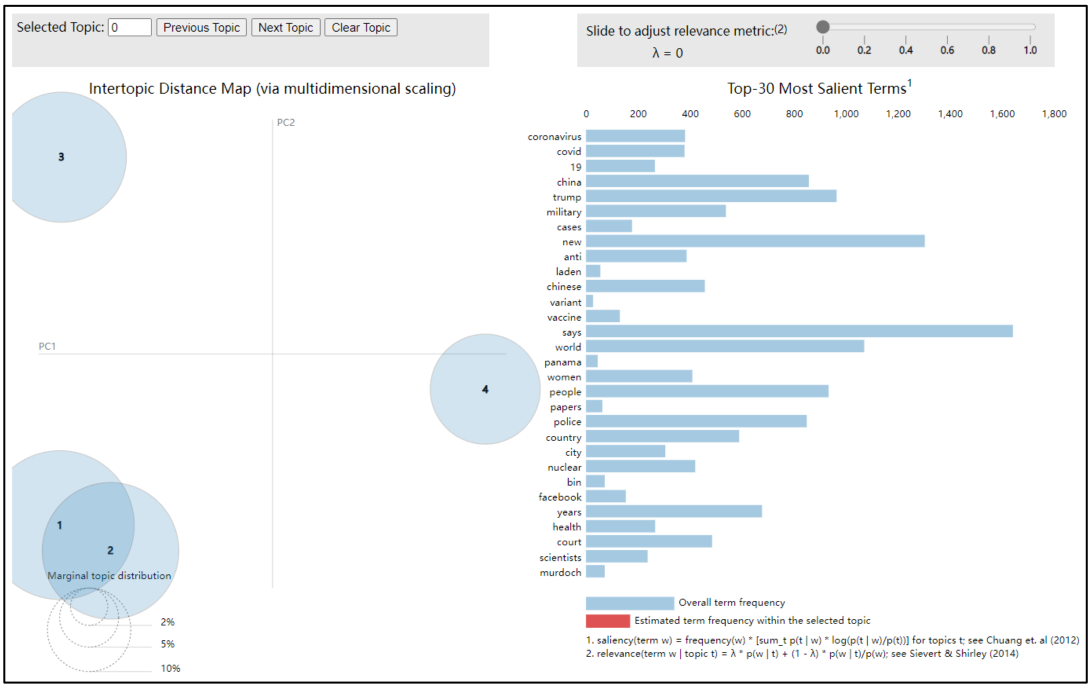


Figure 15 Top 50 words that make the stock market rise frequency chart

Through the statistical analysis of word frequency, we found that most of the high-frequency words have high consistency whether, in the category of up or down, it is impossible to distinguish which words will cause the stock market to rise or fall, but we found that the most concerned topics in the news focus on, international politics, war, oil, China, the United States, Russia, Britain and South Korea, as well as climate, government, nuclear, death, human rights and other issues, It can also be said that these topics have a great impact on the Dow Jones stock market, whether up or down.

Next, we try LDA analysis to categorize these text topics to see if we can find some topics that have an impact on the stock market and categorize them to alert investors to pay special attention to these topics.


Selected Topic: 0

Slide to adjust relevance metric:(2)
λ = 0

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 16 Overview of LDA topic and word frequency distribution chart

The bubble distribution on the left is for different topics, and on the right are the top 30 feature words within the topic. The light blue ones indicate the frequency (weight) of this word in the whole document, and the dark red ones indicate the weight of this word in this topic.

