Using Natural Language Processing Techniques to Analyze the Impact of Covid-19 on Stock Market

MSc Research Project
Data Analytics

Wei He
Student ID: x18144489

School of Computing
National College of Ireland

Supervisor: Dr. Majid Latifi

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Wei He
Student ID:	x18144489
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr. Majid Latifi
Submission Due Date:	23/09/2021
Project Title:	Using Natural Language Processing Techniques to Analyze the Impact of Covid-19 on Stock Market
Word Count:	6429
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Wei He
Date:	23rd September 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Using Natural Language Processing Techniques to Analyze the Impact of Covid-19 on Stock Market

Wei He
x18144489

Abstract

The wide spreading of Covid-19 disease has resulted a high amount of text content being generated on the social network. The probability to explore the impact of social network on the stock price has observed a strong improvement from the utilisation of Natural Language Processing (NLP) in finance and economics domain. This study collected social network comments to calculate whether positive or negative social network text under the category of coronavirus (COVID-19) can affect stock market movement. I take consideration of the period from January to June 2020, analyse 1,630,290 Twitter comments and 10,652 Reddit comments that related to the pandemic during this period. Our optimised topic modelling has delivered more accurate interpretation of the social media context by leveraging the NLP techniques. The text pre-processing process has improved the sentiment analysis result. The result indicates that social network sentiment can impact stock market movement, COVID-19 is the major reason for the movement of the U.S. stock market for the period I studied. This research has found the significant connection between the sentiment and the stock market fluctuation. This research can benefit for a professional fund manager to predict the stock market movement and take actions against an outbreak of COVID-19, as well as offering a different perspective on the effect of an epidemic on the economy and risk avoidance strategies.

1 Introduction

The pandemic of coronavirus (COVID-19) disease across the worldwide have remarkably impacted people's work and lifestyle in general. Financial sector and overall economies around the world are experiencing serious challenges as a result of the COVID-19 outbreak. Up till March 2021 globally there have been 118,471,917 diagnosed cases of and 2,627,207 deaths from COVID-19 [1]. Evaluating and interpreting the economic impact of COVID-19 has turned into an urgent matter. The purpose of this article is to evaluate how the social network sentiment of COVID-19 impact on stock market.

Covid-19 pandemic brings to large amount of text content being generated online, it created unprecedented context for emergency. This provides opportunity to gain understand on a current situation by applying machine learning techniques and performing a text based analysis with NLP techniques.

[1] Data fetched from https://www.worldometers.info/coronavirus/

The project is motivated with personal interests financial response to the Covid-19 crisis. This research provide a good area to explore the domain of Natural Language Processing, in relation to how our word is used for text mining and machine learning perspective. Besides, the Covid-19 pandemic is a new domain of public interest with fast development after its outbreak. With the decrease of people's movements and enhanced lockdown, social media became the major channel for rapid information to be achieved and communication among peoples.

This research paper is to use NLP for the evaluate the impact on the stock market under the influence of the Covid-19 outbreak. Methodology is designed to analyse the any major relationship between the SP500 returns and the comments on the social media platform on a daily basis. In our implementation, I analysed 3 components for daily sentiments i) the average sentiment polarity score. ii) data variance. iii) the volume of the contexts collected. From the stock market side, I used a set of control variables such as i)SP500 index which tracks the movement of 500 large companies listed in US stock exchange, ii)Volatility index (VIX) which tracks the stock market volatility iii) OFR Financial Stress Index as an indication measurement of the stress in the globe market.

1.1 Research Question

The Research Question (RQ) is as follows:

RQ: How to improve the accuracy of predicting Stock Market due to the impact of Covid-19 using NLP techniques?

Sub-RQ: To what extend the topic modelling and sentiment analysis can be efficiently exploited?

1.2 Research Objectives

The research project aims to discover the patterns behind the text data from Twitter and Reddit social media platform to provide the insights on the public sentiments to the Covid-19 crisis, which bring the financial impacts to the stock market. I breakdown the objectives into the below list.

Objective 1: Carry out a critical literature review on the NLP techniques applied on the social media and its connection to the stock market.

Objective 2: Perform the data collection for social media text and stock market data from different platform.

Objective 3: Pre-processing the dataset collected previously.

Objective 4: Applied the exploratory analysis before the modelling.

Objective 5: Extracting the context topics by using LDA topic modelling approach.

Objective 6: Implement sentiment analysis and calculate polarity score from the collect text. Compare Twitter and Reddit result in relation to the sentiment to the Covid-19 crisis.

Objective 7: Analyse the significance of data pre-processing on sentiment score computation to answer **Sub-RQ**. Sentiment score were calculated before and after the pre-processing the Twitter and Reddit text, Paired t-test statistical test was used to compare these outputs.

Objective 8: Measure the impact of sentiment daily average, variance, and volume towards the stock market returns to answer the **RQ**.

The goals of this research to use NLP to carry out analysis on the Twitter and Reddit social media text to tackle the worldwide concern on the Covid-19 crisis, and leverage the correlation matrix to find the determine element which impacts to stock market financially. The major contribution of this research project is to understand the Covid-19 crisis impact on the financial industry. The aim of the research is to help with the financial industry to better deal with Covid-19 crisis during pandemic. Minor contribution is the identification of the use-case of NLP techniques used in the text modelling and sentiment analysis.

This paper is organized as follows. The introduction of the topic is discussed in the current section. In section 2, in-depth literature review of the past related works is discussed. In section 3, methodology is explored and explained. In section 4, model design specification is presented. In section 5, implementation is delivered. In section 6, evaluation is discussed with findings with outcomes obtained from NLP techniques. In section 7, conclusion of the project is delivered.

2 Related Works

Covid-19 crisis has remarkably impacted people and economics globally. Under this unprecedented circumstances, social media plays a vital channel of information and communication channel for human limited from social relationship. Section 2.1 discuss the function of Media Communication during the pandemic Section 2.2 focus on the machine and deep-learning approach based on the sentiment and semantic analysis to analyse the text content using NLP.

2.1 Media Communication During Pandemic

As social network has increased connection to the people nowadays, the news spread much faster than before. Social media become a channel for communicating the crisis updates and events. Research findings show that the news can initially circulate over the social network before they reach to the official media platform such as TV, newspaper (Traylor et al.; 2019). The traditional media receive information later than the social media.

2.1.1 Covid Crisis Information

Crisis Infomatics consists study of the application on information and technology in the various phases of disasters and other emergencies, such as preparation, scaling down, reaction and recovery. Its essential principle is that people use their own information to exchange idea to crisis in preferred ways to deal with crisis. Two related use of Crisis Information for this project are described in the Section 2.1.2 and Section 2.1.3

2.1.2 Communication with Public

Virtual communication was used by the people for creating self-support to the communities through the social media. The platform allow citizen to communicate between each other and share the ideas. With the help of social media, Human being tend to response to the crisis with rational approach, instead of panic. (Helsloot and Ruitenberg;

2004). Social media can share the solidarity among people after crisis such as earth quake and natural disaster. Under the circumstance of uncertainty is caused by additional information and misinformation due to the disorder online activities, the pattern found in response has a great deal of cooperation dealing with social media (Valecha et al.; 2013)

2.1.3 Communication originated from Official to Public

Government and public-sector leaders have been increasingly applied the social media to send the crisis communication to reach the wide range of citizens in a speedy approach. The approach mainly used to advise the public to handle the situation and address the misinformation related to the crisis (Kaewkitipong et al.; 2012). Research suggested that the public usually seek the clarification from social media for confirmation. Tang et al. (2018) concluded the three main approaches for the social media to deal with the medical crisis: (1) evaluate the public's interests and response to the crisis (2) official use of the social media to communicate the crisis (3) evaluate the correctness of the medical information about the crisis.

2.2 Natural Language Processing Techniques

Computers required to understand human in their language and assist with the language-related tasks. The analysis of natural language motivates review of text mining techniques and the application of the Natural language processing.

2.2.1 Sentiment Analysis

English words is general connected with a list of basic emotions category and two type of sentiments (Mohammad and Turney; 2010). Sentiment analysis is applying computation of a polarity measure to categorise data into one of the sentiment group (Bold; 2019). The approach finds the key phases in the sentence and associate these words to the dictionary which computes a similarity score (Hobson Lane, Cole Howard; 2019). lexicon is made from a list of rules which reside in the Sentiment dictionary, text is categorised by examining sentence, order of word and language grammar Beigi et al. (2016).

Dattu and Gore (2015) proposed sentiment analysis on collected text in these techniques such as lexical analysis and hybrid analysis. To carry out a sentiment analysis on social media information, the SVM and Naïve Bayes are used, both machine learning approaches have obtained a very high accuracy. Research indicates the application of the neutral category can increase the validity of the model by study the sentense with distinction of sentiment (Koppel and Schler; 2006; Taboada et al.; 2011). Under the circumstance of not labeled data, it's good to apply the lexical analysis as it uses pre-tagged lexicons for its dictionary (Dattu and Gore; 2015). Unsorted text can be grouped by the Unsupervised Machine learning techniques by studying the unknown pattern without obtain the previous knowledge of the text (Mittal and Patidar; 2019)

There is very little investigating on the sentiment analysis and its association with emotions related to the medicine (Zeng-Treitler et al.; 2008). Natural Language Processing has been used to extract sentiments about a subject from online text documents in Yi et al. (2003). A 7-stage context-aware Natural Language Processing (NLP) techniques was used in the Oyebode et al. (2020) to find related keyphrases and group them into various themes, which helped to reveal negative themes and positive themes during the COVID-19 pandemic. (Aslam et al.; 2020) in their investigation on the Covid-19 pandemic headlines reveals the the sentiment analysis show the 51 percent of headline are associate with negative sentiment, versus to 34 percent positive sentiment and 18 percent neutral sentiment.

2.2.2 Topic Modeling

A meaningful analysis of dataset is possible to be achieved with topics modeling. It helps to easily scan large documents and find out what customers are talking about. Topic Modelling is a set of algorithms that discovers the hidden theme structure inside a document (George and Birla; 2019). This NLP approach includes in investigating the connection between the texts consist in a sentence to form themes. A number of approaches can be used like Latent Semantic Analysis (LSA) or Latent Dirichlet Allocation (LDA).

LSA is essentially a mythology that the hidden patterns can be discovered from the text or word. it is used to identify the related and critical information from the text. The drawback is that it does not consist a sturdy statistical background.

