İSTANBUL BİLGİ UNIVERSITY

INSTITUTE OF GRADUATE PROGRAMS

INTERNATIONAL FINANCE MASTER'S DEGREE PROGRAM

PREDICTION OF S&P500 STOCK MARKET MOVEMENT USING TWITTER SENTIMENT ANALYSIS

OMAR HANNAN HORO

118664018

Prof. Dr. Cenktan ÖZYILDIRIM

ISTANBUL

2021

Prediction of S&P500 Stock Market Movement Using Twitter Sentiment Analysis

Twitter Duyarlılık Analizi Kullanarak S&P500 Hisse Senetleri Hareketi Tahmini

Omar Hannan Horo
118664018

Tez Danışmanı: Prof. Dr. Cenktan Özyıldırım (imza) .....
İstanbul Bilgi Üniversitesi

Jüri Üyesi: Dr. Öğr. Üyesi Ebru Reis (imza) .....
İstanbul Bilgi Üniversitesi

Jüri Üyesi: Prof. Dr. Yaman Ömer Erzurumlu (imza) .....
Bahçeşehir Üniversitesi

Tezin Onaylandığı Tarih: 28/09/2021
Toplam sayfa sayısı: 66

Anahtar Kelimeler (Türkçe)

1) ARCH modeli
2) Duygu Analizi
3) GARCH modeli
4) Ekonometri
5) Oynaklık

Anahtar Kelimeler (İngilizce)

1) ARCH Model
2) Sentiment Analysıs
3) GARCH MODEL
4) Econometrics
5) Volatility

DECLARATION

I hereby declare to be the Author of this dissertation and in no way does my work relate or incorporate with any of the previous work submitted to either this Institution or any other. Hence to my best of knowledge, this Dissertation neither contains any submitted resources and if so, then the document related has been cited to avoid any copyright infringement.

Therefore, I also agree that my dissertation will be revised through plagiarism tools and any failure to comply authenticity may result in consequences.

Omar Hannan Horo

ACKNOWLEDGMENT

First and foremost, I would love to extend my gratitude to my research advisor Prof. Cenktan Özyıldırım, for the full support and commitment he had through this journey. His valuable advice, selflessness and always being available both during working and non-working throughout my journey during the thesis period has been nothing but extraordinary. Lastly,

Special thank you to a few of my colleagues for their support, motivation and the time they put in to work with me through my dissertation. Every piece of information they provided has been nothing but vital to this dissertation.

More importantly, I would love to thank my parents and family as whole, who have been with me not only during this period, but also through every step of my life towards the completion of what I am currently trying to achieve. Their struggle, prayers and kindness have given enough energy to push through different barriers and none of what I am trying to achieve currently would have been possible if it was not for them

Thank you all and may God bless you.

Omar Hannan Horo

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

S&P500	: Standard & Poor 500 Stock Index
FTSE	: The Financial Times Stock Exchange
SVM	: Support Vector Machine
NASDAQ	: National Association of Securities Dealers Automate Quotations
KSE	: Karachi Stock Exchange
NLP	: Natural Language Processing
EMH	: Efficient Market Hypothesis
BoW	: Bag of Words
LWL	: Locally Weighted Learning
NB	: Naïve Bayes
SMO	: Sequential Minimal Optimization
KNN	: K-Nearest Neighbors
SSR	: Sum of Square of Residuals
SST	: Sum of Square Total
NLTK	: Natural Language Tool Kit
NYSE	: New York Stock Exchange
GARCH	: Generalized Autoregressive Conditional Heteroskedasticity
ARCH	: Autoregressive Conditional Heteroskedasticity
AIC	: Akaike Information Criterion
OLS	: Ordinary Least Squares

ABSTRACT

PREDICTION OF S&P500 STOCK MARKET MOVEMENT USING TWITTER SENTIMENT ANALYSIS

Twitter is amongst the biggest social networking services, offering its services to over 350 million active users as of 2021. For over the past few years, Twitter has been the most used social network platform as a tool of interaction between investors and stock market given its convenience and ability to move information from one part to another in a very short time.