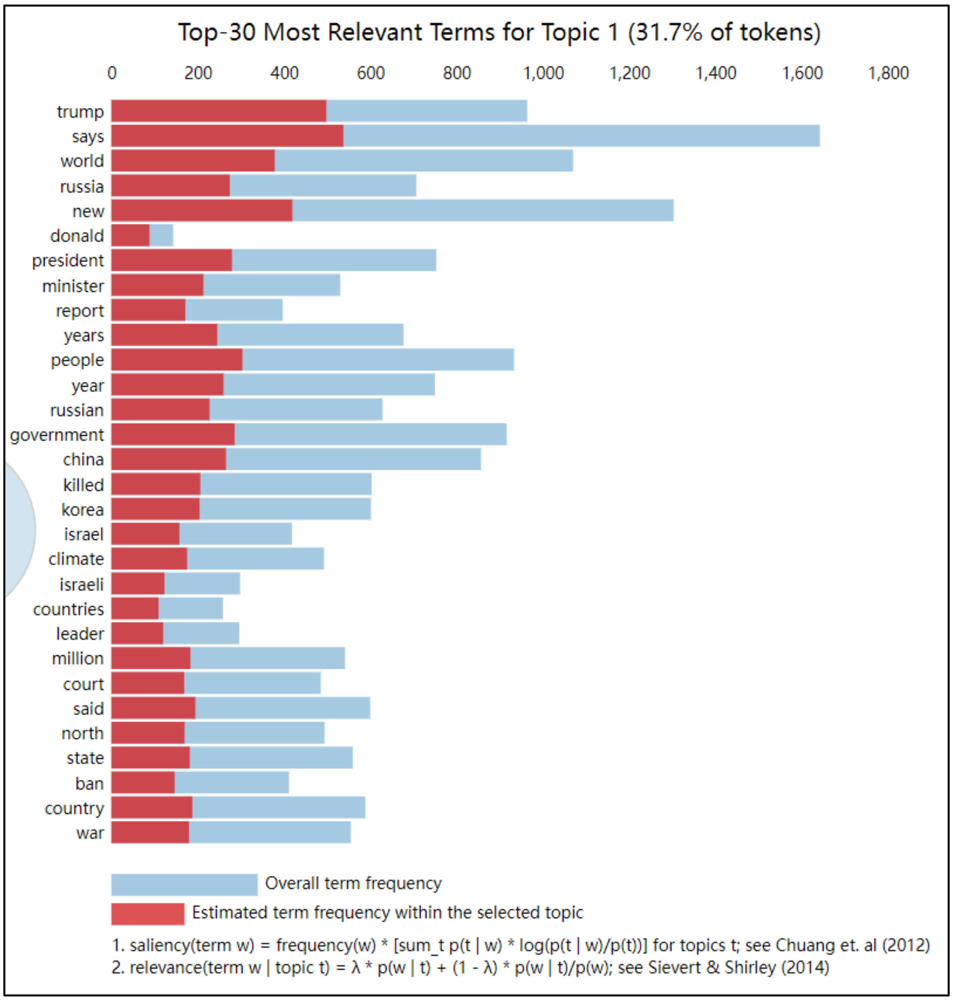


Figure 17 The first category of topics is the word frequency distribution

The main topics covered were Donald Trump and the presidential election, as well as the international politics of the Chinese, American and Russian governments, the war in Syria and the hacking issue.