LDA is the simple and popular statistical topic model. It applies the "bag of words" approach by calculating a topic list from a list of the texts. The result is possible to be interpreted by the people. topic generated from a collection of text are made of a group of keywords which are firmly connected with the topic discovered (George and Birla; 2019).

Major advantage of the LDA is that it applies a rich feature and can use the probability algorithm to fine tune the model. LDA can have the function of dimensional reduction, very easy to use with great size of corpus. LDA is a probability theory model, every sentense in the corpus corresponds to a topic. Every topic is has the distribution information among them. LDA show a great advantage where there is a large size of corpus with a few topics which are distributing in the corpus (Jelodar et al.; 2020).

Research Jelodar et al. (2020) used natural language process (NLP) method found on topic modeling to discover several issues corresponds to COVID-19 from social media, it conclude that the LSTM learning model can find useful latent-topics and sentiment classification of COVID-19 comments. it suggested future study to investigate on other social media like Twitter, using multiple machine learning techniques. Thelwall and Buckley (2012) found the lexical extension method is especially suitable for social media text that is covered by a specific topic, suggesting to test the emotion methods on a list of negative topics to calculate whether moods is aligned with topics. Valle-Cruz et al. (2021) concluded that mixing of a lexicon-based approach is improved by a re-position correlation analysis, where the latent or hidden correlations exists in the data.

2.2.3 Deep Learning and Neural Networks

Machine learning(ML) uses algorithms to exam data, extract the pattern from the data, then make the corresponding decision from what it has studied. it is complex and need lots of domain expertise, human input. Deep learning(DL) promotes algorithms in different levels to establish an "artificial neural network" which can be used to study and decide on its own. By comparison, DL has a major advantage by achieving great result and flexibility and learn to represent the world through a nested hierarchy of concepts. DL can discover the hidden layer architecture in the category of words or sentences. The method model the network in a method that enables the requirements to be carried out efficiently without modelling related input features (Bondielli and Marcelloni; 2019).

Neural Networks consists of a group of algorithms that are modelled towards the human brain. It groups the unlabeled data according to the similarities discovered in the input data and then group the data according to the labelled training dataset. Several types of neural networks in deep learning, such as convolutional neural networks (CNN), recurrent neural networks (RNN) are shift the way that we interact with the wider population. Long short-term memory (LSTM) belongs to RNN architecture that can be applied in deep learning. Different to other neural networks, LSTM has feedback connections. It can study various behaviors and carry out tasks which are not achievable by main stream machine learning methods.

Volkova et al. (2017) applied RNN and CNN techinique to classify suspicious and trustworthy news articles. with linguistic feature enabled, the evaluation of both methods shows average precision is high and outperforms lexical models and logistic regression baseline. Wang (2017) applied hybrid models with a combination of RNN and CNN, outperform other model for fake news discovery. Paredes-Valverde et al. (2017) leverage the the convolutional neural network(CNN) and word2vec to detect the opportunities for improve the product quality through sentiment analysis. Xu and Keselj (2019) applied attention-based LSTM deep neural network in future stock market fluctuation forecast and discovered the finance text post between market closure and market open provide better prediction for the following day stock price. Jelodar et al. (2020) applied LSTM model to discover latent-topics and sentiment comment classification based on COVID-19 pandemic from healthcare platform, such as twitter. it conclude LSTM model can detect useful latent-topics and sentiment analysis of COVID-19 comments. it suggested future study to investigate various social media, such as Twitter, using different deep-learning techniques. Nemes and Kiss (2021) find BERT baseline and RNN were producing more accurate result to determine the emotional values without applying neutral classification.

2.3 Covid impact on Stock Market

Recent growing literature revealed the impact of Covid-19 on the stock market. Research Mamaysky (2020) shows that there is a tight connection of stock markets to news in relation to the Covid-19 pandemic. Baker et al. (2020) investigate the stock market reaction to the previous pandemics like SARS, as well as various underline causes and discovered the result of government introduced mobility prohibition and business trading along with voluntary social distancing between individuals as a major factors that trigger the downward adjustment of the U.S. stock market.Huo and Qiu (2020) and Xiong et al. (2020) both showed proof that China has enforced a rigid segregation policy between individuals, which affected the Chinese stock market negatively, however the affection was more moderate than the impact to US stock market. Gormsen and Koijen (2020) investigate how equity dividend futures and stock market levels enable to get the investor expectation about the economic status. investigate the market reaction to the policy interventions in relation to the Covid-19 crisis. Daniel et al. (1998) and Hong and Stein (1999) proposed that investor under-reactions and over-reactions can also indicates movement of stocks triggered by surprise and their emotions.

3 Methodology

This section delivers the methodology which used to tackle the main results to this study, architecture technical design, and process to analyse the collected data. It follows the approach as illustrated in Figure 3

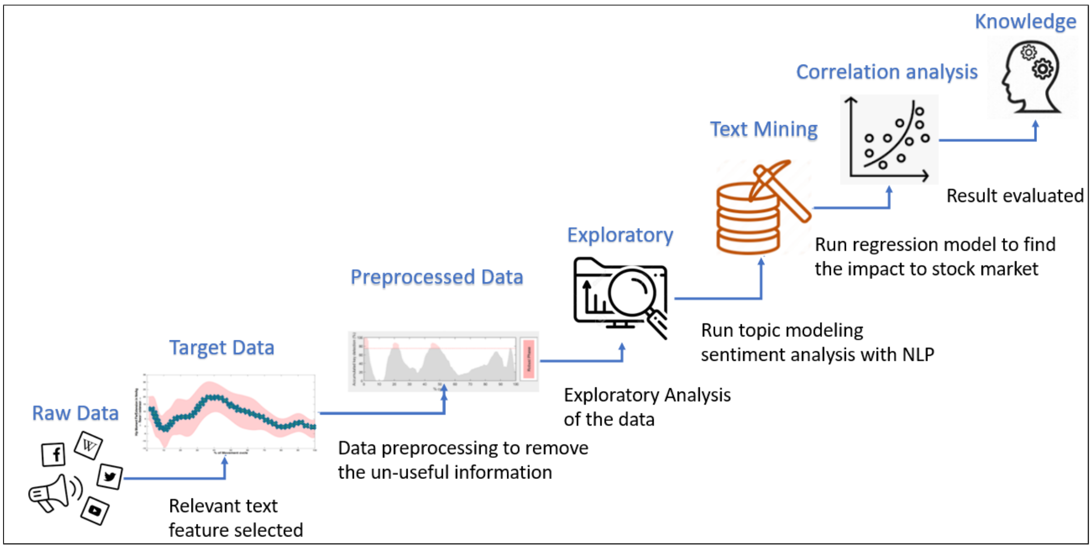

Knowledge
Correlation analysis
Text Mining
Result evaluated
Exploratory
Run regression model to find the impact to stock market
Preprocessed Data
Run topic modeling sentiment analysis with NLP
Target Data
Exploratory Analysis of the data
Raw Data
Data preprocessing to remove the un-useful information
Relevant text feature selected

Figure 1: Methodology approach

3.1 Twitter and Reddit data

3.1.1 Data collection

I have two ways to collect data from different sources as follows:

a) Twitter: media data was fetched using Twitter academic API. The hashtag ”COVID” was used to scrap the tweets during the period from 22 Jan 2021 to 21 Jun 2021. The decision of the hash tag was supported by the popular hashtags search connect with Covid-19 crisis.

Python script was designed to call the Twitter V2 API to fetch the tweets through the TWARC library.The scripts generate 5 unstructured json files with 7.53 GB data. The virus was discovered in China on November 2019, but it gets circulated from January 2020. The date range was selected to capture the first wave of the Covid-19 pandemic worldwide.

b) Reddit data was obtained through pushshift API by using the python PSAW library, which is a minimalist wrapper to search public reddit comments via pushshift.io API. Pushshift is a platform which used to collect Reddit historic data and archive them for future use. These data are made available for the public to use. It is possible to fetch the real time data from pushshift along with the historic data.

Hashtag COVID19 and Cornavirus was selected for fetching the subreddit.Instructions were run with Anaconda jupyter notebook to collect 10,652 subreddit comments between these 2 hashtags between 22 Jan 2021 and 21 Jun 2021.

3.1.2 Transformation

The Data transformation has been performed on the below two data sources:

a) Tweet data were collected and saved to the JSON files which later combine to a large data frame to keep the require columns for analysis. Feature decision was introduced by the finding in the paper (Castillo et al.; 2011). The selected feature were: conversation id, language, timestamp, text, author id, retweet count, like count and sentiment. The feature "sentiment" is associated to the polarity score computed on the un-processed tweet message. The "text" field was exacted with unicode-escape to retain all the tweet text. Each twitter was categorised in to the different language group, e.g "en" is for English. This helped to filter all the English tweets where the English based pre-processing will be applied to. Figure 2 show the visual process of this step.

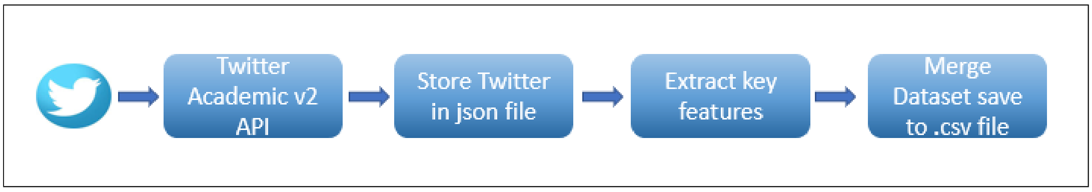


Figure 2: Collect Twitter using Python

b) Reddit text data was extracted based on the hashtag of "Covid19" and "Coronavirus". I generated a separated CSV file for each extraction. A list of the kept features were: "Post ID", "Title", "Url", "Author", "Score", "Publish Date", and "Total No. of comments", " permalink" and "Flair" . this is based on the recommendation from the research paper Jelodar et al. (2020). Figure 3 show the visual process of this step.