For over years, econometrics had always turn to the basic Ordinary Least Regression (OLS) to analyze the relationship between two variables and how much impact one variable could have another. However, during OLS analysis, the assumption is that the squared values of all errors terms at any particular point is constant (Homoskedasticity). Therefore, for a given set of data in a time series with different error terms (Heteroskedasticity) may result to a slight sense of false analysis and forecasting precision, hence leading us to go with ARCH and GARCH models, where heteroskedasticity is basically considered as variance to be accounted for and modeled.

The aim of this study was to determine whether there is a significance impact of sentiment to the stock return in the time series. Despite several related studies, most papers mostly used Ordinary Least Squares with an assumption of linearity. Despite the original OLS showing significant influence of sentiments in our model, ARCH and GARCH models concluded that sentiment was not a significant predictor in the conditional variance.

Keywords: Generalized Autoregressive Conditional Heteroskedasticity, Sentiment Analysis, Autoregressive Conditional Heteroskedasticity model, Econometrics, Volatility.

ÖZET

TWITTER DUYARLILIK ANALİZİ KULLANARAK S&P500 HİSSE SENETLERİ HAREKETİ TAHMİNİ

Twitter, 2021 yılı itibarıyla 350 milyonu aşkın aktif kullanıcıya hizmet sunan en büyük sosyal ağ servisleri arasında yer almaktadır. Twitter, geçtiğimiz yıllarda yatırımcılar ve borsa arasında bir etkileşim aracı olarak en çok kullanılan sosyal ağ platformu olmuştur. Bilgileri çok kısa sürede bir parçadan diğerine taşıma kolaylığı ve yeteneği olan bir platformudur.

Yıllar boyunca, ekonometri, iki değişken arasındaki ilişkiyi ve bir değişkenin diğerini ne kadar etkileyebileceğini analiz etmek için her zaman temel Sıradan En Az Regresyona (OLS) yöneldi. Bununla birlikte, OLS analizi sırasında, herhangi bir noktadaki tüm hata terimlerinin kare değerlerinin sabit olduğu varsayımıdır (Homoskedastisite). Bu nedenle, farklı hata terimlerine (Heteroskedastisite) sahip bir zaman serisindeki belirli bir veri seti için, hafif bir yanlış analiz ve tahmin kesinliği hissi ile sonuçlanabilir, bu nedenle bizi değişen varyanslılığın temelde varyans olarak kabul edildiği ARCH ve GARCH modellerine yönelmemize neden olabilir, hesaplanacak ve modellenecektir.

Bu çalışmanın amacı, zaman serilerinde duyarlılığın hisse senedi getirisi üzerinde anlamlı bir etkisinin olup olmadığını belirlemektir. Birkaç ilgili çalışmaya rağmen, çoğu makale çoğunlukla doğrusallık varsayımıyla Sıradan En Küçük Kareleri kullandı. Modelimizde duyguların önemli etkisini gösteren orijinal OLS'ye rağmen, ARCH ve GARCH modelleri, duyarlılığın koşullu varyansta önemli bir tahmin edici olmadığı sonucuna varmıştır.

Anahtar Kelimeler: ARCH modeli, Duygu Analizi, GARCH modeli, Ekonometri, Oynaklık.

INTRODUCTION

1.1 BACKGROUND OF THE STUDY

The study on how sentiments from social media platforms affect the movement of stocks in the market has become one of the leading studies. There are several ways on how the stock market has been defined by scholars. Sunders (2003) defines the stock market as simply as institution where money flows. Stock Market Exchange and Stock trading are two terms possibly the most outspoken words in the current digital world of today. Abbood (2016) also defined the stock market as Investing, a way that provides one an opportunity to pursue other things while being able to grow their capital at the same time.