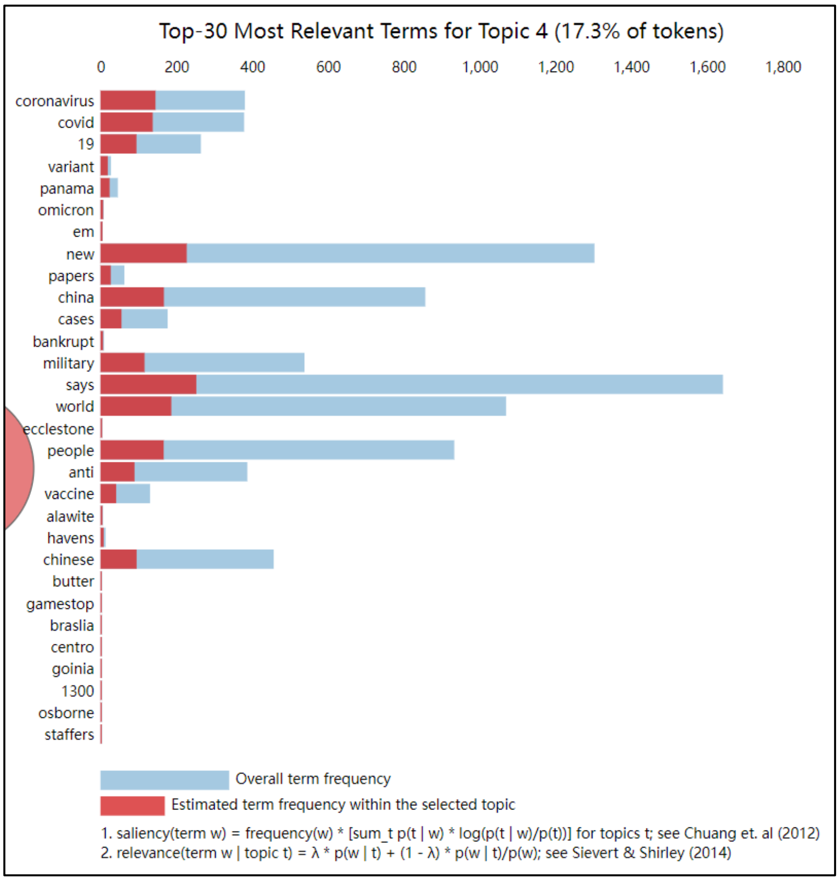


Figure 18 The second category of topics is the word frequency distribution

The COVID-19 virus and its variants and cases, as well as vaccine issues, are described and cover China.

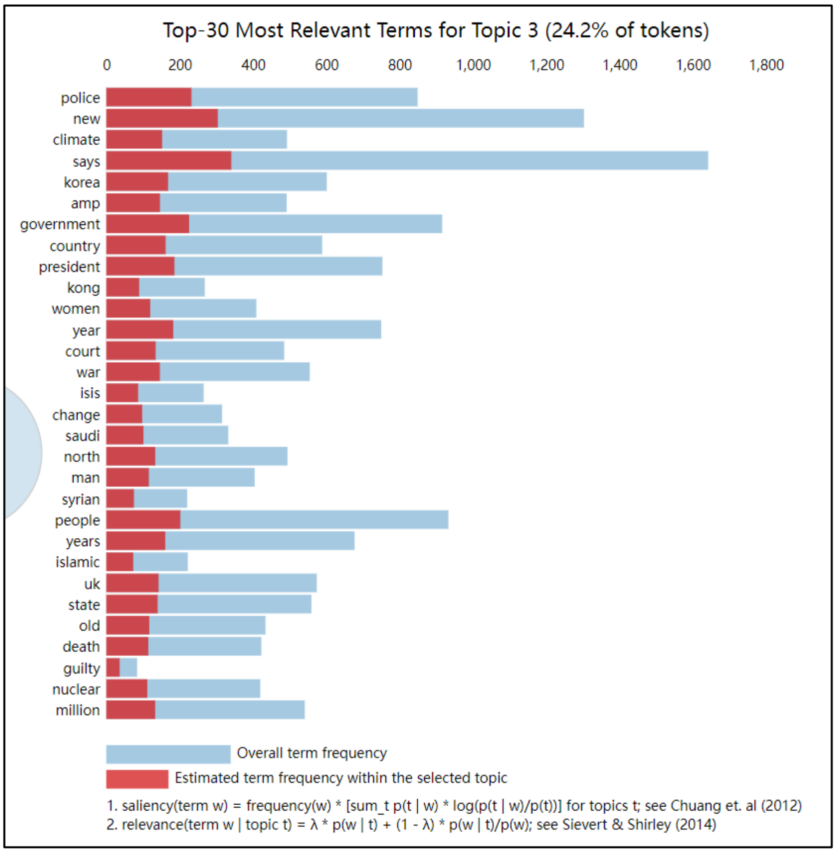


Figure 19 Word frequency distribution of the third category of topics

Crime and climate issues terrorism and North Kkorea'snuclear issues are mainly depicted.

In the above three charts, these three topics have more influence on the Dow Jones stock market, and investors need to pay special attention to the news headline elements that appear in these three topics to influence the Dow Jones stock market.