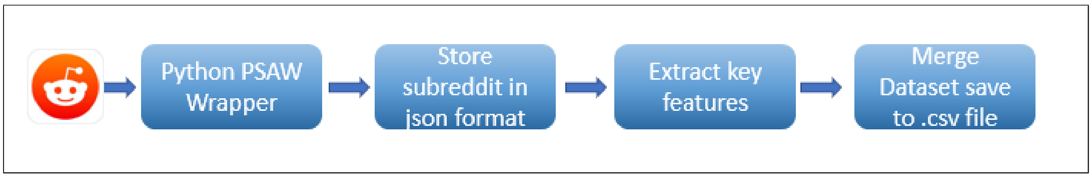

Python PSAW Wrapper
Store subreddit in json format
Extract key features
Merge Dataset save to .csv file

Figure 3: Collect Reddit using Python

3.1.3 Pre-processing

Pre-processing Covid-19 related topic is one of the most important steps. It involves of parsing the text and remove the non-meaningful word such as stop-words. By removing the useless word, I reduced the dimensionality of feature space. I applied the Natural Language Toolkit(NLTK) for text preprocessing.

Twitter and Reddit follows the similar process in relation to the pre-processing. The data cleaning process involves the take the text from English language which is defined to remove the stop words before tokens are lemmatized. It takes the consideration of the language used in social media as informal with emotions expression. The quality of test pre-processing is important to achieve the cleanness of the data and find the useful information. The python library is designed to pre-process tweets in relation to remove the hashtags and URL etc. After the initial clean process by the python library, I have carried out further data cleaning processes as below steps:

1. Remove URL and punctuation.
2. Transform text from uppercase to lowercase.
3. Delete digits plus any words with digits.
4. Delete additional spaces and short words containing only one character.
5. Delete stop words defined in English.
6. Tokenize sentence by splitting text into individual words.
7. Lemmatize to group different inflected forms of words into the root form.

3.2 Exploratory Analysis

Exploratory analysis was carried out to obtain understanding on collected text using python. This is the step before the topic extraction and sentiment analysis.

3.2.1 Word Frequencies

Word Cloud was used to explore the Word Frequencies in Twitter and Reddit data. Word Cloud can present the text data with various size which indicate the frequency of the appears. It provides a directly view of the frequent word in the text. Word Cloud is a popular model what analyses the social media text.

3.2.2 Emotion detection

Text2Emotion is a python library to classify the large amount of text by categorizing it into five different emotions as Happy, Angry, Surprise, Sad, and Fear. it can Processes any textual message and recognizes the emotions embedded in it.Text2Emotion can identify the different emotions from the words obtained from pre-processed text and will keep a count of each and every emotion. a) Find those words which appropriately express emotions or feelings. b) Inspect the emotion category for each word. c) Store the count of all the emotions relevant to all the words which were found.

3.3 Topic Modelling

Topic Modeling is an approach to find the hidden topics from large corpus. There are four popular topic modelling techniques available today : LSA, pLSA, LDA, and lda2vec. All these topic models are based on the same concept: Each document is represented by a group of topics and each topic is made of a list of words. The goal of topic modeling is to find the topics. Topics are effectively shape the understanding of the document and corpus. One of the challenge here is to find a suitable Topic modelling for the data I have.

After the research for these method, I have investigate the pros and cons for each approach. Table 1 presents some comparison made between these topic modelling. Based on the comparison result. With the help of the comparison table, I carefully choose the LDA for my topic modelling implementation, because it is easy to generate new document and achieve the dirichlet distribution.

There are a list of benefits to use the Latent Dirichlet Allocation(LDA) approach which uses Bayesian method. It applies dirichlet priors for the document-topic and word-topic extraction.. It is a useful algorithm for topic modeling with optimised implementations in the Python's Gensim package. LDA algorithm was used to analyse the themes discussed in Twitter and Reddit. LDA can parse the supplied documents and generate the topics from it . It adjust the topics distribution inside the documents and keywords distribution inside the topics to achieve a optimised composition of topic-keywords distribution.

Below are the steps to leverage the LDA for the Topic modelling:

1. Create Dictionary (id2word) which keeps word for each id.

2. Create the Corpus which show the association of (word_id, word_frequency)

3.Build the Topic Model by supplying corpus, id2word and num_topics etc

4.Exam the topics in LDA model. LDA model was built with number of topics where each topic consists a list of keywords and each keyword has a certain weight to the topic.

Method	Description	Pros	Cons
LSA	Take a matrix of documents and decompose it into a separate document-topic matrix and a topic-term matrix.	Quick and efficient to use	Lack of interpretable embeddings
pLSA	Find a probabilistic model with latent topics that can generate the data with document-term matrix	Adds a probabilistic treatment of topics, Flexible	No parameters to model P(D),Prone to overfitting
LDA	Bayesian version of pLSA, bring itself to optimised gereralization.	Generalize to new documents easily, dirichlet distribution	Requires normal distribution assumption on features/predictors
lda2vec	Mix of word2vec and LDA	Combining global document themes with local word patterns	Context/paragraph vectors resemble typical word vectors, making them less interpretable

Table 1: Topic modelling comparison

These keywords indicates what this topic could be.

5. Compute Model Perplexity which seen as a good measure of performance for LDA

6. Compute Model Coherence Score. topic coherence measures result of a single topic by accessing the degree of semantic similarity among high scoring words in the topic

7. Visualize the topics-keywords by examining the generated topics and the associated keywords using pyLDAvis package's interactive chart inside of jupyter notebooks.

8. Find the decisive topic for each sentence

9. Find the most representative text for every topic

3.4 Sentiment analysis of COVID-19

I am going to extract the sentiment from the TWeets and Reddit documents separately, I compared the sentiment result(Polarity scores) before and after the pre-processing to investigate whether there is any improvement gain through the text pre-processing.

Polarity scores for each Tweet and Reddit were calculated using Python TextBlob for sentiment classification. TextBlob is built on top of NLTK for fast-prototyping. Unsupervised technique was used here to capture the text sentiment and investigate the words association and position. Sentiment is generated by associate the keywords from the text with classification from the dictionary established.

3.5 Stock market and Covid-19

This paper investigates the relationship between the stock market and the Covid-19 sentiment from the social media. I have selected a list representative indexes from the stock market. These indexes are : a) SP500[2] b) The volatility index VIX[3] c) OFR Financial Stress Index[4].

[2] SP500 data download from https://www.marketwatch.com/investing/index/spx/download-data

[3] VIX data download from https://www.marketwatch.com/investing/index/vix/download-data

[4] OFR data download from https://www.financialresearch.gov/financial-stress-index/

From the Sentiment side, I have calculated the mean, variance and volume on a daily basis for both Twitter and Reddit respectively. The "mean" represents the daily average sentiment score on a given day. The "variance" of the sentiment is the average of squared difference from the mean value on given day. the "volume" represents the number of Covid-19 relevant comments were fetched from Tweet and Reddit text. I merged these sentiment data into the stock market date before running the Pearson correlation and linear regression to find the correlation and significance.

4 Design Specification

The below 3-tier architecture diagram as Figure 4 describes the approach use to find the correlation between sentiments collected from the social media and the stock market index for future prediction.

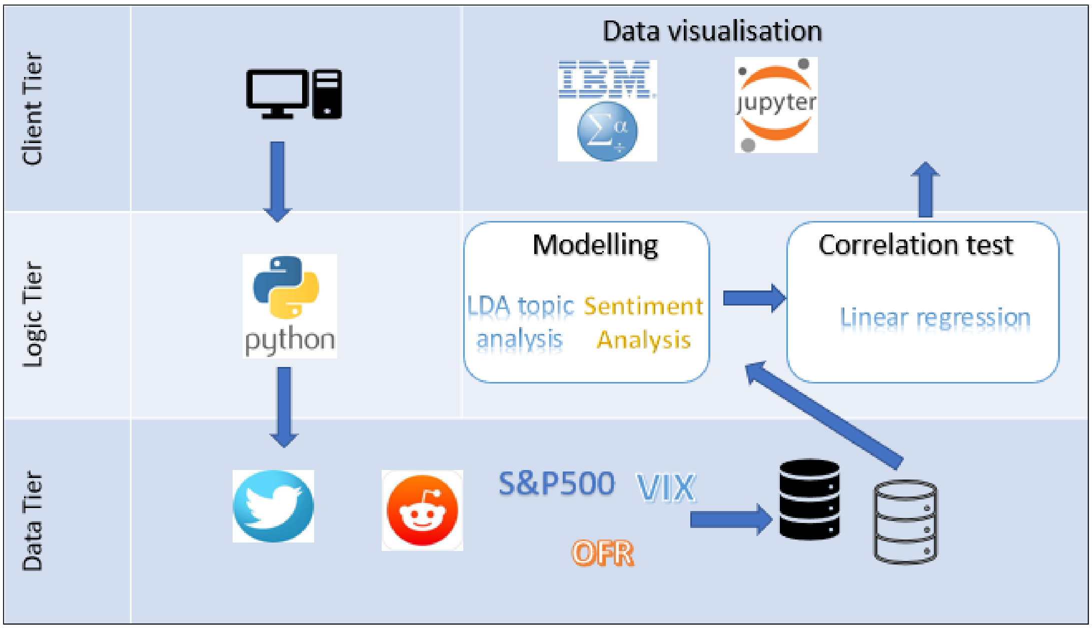


Figure 4: Design Specification

The process flow starts with the client tier where presents the visualisation to the stakeholder who has a vision where the date was collected. The second component is Logic Tier which involves the capture the require data and run the data pre-processing on the text. The third component is the Data Tier which is used to create Pandas data frame to store all the data (in csv format) from Tweet, Reddit platform and the stock market daily data. The data frames are delivered to Logic Tier to run the topic modelling and sentiment analysis algorithm, as well as run the Pearson correlation and linear regression model to calculate the correlation.

The results get returned back to the client tier for the findings to be visualised using IBM SPSS and Jupyter notebook to display the machine learning model to tackle the research question of this project. The below section describes the implementation of this research.

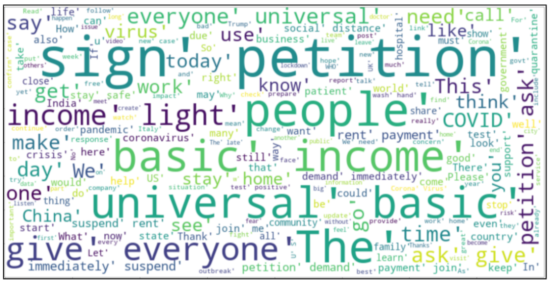


Figure 5: Twitter Frequent words

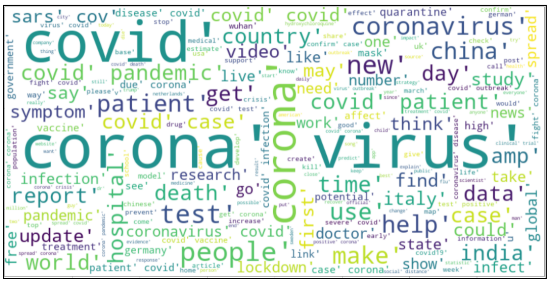


Figure 6: Reddit Frequent words

5 Implementation

This project consists two main objectives. Firstly, to optimise the topic modelling using unsupervised clustering LDA approaches. Secondly, to identify the relationship between Stock market and the sentiment polarity score from the text.