Like any other field, Stock markets have been widely acknowledged to be driven by real unpredictable and predictable events, hence making them so volatile and complex (Duong et. al. 2016). Lue (2010) argued that one of the most complex aspect about the nature of the stock market is whether is linear or non-linear. This in turn, has become one of the biggest problems stock traders face. Al-Augby (2015) argues that despite of the nature of complexity of the stock market, it is still one of the most important sectors in the economic field.

Turner (2007) and Murphy (1999) pointed two of the traditional ways traders used to predict the market namely Technical analysis and Fundamental analysis respectively. Technical analysis covered predicting the stock market through patterns and past stock's prices, whereas Fundamental analysis basically covered the company's basic information.

In the last decade or so, social media platforms have become a home for the public emotions in the every aspect of life. Platforms like Twitter and Facebook having handed the public to express their emotions through different ways such as posting, image sharing as well as the use of emoji. As of current, both investors and stock market use twitter mostly to share news and public emotions. B. Jansen et al. (2009) discussing the power of Twitter depicted that over 180 million tweets are posted by millions of users everyday making the platform an entity with full of valuable data. Such

information extracted from the tweets tend to provide significant information making the data very convenient during prediction process (A. Pak and P. Paroubek, 2010).

Hence, many scholars have found a greater significance in trying to study to what extent public opinion can affect the movement of stock prices in the market. Forecasting and precision of stock market volatility has become very significant for financial institutions and investors. Sutheebanjard and Premchaiswadi (2010) defined the stock market as a non-linear system which is affected by different factors including inflations, social and political events etc. When it comes to the stock market, volatility is the most important factor considered when trying to forecast the stock return and volatility.

Panait and Slăvescu (2012) labelled volatility as an extreme vital variable in stock returns speculation as well as credibility of a given financial institution. Islam et al. (2012) also indicated that volatility is very crucial for the government in its policy making as well as economic and financial sectors since it plays a big role to a country's economy. Ederington and Guan (2005) also pointed the important of volatility in the stock market especially to investors and traders who are trying to analyze the stock market movement. Stock market trading especially on indices has caught most of the traders' attention given its broad dimensions of investment opportunities it provides for traders to invest. This is why we considered focus our study on S&P500 indices rather than individual stock companies.

The most difficult task for econometrics was to predict both the stock return's mean and variance with regard to historical data. Although different analysis were made to predict the stock return and volatility using mean return, barely any study had been conducted using variance until the introduction of ARCH models by Engel (1982). Until the popularity of the ARCH models, error terms derived from the basic known Least Squares Regression were all under assumptions of having a constant variance over a time series.

Among the econometrics tools used in forecasting of volatility is the GARCH model, an extension to the ARCH model which was first introduced by Bollerslev (1987). The main reason to why GARCH model was used in our model study was because there was an existing correlation between

our variables. Hence using Garch Model, we will analyze whether sentiment scores has any effect on log return and whether the variable could be used in the GARCH to predict stock volatility.

1.2 PROBLEM STATEMENT

Stocks and Stock Market are terms that daily heard in the digital technology whether it is through news, social media or even through word of mouth. Kang Zhang et al. (2019) simply defined the stock market as an area where buying and selling can set one's future. Despite all this, the market isn't as straightforward and simple as most think. Duong (2016) depicted the stock market to be of the one of the toughest task to execute given its volatility.

In our study, we tried to focus on trying to find the correlation between our Twitter data and S&P 500 stocks on both stock market level and company stock level using the following stock indicators: closing price, stock return, and daily trading volume. In our dissertation we will study sentiment analysis and machine learning techniques to answer the following questions:

i. Is there any correlation between the variables i.e. sentiment scores and log return
ii. Is the data stationary and linear in the time series? And if so how can we be able to analyze different asymmetric effect the sentiment score has on stock volatility
iii. Forecasting of volatility using ARCH GARCH models and sentiment scores.