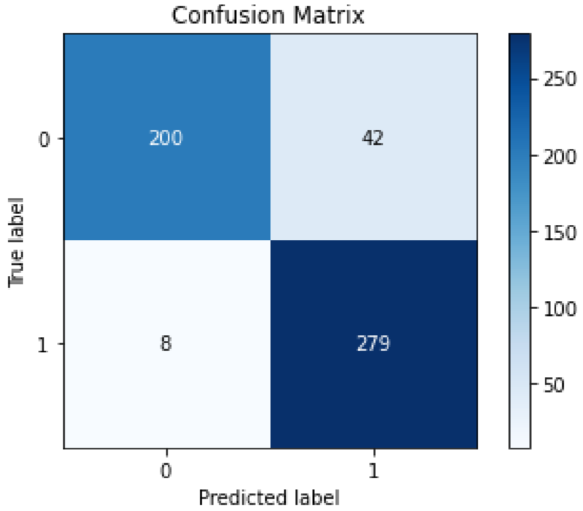
And we can apply the crawler to get the news headlines and classify them into LDA topics to remind stock market investors which topics they need to pay more attention to in order to predict the direction of the Dow Jones stock market and adjust their investment direction and assessment of the current stock market.

The final results will be presented to stock market investors in visualization on the web.

2.5 Evaluation

2.5.1 Evaluate Results

By using sentiment analysis to build a Linear Discriminant Analysis Model, we could see that the results were successful. From our classification report, we can see the Support, which is the number of instances of the given class in the dataset. Since there are 242 of class 0 and 287 of class 1, this can be considered a balanced dataset. With balanced datasets, a 91% accuracy result is meaningful.



2.5.2 Review Process

Currently, it is appropriate to conduct a more thorough review of the data mining engagement to determine if there are any important factors or tasks that have somehow been overlooked. During this phase, we check that data inputs and parameter settings are correct, as well as make sure only attributes that can be used in the future are included in our model.

2.5.3 Determine Next Steps

Based on the 91% accuracy of the linear discriminant analysis model, we could continue the project and proceed with deployment.

2.6 Deployment

2.6.1 Plan Deployment

Taking the evaluation results and applying them to the business is our task. In providing the prediction, we would explain the model's accuracy to the potential stakeholder in terms they can easily understand. We plan to deploy our prototype web application through the Flask framework to visually present the results to the stakeholders

2.6.2 Plan Monitoring and Maintenance

With each passing day, the input data will grow, and there will be both old and recent data available. To avoid incorrect usage of data mining results, We must routinely incorporate new data sets and text content, as well as improve modelling methods to maintain accuracy

2.6.3 Produce Final Report and Review Project

A review of our code with clear comments and model algorithms is necessary from time to time. We can use experience documentation to keep track of our project and to detect any problems that occur in each phase and task. During this phase, conclusions and insights will be provided to potential investors and other stakeholders, allowing them to make informed decisions.

3. Project Tools

3.1 Python

We loop through the /r/worldnews subreddit headlines with the Reddit API wrapper, 'praw' and 'psaw' in Python. With scikit-learn we will use relative algorithms and models to handle our NLP tasks, and sentiment analysis, and to train and test our predictive models.

3.2 Power BI

One-to-one relational concatenation of news headlines and the Dow Joneson Index, based on date columns, using relational database principles and visualisation of model results for comparison

3.3 Tableau

To visualize changes in the Dow Jones Industrial Average, tableau plotted a line graph of the date and DJIA adjusted price. As shown in Figure 10, the DJIA adjusted price generally increases each year. Due to the effect of Covid-19, the DJIA adjusted price decreased in 2020, then gradually recovered to reach its size before the epidemic.