5.1 Word Frequency

I used Word Cloud to pick maximum of 50 words with frequency over 100. Frequencies in Figure 5 and Figure 6 shows that twitter used the term ”sign” ”Petition” more frequent than the ”COVID”, while Reddit has dominate use of ”Covid” and ”Corona Virus”.

5.2 Topic Modelling - Latent Dirichlet Allocation

After carefully researched list of topic modeling options, I have chose the Latent Dirichlet Allocation model to investigate the text extract from the Twitter and Reddit. The reason is listed in chapter 3.3. I applied th Gensim’s inbuilt version of the LDA algorithm with the Mallet’s version to get the better quality of topics. With LDA, I can extract human-interpretable topics from a document sample, where each topic is featured by the words they are most tightly associated with.

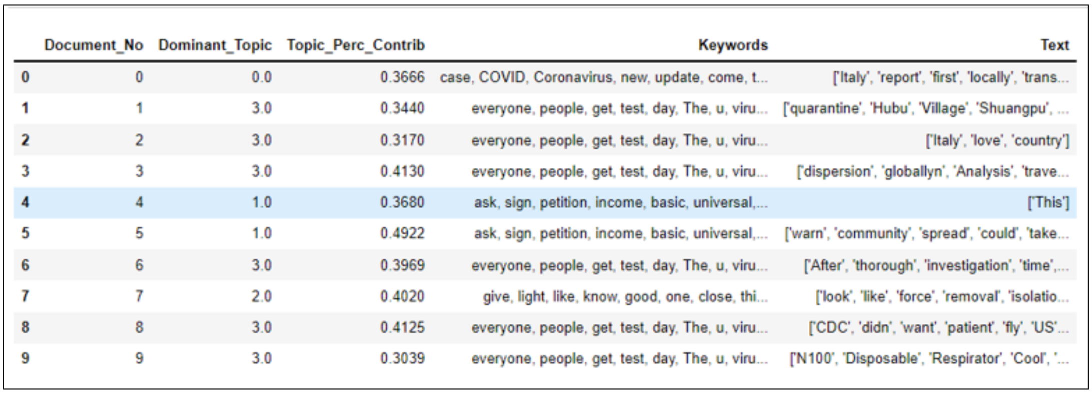
a) Calculate the optimal number of topics for LDA the objective here to find the optimal number of topics (k) and pick the one that gives the highest coherence value. I trained a number of LDA models and exam the corresponding coherence scores. I tested from 2 topic up till 40 topic with step of 6.

b) Discover the dominant topic in each sentence. I applied the format topics sentences function to find the topic number that has the highest percentage contribution in each document using unsupervised learning. Figure 7 gives us the feedback which topic is associated to the current text content.

c) Discover the most representative document for each topic. In some circumstance, topic keywords does not clearly reflect what the topic is about. To better understand the topic, I can locate those documents a particular topic has contributed. I used Python to group top 5 senescence under each topic by reusing the result from the format topics sentences function as shown in Figure 8.

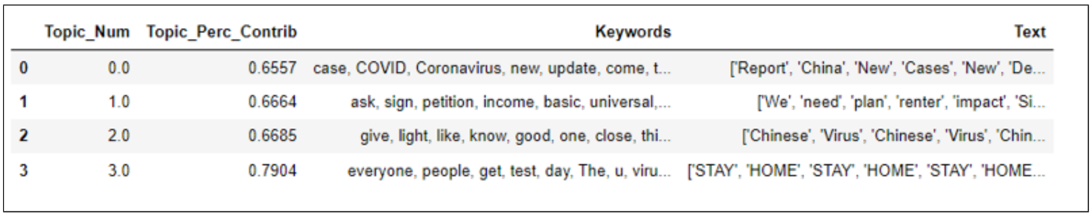
5.3 Sentiment Analysis

I used the Python TextBlob lexicon to compute the polarity score for each tweet. It leverages the unsupervised technique to calculate the sentiment score based on the



	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	0.0	0.3666	case, COVID, Coronavirus, new, update, come, t...	['Italy', 'report', 'first', 'locally', 'trans...
1	1	3.0	0.3440	everyone, people, get, test, day, The, u, viru...	['quarantine', 'Hubu', 'Village', 'Shuangpu', ...
2	2	3.0	0.3170	everyone, people, get, test, day, The, u, viru...	['Italy', 'love', 'country']
3	3	3.0	0.4130	everyone, people, get, test, day, The, u, viru...	['dispersion', 'globallyn', 'Analysis', 'trave...
4	4	1.0	0.3680	ask, sign, petition, income, basic, universal,...	['This']
5	5	1.0	0.4922	ask, sign, petition, income, basic, universal,...	['warn', 'community', 'spread', 'could', 'take...
6	6	3.0	0.3969	everyone, people, get, test, day, The, u, viru...	['After', 'thorough', 'investigation', 'time',...
7	7	2.0	0.4020	give, light, like, know, good, one, close, thi...	['look', 'like', 'force', 'removal', 'isolatio...
8	8	3.0	0.4125	everyone, people, get, test, day, The, u, viru...	['CDC', 'didn', 'want', 'patient', 'fly', 'US'...
9	9	3.0	0.3039	everyone, people, get, test, day, The, u, viru...	['N100', 'Disposable', 'Respirator', 'Cool', '...

Figure 7: Dominant topic in sentence



	Topic_Num	Topic_Perc_Contrib	Keywords	Text
0	0.0	0.6557	case, COVID, Coronavirus, new, update, come, t...	['Report', 'China', 'New', 'Cases', 'New', 'De...
1	1.0	0.6664	ask, sign, petition, income, basic, universal,...	['We', 'need', 'plan', 'renter', 'impact', 'Si...
2	2.0	0.6685	give, light, like, know, good, one, close, thi...	['Chinese', 'Virus', 'Chinese', 'Virus', 'Chin...
3	3.0	0.7904	everyone, people, get, test, day, The, u, viru...	['STAY', 'HOME', 'STAY', 'HOME', 'STAY', 'HOME...

Figure 8: Representative document per topic

analysis of words context, extraction of noun phrase and tagging part-of-speech etc. Sentiment analysis is carried out based on the determination of the attitude or the emotion of the text, in another word it can be either positive or negative or neutral.

The sentiment function of TextBlob resulted in 2 attributes, polarity, and subjectivity. Subjective sentences usually corresponds to personal opinion, emotion or decision whereas objective corresponds to partial information. Subjectivity is a float variable which is in the range of [0,1].

In order to find the Pre-processing effect on the Tweets and Reddit, the sentiment analysis has been carried out twice. The first time the analysis was run on the raw data and saved as ”sentiment” feature while the second time the analysis was run on the pre-processed data with result saved as ”sentiment2” feature. I have separated the polarity score into 3 levels. positive with score greater than 0.3. Negative with score less than -0.3. Neutral with score between +0.3 and -0.3. the original ”sentiment” feature has kept the same boundary. I am going to the use the output to work out the confusion matrix to obtain the prediction model accuracy. From the confusion model, it provides the quality of the sentiment analysis with the effective pre-processing process onto the tweet text.

5.4 Sentiment statistics with the stock market movement

I investigated the correlation between the extracted Covid-19 text sentiment and the financial markets data. I mainly focused on the financial market in particular the returns on the SP500. Covid-19 daily sentiment consists three daily variable i) The average daily sentiment. ii) the daily variance of the sentiment. iii) the volume of the comments for the day. I included a list of control variables that potentially incur a impact on the stock market returns: i) SP500 ii) Volatility index (VIX) iii) OFR financial stress index which tracks the stress in the worldwide stock markets. The goal here is to find the determine factor that could impact the stock market movement. I am aiming to apply the Pearson correlation and linear regression to find the key contributors to the stock market. The reason to use both Pearson correlation and linear regression is to have cross verification on the findings.

6 Evaluation

The evaluation section is split into 2 parts. First part focuses on the results of the topic modelling output and sentiment analysis. Second part focuses on the relation of Stock market movement with the sentiment text.

6.1 Topic modelling Evaluation

I used topic coherence score to evaluate the LDA topic model. Topic Coherence calculate score a single topic by finding the degree of semantic similarity between high scoring words in the topic. These calculation is benefit in terms of find the difference among the topics

First, I found the optimal number of topics for LDA. The objective here to find the optimal number of topics (k) and select the one that provides the highest coherence value. I fine tuned a multiple LDA models and calculate the models with their coherence scores accordingly.I test from 2 topic up till 40 topic with step of 6. For Twitter data, the best topic number is 14 with coherence metrics score of 0.4197 as indicated in Figure 9.