1.3 SIGNIFICANCE OF THE RESEARCH

Weston et al. (1996) simply defined the stock market as simply a structure where people can lend or borrow money. Hence given the importance of money to the society, we can already see how significant this research could be. In the recent years, Stock trading has been touted as one of the most if not the most profitable business out there if done with precision and patience. However, with the growth of social media platforms as means of communication, most of the traders have been at loss, blaming it all on the volatility of stocks. Even despite its complexity, Al-Augby (2015) dictated that the stock market was and is still one of the most important issue in the economy sector.

The stock markets' importance is not only to the companies but also the experts as well as common people (Nayak et al. 2016). Companies' economic value tend to be measured according to their stock market value (Kumar & Ravi, 2016) whereas common people and experts tend to see the stock market as a money making machine, making it such a significant study for scholars.

The creation and rise of social media platform like Twitter, a social media platform mostly used for quick information digestion, has enabled an even faster interaction between stock market investors. Brown and Cliff (2004) argued the subject on if the sentiments by investors on the prices of the stocks is crucial since the sentiments may contribute to market bursts and large recessions.

Daniel et al (2017) previously discussed how vital people's sentiments are towards decision making. This brings out the significance of sentiment analysis towards prediction of any movement of the stock market. Sentiment analysis using NLP is able to classify any given statement to find out whether the statement is positive, neutral or negative using polarity (Aqel &Vadera, 2013). Pang and Lee (2004) continue to argue that the ability of sentiment analysis given its ability to classify any given statement makes it very significant.

1.4 OBJECTIVES OF THE RESEARCH

This study was conducted to observe whether there was any correlation of sentiment scores and stock return and whether we could use the sentiment scores to forecast the stock volatility of S&P500 indices using sentiment scores we extracted and analyzed from Twitter data. Other objectives of the dissertation include the following:

- Determine whether data in a time series is stationary or not.
- To determine whether the autocorrelation between our variables (sentiment scores and log return and whether sentiment scores has any effect on log return and volatility.
- To determine the limitations and setbacks of the research. And if found, to discuss what could be done to get better and sufficient results

1.5 STRUCTURE OF THE RESEARCH

The dissertation paper will consist of five (5) chapters namely Introduction, Literature Review, Data and Methodology, Findings and Conclusion. The introduction section will cover a detailed study research background, the purpose and significance of the study, research hypothesis as well as the current section, structure of the research. The section will also cover few of the limitations or setbacks to the research study.

The Literature review will cover some of the related works some of the scholars have undertaken. Since we are trying to look on the effect of Twitter sentiment analysis on stock market prediction, this section will cover studies done on sentiment analysis either from Twitter or any other source, stock market prediction and also few previous studies on Sentiment analysis on stock market prediction.

Data and Methodology section covers both data collection methods as well data analysis methods. This section will briefly explain different types of research methodologies before elaborating data collection methodologies for both Tweets and Stock market data. Furthermore, the section also covers data pre-processing methods for both Twitter sentiments and Stock return values. This section will also explain the methodology and model of the study on how we achieve our results and prepare our data for forecast in a time series.

The Findings section will cover the correlation between the variables in a time series using Autocorrelation and Partial autocorrelation test, stationarity test for a time series using Unit Root and Dicky-Fuller Test. Furthermore, this section will also cover Heteroskedacity and Arch effect tests as well as identify which model gives us the best fitting to predict volatility using ARCH model, GARCH model and sentiment scores to predict stock volatility. We will also analyze the results of the basic Ordinary Least Regression as well as whether there is linear relationship between our given variables.

The Conclusion section will summarize the results analyzed from our finding section. The section will also analyze which model fitted best towards precision of stock volatility in the time series.

Furthermore, we will also analyze various limitation we faced in our study and what could be done to achieve better results and precision in stock volatility using another variable in sentiment scores.