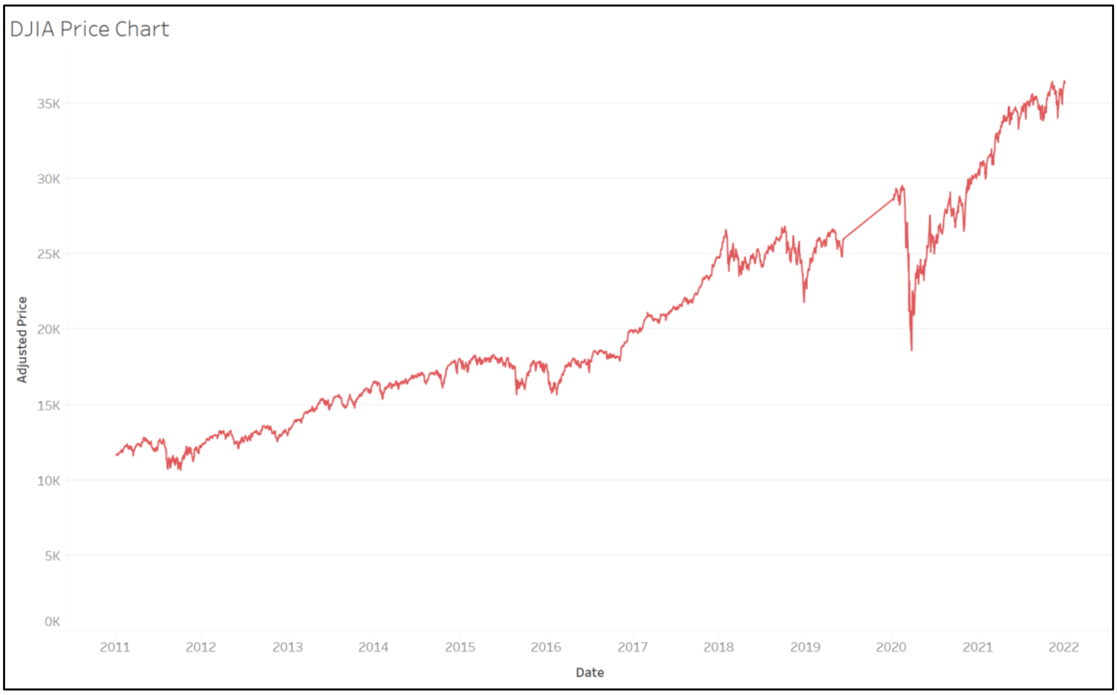
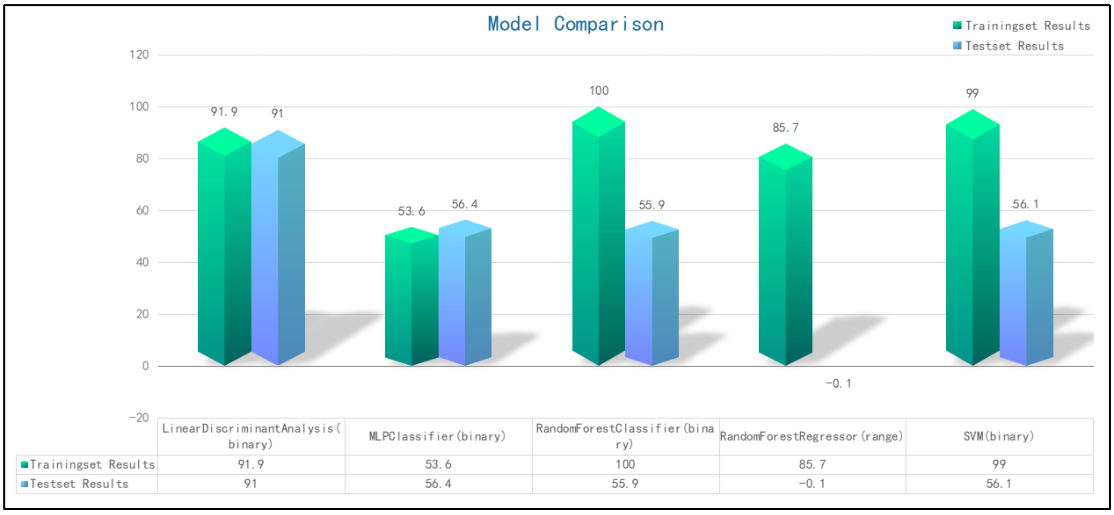


Figure 20 Tableau DJIA Price Chart

4. Results

In this section, the performance of the prediction models will be evaluated and compared. Figure 21 shows the results of both the training and testing sets of the MLP, Random Forests, SVM, and Linear Discriminant Analysis Model with sentiment analysis of the binary classification task. In comparison to the other three models, the Linear Discriminant Analysis Model performs the best. As shown in Figure 9, the model was able to detect the decline in DJIA with 96 percent accuracy while predicting the rise in DJIA with 87 percent accuracy.