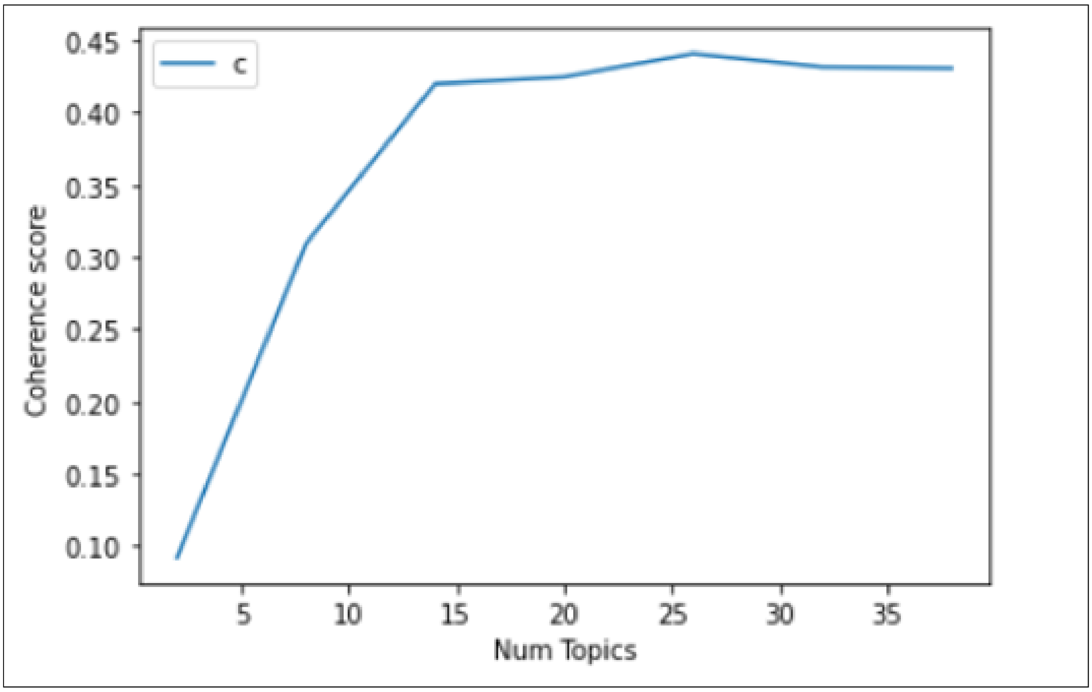


Figure 9: Coherence score for num of Topics

Intertopic Distance map is a visualization of the topics in a two-dimensional space . The area of these topic circles is corresponding to the volume of words that belong to each topic within the dictionary. The circles are plotted using a multidimensional scaling algorithm. it can be accessed that using the 14 selected topics without overlap introduced, in Figure 10 size of each topic suggests the volume of the each topic in the corpus.

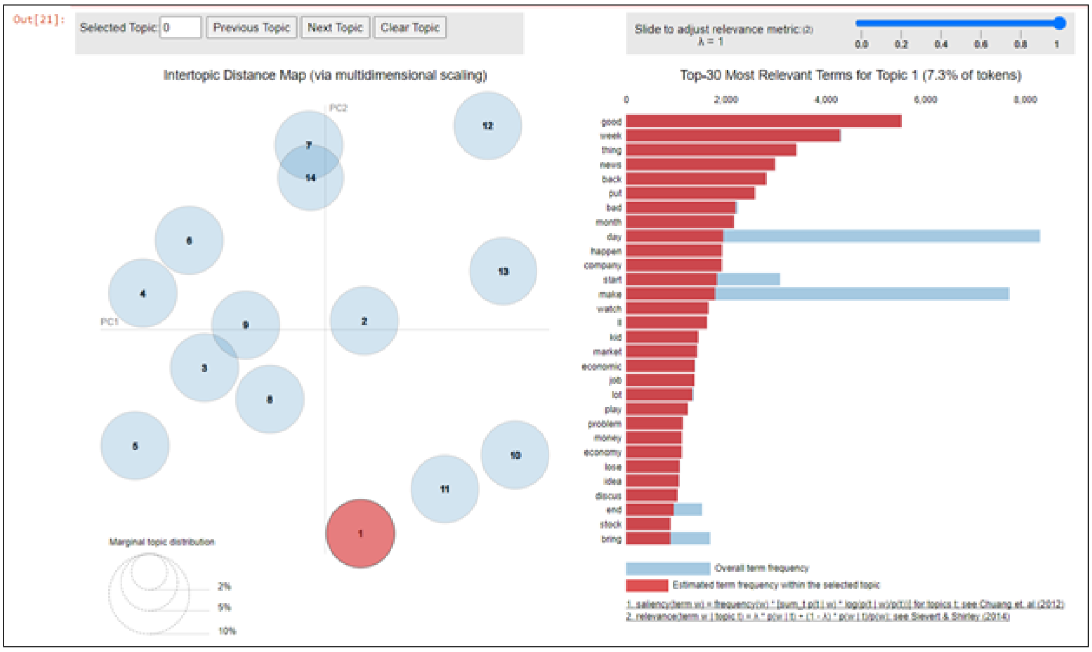
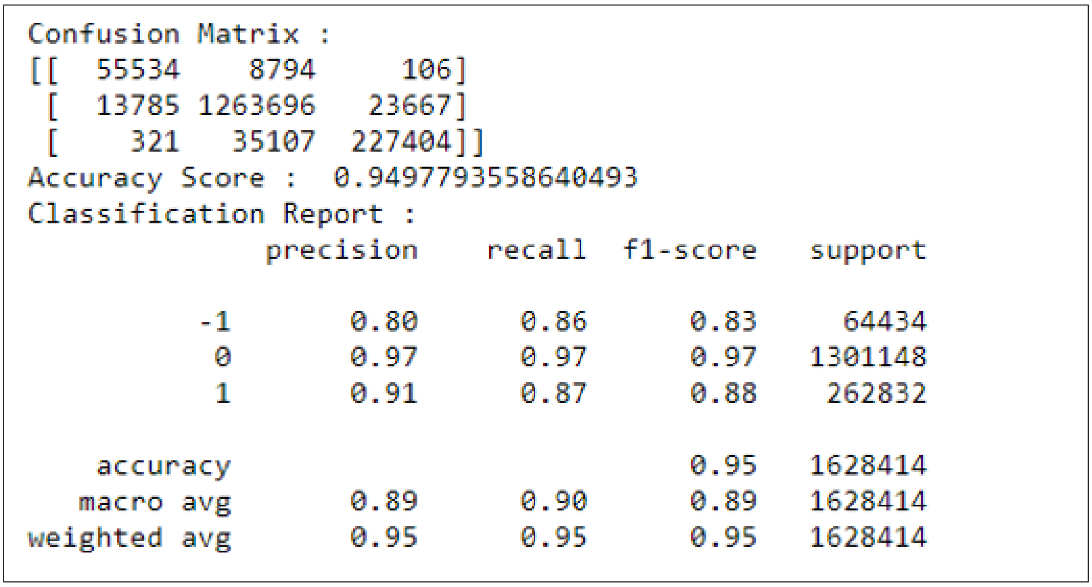


Figure 10: Intertopic Distance map

6.2 Sentiment Analysis Evaluation

The prediction (Figure 11) results show the accuracy has reached 94.9%. Breakdown to lower level, negative tweets has the 80% accuracy, neutural tweets has 97% accuracy and positive tweet has 91% accuracy. The higher percentage shows the close result before and after pre-processing.



```
Confusion Matrix :
[[  55534     8794     106]
 [  13785  1263696   23667]
 [    321    35107  227404]]
Accuracy Score :  0.9497793558640493
Classification Report :
              precision    recall  f1-score   support

          -1       0.80      0.86      0.83     64434
           0       0.97      0.97      0.97   1301148
           1       0.91      0.87      0.88    262832

    accuracy                           0.95   1628414
   macro avg       0.89      0.90      0.89   1628414
weighted avg       0.95      0.95      0.95   1628414
```

Figure 11: Confusion Matrix

It is found in our case that most of the sentiment for Covid-19 in tweets is in the neutual category. There are a few reasons contribute to this result. a) remove the stop word affect some of the sentiment calculation. b) Some text does not transfer into the sentiment. c) a number of texts in the same document contains positive and negative

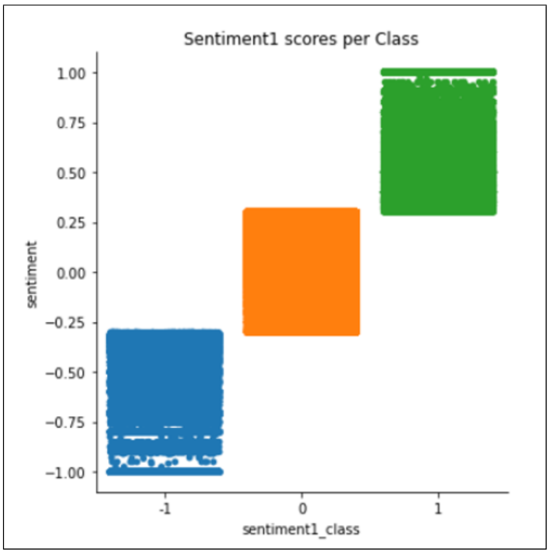


Figure 12: sentiment1

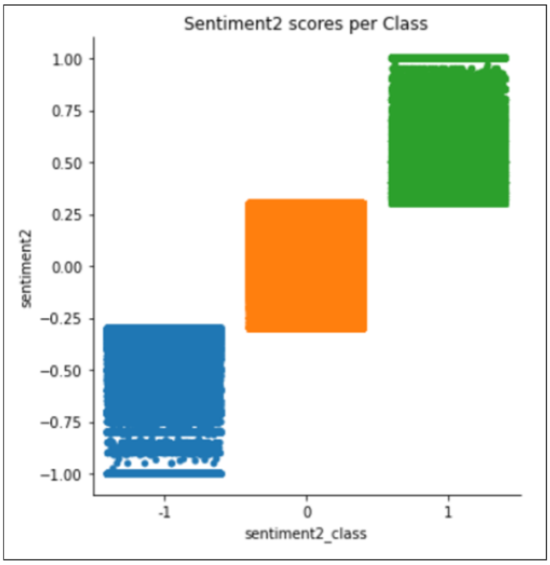


Figure 13: sentiment2

sentiment, which bring the result to neutral.

To measure the same amount of tweets message with and without the pre-processing, the paired t-test was introduced to judge whether the text pre-processing impact the sentiment result.

With the significance $\alpha = 0.05$, the paired t-test shows the statistics result = 98.349 which exceed the critical value from the t-statistics table. The null hypothesis(both statistics are equal) is rejected. This indicate the pre-processing has make the impact in relation to the calculation of the sentiment score.

The difference in the distribution of sentiment polarity score was plotted in the Figure 12 and Figure 13. Twitter category was plotted on the x-axis and the polarity score is scattered on the y-axis.

6.3 Stock market and Covid-19 sentiment

Either Pearson correlation or linear regression analysis is capable of determine whether the numeric variables are significantly linearly related or not. A correlation analysis presents feedback on the strength and direction of the linear relationship among the variables, also a linear regression analysis evaluate parameters in a linear equation that is possible to predict certain variable based on the other variables.