1.6 TECHNOLOGY USED IN THE RESEARCH

Since Twitter only limits to mining of 3200 tweets using their official API, we used a twitter intelligent tool known as *Twint* to scrape enough tweets corresponding to S&P500 stocks. Twint can simply be defined as a python scraping tool that was created to scrape data on a given topic from given users. To avoid any confusion with non-stock market related tweets, we added a *cash tag* (hashtag with a $ sign before the stock name) to only acquire stock related data. Other advantages of using Twint over the twitter API include: No API requirement, easy to set up through python and totally cost free.

E-views software was used to conduct econometrical analysis from the correlation analysis to the forecasting of stock returns with regard to the sentiment scores. Despite existence of free of charge software applications for econometrics such as Gretl, the following are some of the reasons as to why we chose to utilize the E-views software over software applications for our time series analysis:

- Very easy to use given its 'natural' interface
- Very efficient for Time Series analysis i.e. GARCH models, ARCH models, Stationarity test, Unit Root etc.
- Easy to keep your record of work and access it
- Ability to save the output into different formats.
- Despites its cost, the service and customer support are excellent.

2. LITERATURE REVIEW

This section will go through the main keywords in our study and studies previously done by other scholars. This includes different studies done on sentiment analysis, prediction of stock volatility using both machine language approaches as well as ARCH and GARCH models. Finally the section will also cover short summary and analysis of those past related studies.

2.1 SENTIMENT ANALYSIS

Sentiment analysis can simply be defined as opinion analysis (Lee et. al. 2008). Cakra (2015) also defined sentiment analysis a process of determining people's opinion through one's emotions. Sentiment analysis is very important for sectors mainly in need of public opinion to thrive. Sentiment analysis is a field of study in which the one's sentiment in a given article/sentiment is considered as his/her (Turney, 2002). Lee et al. (2008) argued that Opinion analysis can be used as a synonym to sentiment analysis. Through their several studies which covered aspects such as history and terminology, the aforementioned terms above were concluded to be synonymous.

Lee et al. (2008) described three (3) methods used to analysis any given sentiments which are Linguistic methods, Machine learning and Lexicon approaches. Sentiment identification can be processed at three (3) different levels i.e. sentence level, feature level and document level (Liu, B., 2012). Given the significance of public's opinions in processing of decisions (Horta et. al. 2017), sentiment analysis is used by different sectors including political sectors, marketing and financial analysts (Taboada, 2016). Some of the types of sentiment analysis include the following:

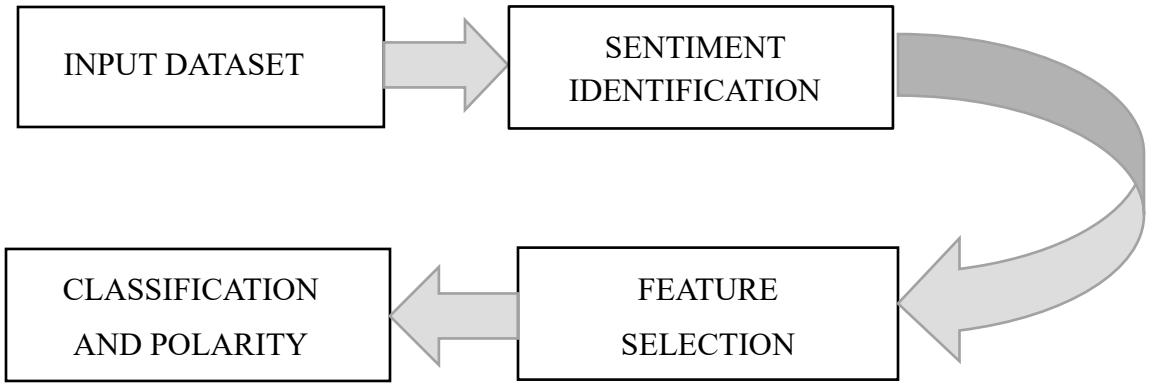
- Emotion detection which determines sentiments using users' emotion.
- Aspect based sentiment analysis determines a particular aspect where a user's emotion on a given product is based (E.g. This laptop is too *heavy*)
- Fine-grained sentiment analysis where polarity level could be expanded to actually get users' sentiment. Ratings could be expanded from 1 to 5 (Very negative to very positive) to determine one's exact emotional response.