	LinearDiscriminantAnalysis(binary)	MLPClassifier(binary)	RandomForestClassifier(binary)	RandomForestRegressor(range)	SVM(binary)
🟩Trainingset Results	91.9	53.6	100	85.7	99
🟦Testset Results	91	56.4	55.9	-0.1	56.1

Figure 21 Model Comparison

5. Discussion

5.1 Findings

The stock market prediction has been an appealing field, Because of its potential effect on investment income, However, financial Markets include noisy data and are nonlinear in nature.

Textual data can be used to forecast prices. The research focuses on stock market analysis using textual data.

More accurate results can be found by combining sentiment analysis, and text information such as news headlines, as it does affect the trend of the stock market, and it is necessary to add them into the model when analysing the data.

Here are three findings from our research

1. a proof of concept that textual information can be used to complement stock indices from a technical and analytical point of view, and to model machine learning to improve the accuracy of machine learning models

2. The sentiment scores derived from textual information can help us improve the accuracy of our models, which inspires us to introduce more textual data into the machine learning modelling process, such as the Wall Street Journal headlines, which are highly correlated with the Dow Joneson Index.

3. For most stakeholders, who may not know how to perform complex stock index analysis and how to extract helpful stock market news from text messages, we designed a prototype application for stakeholders by focusing on the usefulness and benefits of our work for stakeholders.

5.2 Limitation

Technical Limitations:

Deep Learning is commonly used in finance, despite we have substantial advances. But we did not use deep learning for this modelling, Traditional machine learning and text mining are our primary approaches of us.

For us despite the availability of various data sources, Our model system for stock market analysis employs textual data from a single source.

For us to textual data, simplistic text representations such as bag-of-words representation approaches are still extensively utilized. There may now be better techniques and thesauri to supplement the missing semantics

News gets updated on an hourly basis, it’s not constant so if we are predicting the stock price based on news, then there are chances that the news change and we as a customer are unable to understand the change and fail to predict accurately.

(Black Swan Event) An unpredictable and significant event. It happens rarely, but when it does, it has a huge impact. Almost everything that matters is influenced by the Black Swan, and the modern world has been shaped by the Black Swan. From the sub-prime crisis to the tsunami in Southeast Asia, from 9/11 to the sinking of the Titanic, the Black Swan is present in all spheres, whether in the financial markets, in business, in the economy or in personal life.

Therefore, the model needs to be iterated continuously, otherwise, there is a high probability of failure after a certain period of time.

High noise levels in financial data: real information is drowned out in the noise and it is difficult to get valid information without processing the noise

The amount of information that can be expressed in the textual content is limited, and given the excessive amount of features in textual information and the use of only news headlines from Reddit, the amount of data in the corpus may not be sufficient to support modelling to accommodate the rapidly changing Dow Jones index, while the ups and downs of the stock market may be influenced by more than just the single factor of news headlines.

Weak smoothness of financial markets: Financial markets are inherently weakly smooth.

When we or a large number of companies use models to forecast and thus participate in trading in the evidence markets, the models themselves are inside the financial system and these forecasts themselves can interfere with the models' predictions of outcomes.

Non-technical limitations:

Some of the external factors which effects the stock price are, high inflation, low inflation, low-interest rates, and negative interest rate. This noise effect a lot when it comes to fluctuations in stock price

It is likely that more changes in indicators will be reflected in the stock market, such as the gold market, the crude oil market, the international foreign exchange market, and adjustments to benchmark interest rates by central banks, led by the Federal Reserve, could all play a role in the stock market, so we have added the Dow Joneson Index to our sentiment analysis to supplement our modelling to reduce external disruptions.

As discussed above, emotions of humans also affect the stock price, i.e. sometimes news can be manipulative, we can’t be directly dependent on news

We cannot get 100% accuracy as the news can change in every minute affecting the trend of stock price.

Some news broadcasters focusing only on stock news can be bought by companies and never show bullishness in that company's stock.

Non-technical limitations are something out of control hence we can’t predict 100% accuracy anytime. These are external factors we need to accept that it would hamper our accuracy.