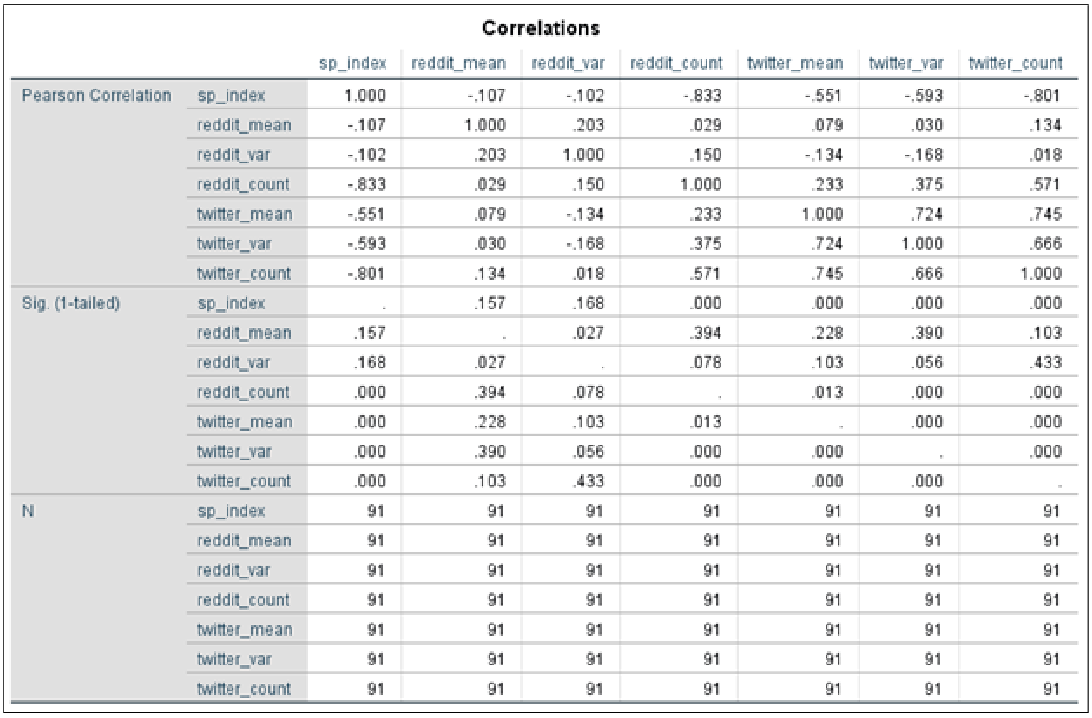
I have applied both Pearson Correlation and Linear Regression to analyse the correlation between the social media texts and the stock market movement.

6.3.1 Pearson Correlation

I used Pearson Correlation to estimate the social media impacts to the stock market return. Firstly, I built the model for SP500 by including the daily sentiment, variance and volume from the Twitter and Reddit. There are a number of approaches for generating a correlation value to measure correlation of multiple numerical variables, in order to find the insight about their relationship. The most popular one is Pearson Correlation Coefficient, it measures linear relationship among a list of variables. Pearson correlation ranges from -1 to +1, where +/-1 describes a perfect positive/negative correlation and

0 means no correlation. The r values between 0.50 to 0.75 and -0.50 to -0.75 indicate moderate to good correlation, and r values between 0.75 to 1 and from -0.75 to -1 point to very good to excellent correlation among the evaluated variables.

The result of the model revealed that indicators like Reddit volume, twitter daily average, twitter daily variance and volume are significant related to the return of SP500. Among these four variables, Twitter daily average and Twitter daily variance shows the moderate to good correlation, while Reddit volume and Twitter volume show very good to excellent correlation. All these variables are showing the negative relation to the SP500, indicating that the increase in the sentiment implies a fall in comments results in a positive return in SP500. Figure 14 presented a correlation of the sentiment variable against SP500 index.



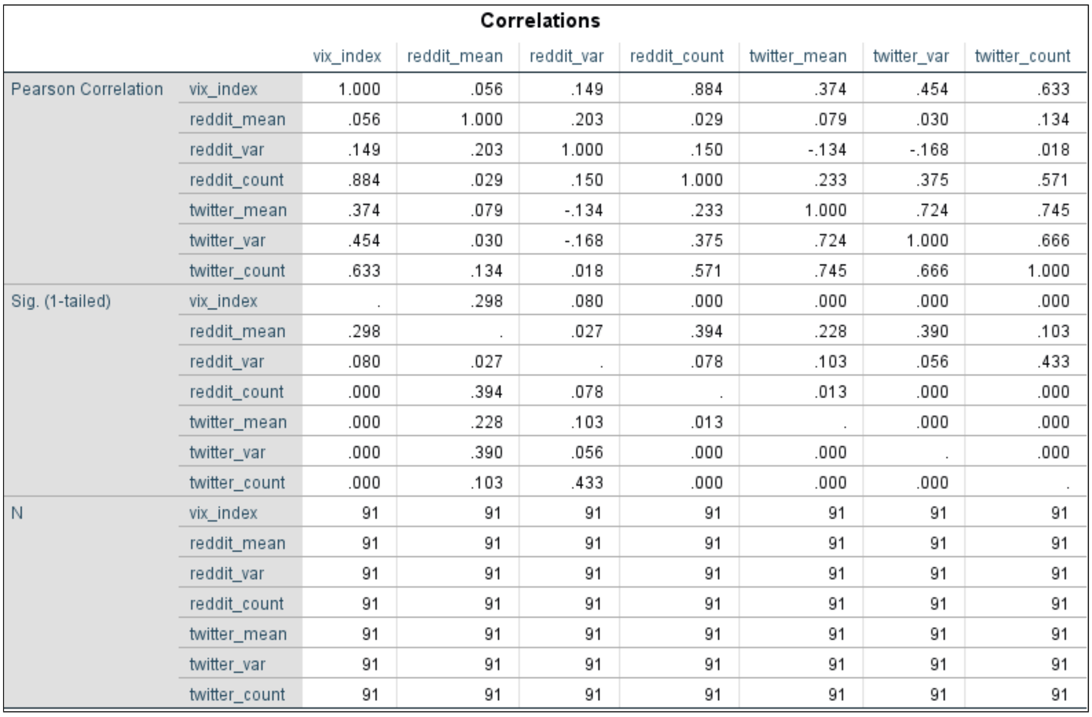
Correlations

		sp_index	reddit_mean	reddit_var	reddit_count	twitter_mean	twitter_var	twitter_count
Pearson Correlation	sp_index	1.000	-.107	-.102	-.833	-.551	-.593	-.801
	reddit_mean	-.107	1.000	.203	.029	.079	.030	.134
	reddit_var	-.102	.203	1.000	.150	-.134	-.168	.018
	reddit_count	-.833	.029	.150	1.000	.233	.375	.571
	twitter_mean	-.551	.079	-.134	.233	1.000	.724	.745
	twitter_var	-.593	.030	-.168	.375	.724	1.000	.666
	twitter_count	-.801	.134	.018	.571	.745	.666	1.000
Sig. (1-tailed)	sp_index	.	.157	.168	.000	.000	.000	.000
	reddit_mean	.157	.	.027	.394	.228	.390	.103
	reddit_var	.168	.027	.	.078	.103	.056	.433
	reddit_count	.000	.394	.078	.	.013	.000	.000
	twitter_mean	.000	.228	.103	.013	.	.000	.000
	twitter_var	.000	.390	.056	.000	.000	.	.000
	twitter_count	.000	.103	.433	.000	.000	.000	.
N	sp_index	91	91	91	91	91	91	91
	reddit_mean	91	91	91	91	91	91	91
	reddit_var	91	91	91	91	91	91	91
	reddit_count	91	91	91	91	91	91	91
	twitter_mean	91	91	91	91	91	91	91
	twitter_var	91	91	91	91	91	91	91
	twitter_count	91	91	91	91	91	91	91

Figure 14: Correlations with SP500 index

Secondly, I built the model for VIX index respectively by including the sentiment variables. The result of the model revealed that indicators like Reddit volume, twitter volume are significant related to the movement of VIX. Among these two variables, Twitter volume shows the moderate to good correlation, while Reddit volume shows very good to excellent correlation. All these variables are showing the positive relation to the VIX index, indicating that the increase in the Twitter and Reddit volume results in a positive movement of VIX index. Figure 15 presented a correlation of the sentiment variable against VIX index.

Thirdly, I built the model for FSI index respectively by including the sentiment variables. The result of the model revealed that indicators like Reddit volume, twitter daily average, twitter daily variance and twitter volume are significant related to the movement of FSI. Among these four variables, Twitter daily average, Twitter daily variance shows the moderate to good correlation, while Twitter volume and Reddit volume shows very good to excellent correlation. All these variables are showing the positive relation to the FSI index, indicating that the increase in the Twitter and Reddit volume results