Pang and Lee (2004) argued that the sentiment analysis being capable of expressing polarity i.e. Positive, negative and neutral to classify a given dataset makes it even very significant in prediction models. Sentiment analysis process includes two process named sentiment identification and sentiment classification.

Several attempts have been made in past few years to use sentiment analysis to analyze stock market movement. According to Malkiel (2003), EMH explicitly indicates that all the prices of the given shares fully reflect all available and with that, any new insightful info can cause a change in the stock price.

Since then, scholars studying the stock market movement have begun to use sentiment analysis on the datasets to try forecasting the upcoming prices. The term sentiment first appeared as an investor's opinion (De Long et al., 1990). With further progression of research in the stock market, movement of stock prices in the market has been attributed to the people's sentiments whether at an individual, corporate or at an institutional level.

Figure 2.1: Sentiment Analysis Flow



As per Bing Liu (2012), sentiment analysis, also recognized as opinion extraction in several contexts, is a study in a given research that assesses public's perception, thoughts, or mood. Trying to get what the public or a specific group feels towards any given situation can be very hard especially if the size of the dataset is very large. As previously stated, assessment of enormous

datasets can be both laborious and an arduous task. Hence, Pang Lee et al. (2002) argued that to examine such large datasets, sentiment analysis became a necessary outlet given its ability to identify one's mood in a given article or statement using machine learning techniques.

The said task necessitates the analysis of text-based information. However, manually sifting it through enormous datasets became a difficult challenge. The answer was to dynamically decode the context of the provided article if it is either positive, neutral or negative. The methods therefore used to analyze the sentiments of the given data can be divided into three branches i.e. using set of keywords strategy, Sentiment analysis approach and Bag of Words approach.

Processing raw data is simply almost impossible using Machine learning approach. Text available in a given article/statement is ought to be broken into in a numerical format for the machine to be able to read it (the text has to be broken to features and then to vectors. The Bag of Word model is widely utilized in classifying a given statement or a document in which the recurrence of every word is utilized as an attribute to train a given classification model (Dani Yogatama and Noah A. Smith, 2014).

BoW mainly involves the following aspects i.e. a set of vocabulary of words and weight or frequency of the presence of the known words in a statement. In the bag of words model, grammar and order of the words are not important at all hence they are not considered. Both Fung et al. (2002) and Schumaker et al. (2009) continue to express that word count in a given article is used as a feature during classification. The other approach involves using a ready set of vocabulary for classification.

2.2 MACHINE LEARNING

Evolution of the human race has been growing rapidly thanks to eagerness of human beings trying to attain perfection. Machine learning, derived as a field under artificial intelligence, can be simply be defined as automated processes that learn how to perform a given task through a serious of examples. Another way of defining Machine Learning is a process where a given system constructs patterns and models by using a given dataset (George Tzanis, et. al. 2006). Lu (2016) argued that