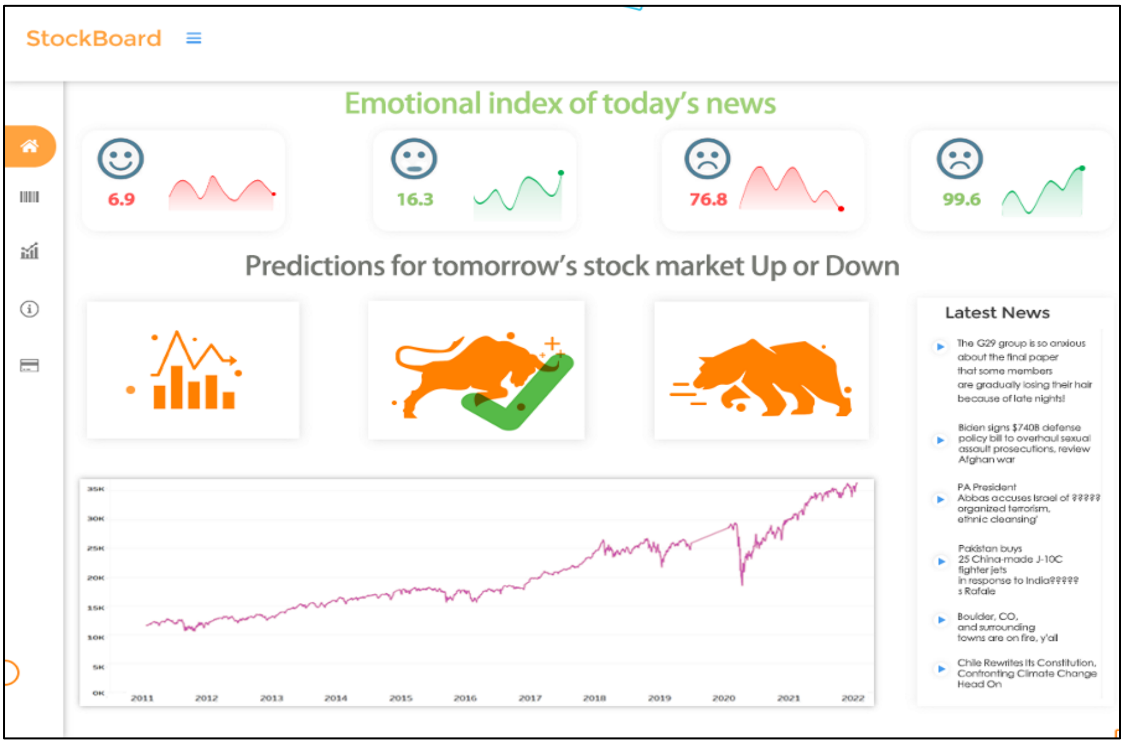
6. Conclusions

The goal of this project is to use sentiment analysis and machine learning to determine the best forecasting model for the DJIA movement and to assist stock participants in making more informed decisions. We observed that sentiment analysis and Linear Discriminant Analysis Model were able to provide more accurate predictions than other models. Based on the accuracy score of 91%, we can conclude that stock metrics and headline sentiment can be used together to enhance the accuracy of other models.

We performed sentiment analysis on news headlines to measure subjectivity and polarity and then calculate sentiment scores for Linear Discriminant Analysis classification models. To remind stock market investors which words need more attention and adjust their investment strategies according to the overall sentiment scores, Linear Discriminant Analysis Model classified news headlines and identified words that impacted the stock market based on their overall sentiment scores.

It is possible that the amount of data in the Python corpus is not sufficient, which will affect the modelling results. Furthermore, market sentiment is one of many factors that influence the stock market. Therefore, this project is only designed to provide stock market participants with the positive or negative impact of recent market sentiment on the stock market and our binary extrapolation of tomorrow's stock market rise and fall, rather than providing precise results.