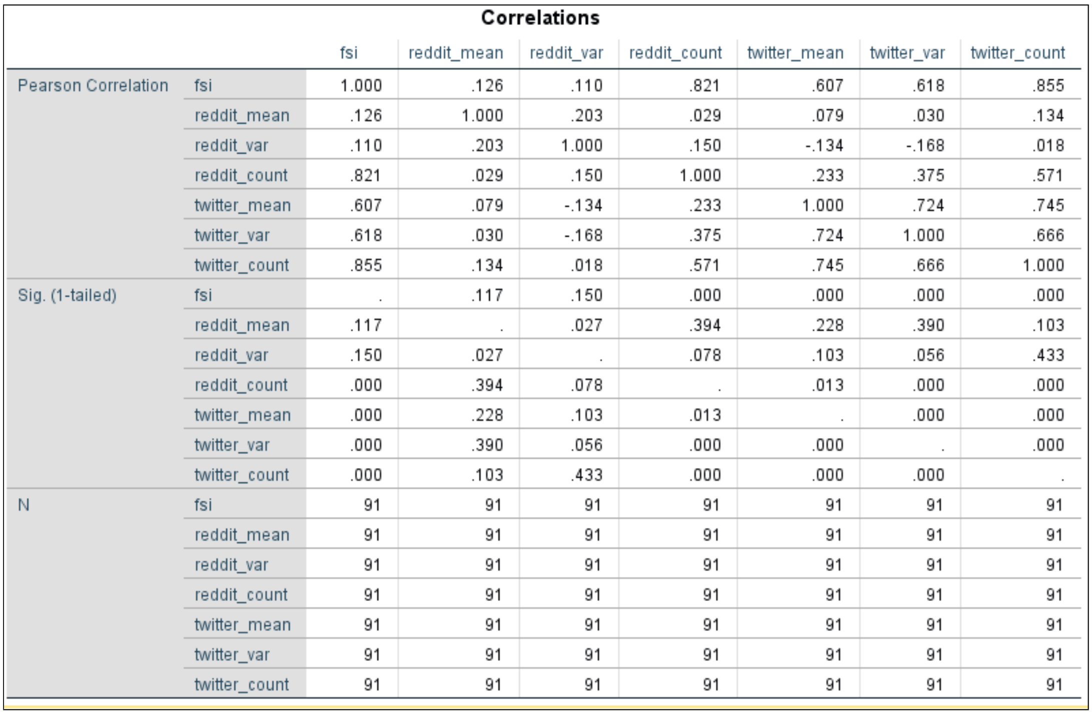


Correlations

		vix_index	reddit_mean	reddit_var	reddit_count	twitter_mean	twitter_var	twitter_count
Pearson Correlation	vix_index	1.000	.056	.149	.884	.374	.454	.633
	reddit_mean	.056	1.000	.203	.029	.079	.030	.134
	reddit_var	.149	.203	1.000	.150	-.134	-.168	.018
	reddit_count	.884	.029	.150	1.000	.233	.375	.571
	twitter_mean	.374	.079	-.134	.233	1.000	.724	.745
	twitter_var	.454	.030	-.168	.375	.724	1.000	.666
	twitter_count	.633	.134	.018	.571	.745	.666	1.000
Sig. (1-tailed)	vix_index	.	.298	.080	.000	.000	.000	.000
	reddit_mean	.298	.	.027	.394	.228	.390	.103
	reddit_var	.080	.027	.	.078	.103	.056	.433
	reddit_count	.000	.394	.078	.	.013	.000	.000
	twitter_mean	.000	.228	.103	.013	.	.000	.000
	twitter_var	.000	.390	.056	.000	.000	.	.000
	twitter_count	.000	.103	.433	.000	.000	.000	.
N	vix_index	91	91	91	91	91	91	91
	reddit_mean	91	91	91	91	91	91	91
	reddit_var	91	91	91	91	91	91	91
	reddit_count	91	91	91	91	91	91	91
	twitter_mean	91	91	91	91	91	91	91
	twitter_var	91	91	91	91	91	91	91
	twitter_count	91	91	91	91	91	91	91

Figure 15: Correlations with VIX index

in a positive movement of FSI index. Figure 16 presented a correlation of the sentiment variable against FSI index.



Correlations

		fsi	reddit_mean	reddit_var	reddit_count	twitter_mean	twitter_var	twitter_count
Pearson Correlation	fsi	1.000	.126	.110	.821	.607	.618	.855
	reddit_mean	.126	1.000	.203	.029	.079	.030	.134
	reddit_var	.110	.203	1.000	.150	-.134	-.168	.018
	reddit_count	.821	.029	.150	1.000	.233	.375	.571
	twitter_mean	.607	.079	-.134	.233	1.000	.724	.745
	twitter_var	.618	.030	-.168	.375	.724	1.000	.666
	twitter_count	.855	.134	.018	.571	.745	.666	1.000
Sig. (1-tailed)	fsi	.	.117	.150	.000	.000	.000	.000
	reddit_mean	.117	.	.027	.394	.228	.390	.103
	reddit_var	.150	.027	.	.078	.103	.056	.433
	reddit_count	.000	.394	.078	.	.013	.000	.000
	twitter_mean	.000	.228	.103	.013	.	.000	.000
	twitter_var	.000	.390	.056	.000	.000	.	.000
	twitter_count	.000	.103	.433	.000	.000	.000	.
N	fsi	91	91	91	91	91	91	91
	reddit_mean	91	91	91	91	91	91	91
	reddit_var	91	91	91	91	91	91	91
	reddit_count	91	91	91	91	91	91	91
	twitter_mean	91	91	91	91	91	91	91
	twitter_var	91	91	91	91	91	91	91
	twitter_count	91	91	91	91	91	91	91

Figure 16: Correlations with FSI index

Table 2 shows the summary of the Pearson correlation for the SP500, VIX and FSI.

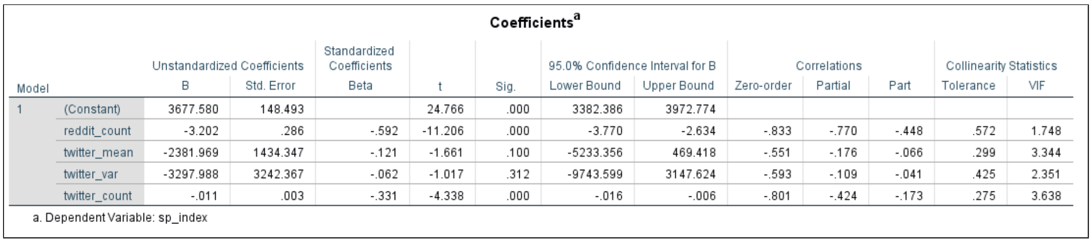
Index	**Good to excellent correlation**	**Moderate to good correlation**
SP500	Reddit volume, Twitter volume	Twitter daily average, Twitter daily variance
VIX	Reddit volume	Twitter volume
FSI	Twitter volume, Reddit volume	Twitter daily average, Twitter daily variance

Table 2: Summary of Index correlation

6.3.2 Linear Regression

Besides the Pearson correlation, I also investigate the linear regression result.for the p-value. The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (¡ 0.05) indicates that I can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to my model because changes in the predictor's value are related to changes in the response variable.

Figure 14 shows the Reddit Count, Twitter average, Twitter variance and Twitter Count are significant to the SP500, as all these P-values are less than 0.05. After rerun the linear regression with the above variables, coefficient matrix has been recreated. Figure 17 show the coefficients matrix that all these variables show the negative relationship to the Sp500 index.



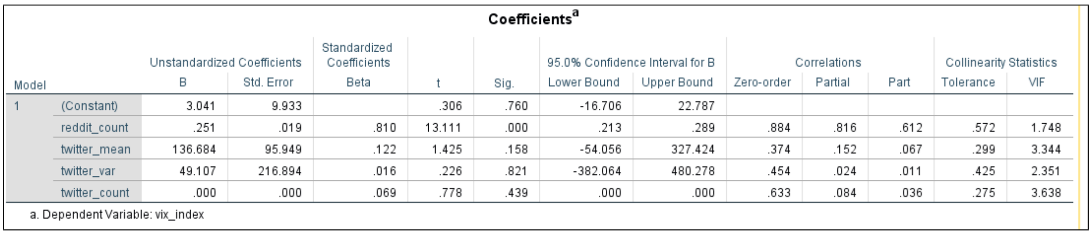
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	3677.580	148.493		24.766	.000	3382.386	3972.774					
	reddit_count	-3.202	.286	-.592	-11.206	.000	-3.770	-2.634	-.833	-.770	-.448	.572	1.748
	twitter_mean	-2381.969	1434.347	-.121	-1.661	.100	-5233.356	469.418	-.551	-.176	-.066	.299	3.344
	twitter_var	-3297.988	3242.367	-.062	-1.017	.312	-9743.599	3147.624	-.593	-.109	-.041	.425	2.351
	twitter_count	-.011	.003	-.331	-4.338	.000	-.016	-.006	-.801	-.424	-.173	.275	3.638

a. Dependent Variable: sp_index

Figure 17: Coefficients with SP500 index

Figure 15 shows the Reddit Count, Twitter average, Twitter variance and Twitter Count are significant to the VIX index, as all these P-values are less than 0.05. After rerun the linear regression with the above variables, coefficient matrix has been recreated. Figure 18 show the coefficients matrix that all these variables show the positive relationship to the VIX index.



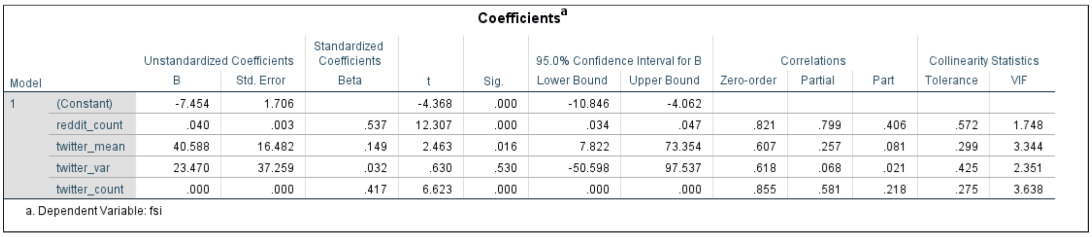
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	3.041	9.933		.306	.760	-16.706	22.787					
	reddit_count	.251	.019	.810	13.111	.000	.213	.289	.884	.816	.612	.572	1.748
	twitter_mean	136.684	95.949	.122	1.425	.158	-54.056	327.424	.374	.152	.067	.299	3.344
	twitter_var	49.107	216.894	.016	.226	.821	-382.064	480.278	.454	.024	.011	.425	2.351
	twitter_count	.000	.000	.069	.778	.439	.000	.000	.633	.084	.036	.275	3.638

a. Dependent Variable: vix_index

Figure 18: Coefficients with VIX index

Figure 16 shows the Reddit Count, Twitter average, Twitter variance and Twitter Count are significant to the FSI index, as all these P-values are less than 0.05. After rerun the linear regression with the above variables, coefficient matrix has been recreated. Figure 19 show the coefficients matrix that all these variables show the positive relationship to the FSI index.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-7.454	1.706		-4.368	.000	-10.846	-4.062					
	reddit_count	.040	.003	.537	12.307	.000	.034	.047	.821	.799	.406	.572	1.748
	twitter_mean	40.588	16.482	.149	2.463	.016	7.822	73.354	.607	.257	.081	.299	3.344
	twitter_var	23.470	37.259	.032	.630	.530	-50.598	97.537	.618	.068	.021	.425	2.351
	twitter_count	.000	.000	.417	6.623	.000	.000	.000	.855	.581	.218	.275	3.638

a. Dependent Variable: fsi

Figure 19: Coefficients with FSI index

The linear regression result is very close to the Pearson correlation result. The list of the variables that correlated to the index are matching between these two approaches. It is further confirmed that there is a strong relationship between the index and the sentiment data collected from the social media platform.