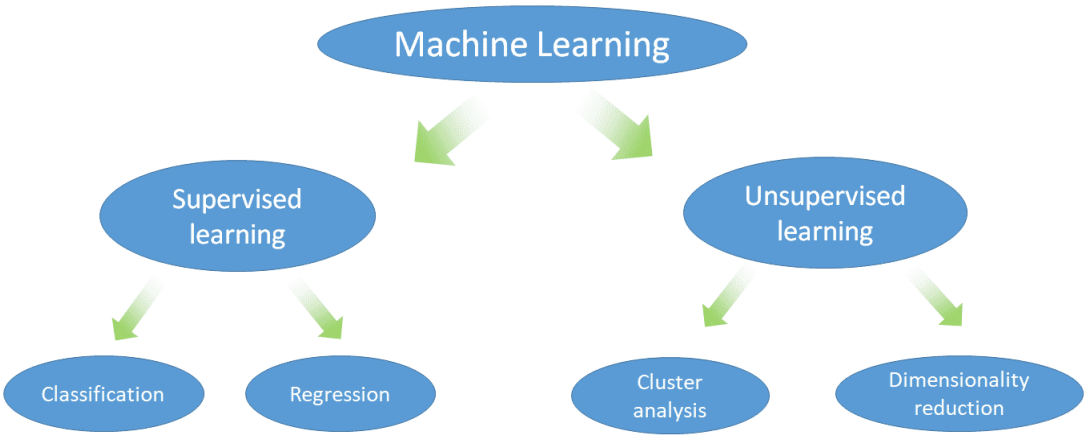
machine learning process can help in prediction of stock market prices given their ability to learn recognize the models and patterns from a given trained data.

Machine learning is classified into the following:

- Supervised Learning: This is type of learning basically through a given series of examples. In this type of learning, the result is obtained mainly by the given input set. Mitchell (2006) defined supervised learning as a technique where datasets are passed to an algorithm along with preferred results expectation so that the algorithm can learn. Supervised learning is divided into types; Classification, where the given input datasets are classified according to their features and Regression, which is used for prediction other features of the datasets using classified datasets.
- Unsupervised Learning: Unsupervised learning also known as learning from observation, differs from supervised learning given the fact that output/desired results are not provided. Here, the machine has to find correlation between the input training dataset. Russell (2003) argues that the setback of this Learning involved having to learning the patterns without having an outcome provided. Unsupervised learning is divided into two types; Clustering involves grouping the data with similar characteristics whereas Association involves determination of presence of any connection between the data in a given cluster.
- Reinforcement Learning: This is a kind of learning when during the absence of training dataset, the algorithm is ought to learn from its own experience. In this type of learning, previous collected data are used to process a feedback of a given task which in turn created a closed loop of a behavior.

Amongst the most common used Machine learning algorithms includes Random Forest, K-Means Algorithm, Naïve Bayes, Logistic Regression Algorithm, Decision Tree and SVM (Support Vector Machine).

Figure 2.2: Machine Learning Classification



Kofi Nti et al. (2020) carried out a research study to try predicting the Ghana stock market movement using different datasets combined of news articles, forums, tweets etc. using MLP, a class of Artificial Neural Network (ANN) which was thought to be more accurate and efficient. The research focused on the datasets collected from GSE between the months of January 2020 and September 2019, with the results on the collective dataset showing an accuracy of around 76%, however a significant drop in the accuracy results to 42%-60& was noticed when datasets were analyzed separately. This further concluded that an even bigger dataset will improve the model's accuracy. Furthermore, Kofi Nti et al (2020) went on to achieve a high correlation between social media platforms and the stock market.

Fung et al. (2002) applied a machine learning algorithm named guided clustering towards trying to study the impact news articles have on the stock market movement. The news articles were then filtered into "Rise" and "Drop sections in an attempt to try predicting the stock market prices. The surprising fact is that the study did not have a fixed time frame and it rather focused on the incidents from the news. Further training was then conducted on the events including T-test to try discovering different trends followed by the clustering algorithm. K-mean algorithm was the applied to classify the clusters with respect to their trends which unfortunately couldn't support the hypothesis. Fung et al. (2002) then came up with the idea to use an altered weighting method that assigns larger

weights to features that appear often in one set of articles and are only present in one trend collection.

Cruz et al. (2021) conducted another research to determine the effect of the Covid-19 pandemic on the stock market. This results were to be compared to the results of another pandemic, H1N1. Their findings reveal that during the COVID-19 pandemic, the financial markets mirrored information from Twitter within 0-10 days, whereas it required 0-15 days for the news on H1N1 to be reflected in the financial markets.