This project envisions the establishment of a start-up company to design a website. See figure 12 for a prototype of the proposed website. This website not only provides model-based stock forecast trends for stock participants but also aggregates professional news information for stock participants. Insights can then be presented to investors and stakeholders via this recommendation website, using the Linear Discriminant Analysis classification model with sentiment analysis, which is presented in a more visually appealing way. And by showing the most frequently used words that lead to negative, positive, neutral, and composite sentiments on the user interface rather than reading the news headlines directly, viewers can understand news topics faster. By subscribing to this website, stock participants can save time collecting and analyzing the market and gain good professional insights.



Latest News

- The G29 group is so anxious about the final paper that some members are gradually losing their hair because of late nights!
- Biden signs $740B defense policy bill to overhaul sexual assault prosecutions, review Afghan war
- PA President Abbas accuses Israel of ?????organized terrorism, ethnic cleansing'
- Pakistan buys 25 China-made J-10C fighter jets in response to India?????s Rafale
- Boulder, CO, and surrounding towns are on fire, y'all
- Chile Rewrites Its Constitution, Confronting Climate Change Head On

Figure 23 User Interface of prototype web application

7. Appendices

Appendix A - Python Code for News Headlines Extraction in Reddit

The code will be attached to the final submission.

Appendix B - Python Code for Sentiment Analysis and Stock Market Movement Prediction Modeling

The code will be attached to the final submission.

8. References

Adam Atkins, Mahesan Niranjan, Enrico Gerding, 2018. *Financial News Predicts Stock Market Volatility Better Than Close Price,* s.l.: Electronics and Computer Science, University of Southampton, UK.

Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, Dan Jurafsky, 2014. *On the Importance of Text Analysis for Stock Price Prediction,* s.l.: Proceedings of the 9th Edition of the Language Resources and Evaluation Conference (LREC).

JUSTINA DEVEIKYTE, HELYETTE GEMAN, CARLO PICCARI, ALESSANDRO PROVETTI, 2020. *A Sentiment Analysis Approach to the Prediction of Market Volatility.* [Online] Available at: https://arxiv.org/pdf/2012.05906.pdf [Accessed 12 May 2022].

Khedr, Ayman E, Nagwa Yaseen, 2017. *Predicting stock market behavior using data mining technique and news sentiment analysis,* s.l.: International Journal ofIntelligent Systems and Applications.

Marc Velay, and Fabrice Daniel, 2018. *Using NLP on News Headlines to Predict Index Trends.* [Online] Available at: https://arxiv.org/pdf/1806.09533.pdf [Accessed 13 5 2022].

Shearer, C., 2000. The CRISP-DM Model: The New Blueprint for Data. *Journal of Data Warehousing,* 5(A 101commnuications Publication), pp. 13-22.

Sun, J., 2016, August. *Daily News for Stock Market Prediction, Version 1.* [Online] Available at: https://www.kaggle.com/aaron7sun/stocknews [Accessed 5 Apirl 2022].

V .V .Ramalingam, A. Pandian, shivam Dwivedi, Jigar P.Bhatt, 2018. *Analysing news for stock market prediction,* s.l.: Journal of Physics Conference Series.

Wikipedia, 2022. *Linear discriminant analysis.* [Online] Available at: https://en.wikipedia.org/wiki/Linear_discriminant_analysis [Accessed 8 Feb 2022].

Ding, X., Zhang, Y., Liu, T. and Duan, J. (2016). Knowledge-Driven Event Embedding for Stock Prediction. [online] pp.2133–2142. Available at: https://aclanthology.org/C16-1201.pdf [Accessed 8 Jun. 2022].

Mourelatos, M., Alexakos, C., Amorgianiotis, T. and Likothanassis, S. (2018). Financial Indices Modelling and Trading utilizing Deep Learning Techniques: The ATHENS SE FTSE/ASE Large Cap Use Case. [online] IEEE Xplore. doi:10.1109/INISTA.2018.8466286.

www.sciencedirect.com. (n.d.). Lexicon-Based Approach - an overview | ScienceDirect Topics. [online] Available at: https://www.sciencedirect.com/topics/computer-science/lexicon-based-approach#:~:text=The%20Lexicon%2Dbased%20approach%20uses.

Usmani, S. and Shamsi, J.A. (2021). News sensitive stock market prediction: literature review and suggestions. PeerJ Computer Science, 7, p.e490. doi:10.7717/peerj-cs.490.