6.4 Discussion

In this section, I am going to assess our model. Our study indicated that incorporating LDA features does have a positive effect in find the optimised number of topics. The topic coherence score provided the measurement for each model tuning. The Intertopic distance map has further proved this concept. LDA model has helped to identify the dominant topic in each sentence and the most representative document for each topic. Meantime I found a number of limitation on the LDA feature, such as the number of topics is fixed and must be known in advance, and Dirichlet topic distribution is not able to detect correlations.

Sentiment analysis on Twitter data presents limitation that the majority of the tweets are falling into the neutral sentiment. The distribution across the sentiments groups are unbalanced. This result the difficult in extracting the accurate emotional from people given the most of the tweets sentiment fall in the neutral category.

It is discovered that there is a connection between stock market movement and sentiment collected from Twitter and Reddit. Both Pearson correlation and linear regression model has shown the connection between the SP500/VIX/FSI and the Tweet volume, Reddit mean, Reddit variance and Reddit volume at the daily level. This is aligned with the found presented in Costola et al. (2020)

7 Conclusion

This research project emphasize on investigating the social media text using NLP techique for text modeling and achieve the more accurated sentiment score, as well as check the stock market impact based on the extracted sentiment. For the delivery of this project, all the objectives listed in the section 1.2. Objectives 1 to objective 4 were achieved by the literature review, data collection, text pre-processing and exploratory analysis. In order to resolve the SubRQ, the LDA topic model was designed to achieve the objective 5. Objective 6 was supported by the sentiment analysis applied in the section 3.4. Objective 7 of analysing the impact of test data pre-processing on polarity calculation was achieved in section 5.2. In order to tackle the RQ, the analysis of the linear regression model was applied to meet the objective 8.

Based on the analysis presented in the previous section, the following main results can be concluded: First, the impact of Covid-19 on the stock market shows significant effect in the U.S. market. When the stock market volatility increases, COVID-19 create a greater effect on the stock market fluctuation. Second, Latent Dirichlet allocation(LDA) as an unsupervised topic modeling has improved the exploration of the topic information and resulting content clusters by tuning the LDA parameters. I used the Mallet and Gensim's standard LDA to find the most suitable amount of topics and match the topic for each sentence I processed. Third, Pre-processing of the text has shown the clear evidence of the improvement of sentiment prediction by providing the more accurate polarity score.

This research contributes to the improve knowledge in the domain of financial impact base on the crisis by analysing the stock market movement with Covid-19 sentiment analysis. Covid-19 brought the area for research the association with various industrial including financial domain. These results are important in relation to the creation of emergency management approachs, in reaction to major public health crisis, and for the future management of the financial market in U.S.

Challenges in terms of time brought a few limitations to this project. Some other techniques can be applied to analyse the stock market data with the sentiment data, e.g Principle component analysis(PCA) can be applied to run the dimensionally-reduction algorithm to locate the key sentiment variables for the predict the stock market movement.

8 Acknowledgement

I would like to thank my supervisor, Dr. Majid Latifi, for his guidance and support during this research project. Dr. Majid Latifi has provided me with his domain knowledge in the NLP area and help me with my learning process. Personally, I would like to thank my family for helping me to complete this project.

References

Aslam, F., Awan, T. M., Syed, J. H., Kashif, A. and Parveen, M. (2020). Sentiments and emotions evoked by news headlines of coronavirus disease (COVID-19) outbreak, *Humanities and Social Sciences Communications* **7**(1): 1–10.
URL: *http://dx.doi.org/10.1057/s41599-020-0523-3*

Baker, S. R., Bloom, N., Davis, S. J., Kost, K., Sammon, M. and Viratyosin, T. (2020). The Unprecedented Stock Market Reaction to COVID-19, *The Review of Asset Pricing Studies* **10**(4): 742–758.
URL: *https://doi.org/10.1093/rapstu/raaa008*

Beigi, G., Hu, X., Maciejewski, R. and Liu, H. (2016). An overview of sentiment analysis in social media and its applications in disaster relief, *Studies in Computational Intelligence* **639**(February 2016): 313–340.

Bold, A. (2019). Sentiment analysis - the lexicon based approach.
URL: *https://www.alphabold.com/sentiment-analysis-the-lexicon-based-approach/*

Bondielli, A. and Marcelloni, F. (2019). A survey on fake news and rumour detection techniques, *Information Sciences* **497**.

Castillo, C., Mendoza, M. and Poblete, B. (2011). Information credibility on twitter, p. 675–684.
URL: *https://doi.org/10.1145/1963405.1963500*

Costola, M., Iacopini, M. and Santagiustina, C. R. M. A. (2020). Public concern and the financial markets during the covid-19 outbreak.

Daniel, K., Hirshleifer, D. and Subrahmanyam, A. (1998). Investor psychology and security market under- and overreactions, *The Journal of Finance* **53**(6): 1839–1885.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/0022-1082.00077*

Dattu, B. and Gore, D. (2015). A Survey on Sentiment Analysis on Twitter Data Using Different Techniques, *International Journal of Computer Science and Information Technologies (IJCSIT)* **6**(6): 5358–5362.

George, L. E. and Birla, L. (2019). A Study of Topic Modeling Methods, *Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems, ICICCS 2018* (Iciccs): 109–113.

Gormsen, N. J. and Koijen, R. S. J. (2020). Coronavirus: Impact on stock prices and growth expectations, *Working Paper 27387*, National Bureau of Economic Research.
URL: *http://www.nber.org/papers/w27387*

Helsloot, I. and Ruitenberg, A. (2004). Citizen response to disasters: A survey of literature and some practical implications, *Journal of Contingencies and Crisis Management* **12**(3): 98–111.

Hobson Lane, Cole Howard, H. H. (2019). *Natural Language Processing in Action*, Manning Publications.

Hong, H. and Stein, J. C. (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets, *The Journal of Finance* **54**(6): 2143–2184.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1111/0022-1082.00184*

Huo, X. and Qiu, Z. (2020). How does china's stock market react to the announcement of the covid-19 pandemic lockdown?, *Economic and Political Studies* **8**(4): 436–461.
URL: *https://doi.org/10.1080/20954816.2020.1780695*

Jelodar, H., Wang, Y., Orji, R. and Huang, H. (2020). Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach, *arXiv* **24**(10): 2733–2742.

Kaewkitipong, L., Chen, C. and Ractham, P. (2012). Lessons learned from the use of social media in combating a crisis: A case study of 2011 thailand flooding disaster, *International Conference on Information Systems, ICIS 2012* **1**: 766–782.

Koppel, M. and Schler, J. (2006). The importance of neutral examples for learning sentiment, *Computational Intelligence* **22**(2): 100–109.

Mamaysky, H. (2020). Financial markets and news about the coronavirus.
URL: *https://voxeu.org/article/financial-markets-and-news-about-coronavirus*

Mittal, A. and Patidar, S. (2019). Sentiment analysis on twitter data: A survey.
URL: *https://doi.org/10.1145/3348445.3348466*

Mohammad, S. and Turney, P. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon, *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* .

Nemes, L. and Kiss, A. (2021). Prediction of stock values changes using sentiment analysis of stock news headlines, *Journal of Information and Telecommunication* **0**(0): 1–20.
URL: *https://doi.org/10.1080/24751839.2021.1874252*

Oyebode, O., Ndulue, C., Mulchandani, D., Suruliraj, B., Adib, A., Orji, F. A., Milios, E., Matwin, S. and Orji, R. (2020). COVID-19 pandemic: Identifying key issues using social media and natural language processing, *arXiv* .

Paredes-Valverde, M. A., Colomo-Palacios, R., Salas-Zárate, M. D. P. and Valencia-García, R. (2017). Sentiment Analysis in Spanish for Improvement of Products and Services: A Deep Learning Approach, *Scientific Programming* **2017**.

Taboada, M., Brooke, J. and Voll, K. (2011). Lexicon-Based Methods for Sentiment Analysis, (August 2010).

Tang, L., Bie, B., Park, S. E. and Zhi, D. (2018). Social media and outbreaks of emerging infectious diseases: A systematic review of literature, *American Journal of Infection Control* **46**(9): 962–972.
URL: *https://doi.org/10.1016/j.ajic.2018.02.010*

Thelwall, M. and Buckley, K. (2012). Topic-Based Sentiment Analysis for the Social Web: The role of Mood and Issue-Related Words, pp. 1–17.

Traylor, T., Straub, J., Gurmeet and Snell, N. (2019). Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator, *Proceedings - 13th IEEE International Conference on Semantic Computing, ICSC 2019* pp. 445–449.

Valecha, R., Oh, O. and Raghav Rao, H. (2013). An exploration of collaboration over time in collective crisis response during the Haiti 2010 earthquake, *International Conference on Information Systems (ICIS 2013): Reshaping Society Through Information Systems Design* **1**: 378–387.

Valle-Cruz, D., Fernandez-Cortez, V., López-Chau, A. and Sandoval-Almazán, R. (2021). Does Twitter Affect Stock Market Decisions? Financial Sentiment Analysis During Pandemics: A Comparative Study of the H1N1 and the COVID-19 Periods, *Cognitive Computation* (0123456789).
URL: *https://doi.org/10.1007/s12559-021-09819-8*

Volkova, S., Shaffer, K., Jang, J. Y. and Hodas, N. (2017). Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter, *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* **2**(August): 647–653.

Wang, W. Y. (2017). “liar, liar pants on fire”: A new benchmark dataset for fake news detection, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, Vancouver, Canada, pp. 422–426.
URL: *https://www.aclweb.org/anthology/P17-2067*

Xiong, H., Wu, Z., Hou, F. and Zhang, J. (2020). Which firm-specific characteristics affect the market reaction of chinese listed companies to the covid-19 pandemic?, *Emerging Markets Finance and Trade* **56**(10): 2231–2242.
URL: *https://doi.org/10.1080/1540496X.2020.1787151*

Xu, Y. and Keselj, V. (2019). Stock Prediction using Deep Learning and Sentiment Analysis, *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019* pp. 5573–5580.

Yi, J., Nasukawa, T., Bunescu, R. and Niblack, W. (2003). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques, *Proceedings - IEEE International Conference on Data Mining, ICDM* pp. 427–434.

Zeng-Treitler, Q., Goryachev, S., Tse, T., Keselman, A. and Boxwala, A. (2008). Estimating Consumer Familiarity with Health Terminology: A Context-based Approach, *Journal of the American Medical Informatics Association* **15**(3): 349–356.