Cruz et al. (2021) used Twitter to collect data and evaluated several financial indicators to reflect each continent. The findings indicated that during the COVID-19 pandemic, information on Twitter had a higher negative impact than during the H1N1 pandemic. The sentiment was extracted using a lexicon-based technique, in which the data was tokenized and split before special characters were removed. Five distinct lexical elements were utilized to determine the text's sentiment value.

Qiu et al. (2013) in his study also analyzed the impact played by social media towards the prediction of the stock market where all the participants in the study embraced the threshold strategy in the Bayes-Nash equilibrium. In their study, they came up with the conclusion that the information from Twitter and other social media platforms can better the precision of network-implanted prediction together.

Another related work on sentiment analysis includes the study proposed by Xu and Keelj using the collected tweets between May 13th and 31st of 2012 to predict the following day's stock market prices (Xu & Keelj, 2014). The scholars initiated two approaches: Natural Language Processing (hand-labeled classification) & Machine learning (Naïve Bayes) for detecting and classifying the sentiments. Tweets were classified into positive and negative using their polarity degrees giving two satisfactory accuracy results of 72% and 75%.

Khan et al. (2019, p.11040) also conducted the study on how both public's sentiment and political events affected the movement of the stock market. They were especially eager to examine how

political events affected the stock market, at both company and market level. They used the KSE to determine how political events affected the stock market.

Ten (10) different machine learning algorithms were then applied on the collected datasets to determine how far the public' sentiments affected the movement of the stock market with all of the achieving the highest prediction accuracy on the seventh (7^{th}) day. The learning methods to achieve the highest accuracy included Naive Bayes (NB), K-Nearest Neighbors (KNN), Sequential Minimal Optimization (SMO) and LWL. The same algorithms were also to a sum of Ninety Eight (98) of political events to determine the effect they had on the movement of stock prices in the Karachi Stock Exchange.

With the highest accuracy being achieved on the fifth (5) day, a conclusion showed that these events had the most impact on the stock market on the Fifth (5) day. In the end, he also depicted that the accuracy would have fallen by 20% if analysis on political events wasn't included employing how influential political events were on the stock market (Khan et al. 2019, p.11040).

Şimşek and Özdemir (2012) also conducted a simple research study where they only tried to find the frequency of only two words i.e. happiness and unhappiness and determine on their impact on the stock market. Tweets and data from Turkish Stock market were collected and the authors conducted few data visualization technique e.g. frequency distribution and determined there was a positive correlation between the variables (Tweets and the Stock Market).

Zhou et al. (2016) established a positive correlation to be existing between Twitter moods and the stock market movement and such correlation can be very useful towards prediction of the stock market. Using a self-organizing fuzzy neural network, Ahuja et al. (2015) used sentiment analysis of public sentiment from Twitter to establish a correlation between the stock market and emotions from tweets. Nguyen et al. (2015) proposed a model to forecast stock market price movement using company based sentiments from social media.

Another study regarding the impact of Twitter sentiment analysis on the stock market was conducted by Xia and Song (2016). The scholars used student datasets from a university then

divided them into different clusters i.e. departments, semesters etc. They came up with a surprising results that students belonging to the social science department had more positive sentiments than the other departments. Negative sentiments were also more depicted during the ending of a given semester. Having established this, they concluded that studying timing and users' region must be prioritized.

Cakra and Trisedya (2015, pp.147-154) established an algorithm to predict the Indonesian Stock Market using public's sentiment on Twitter. The authors applied two machine learning algorithms i.e. Naïve Bayes and Random Forest to classify the collected sentiments from the given companies in Indonesia. Linear regression algorithm was the applied to build a prediction model which in turn gave an accuracy of 57 % with NB and 61% with Random Forest respectively.

In Saudi Arabia, Hamed et al. [2015] conducted a research study to see if there was any correlation between public mood from Twitter and the stock market movement. Using a dataset of around 3300 tweets collected over a period of two (2) months, opinion mining algorithms were then trained on the variables to find the correlation. The results depicted a strong correlation between public's mood and the Saudi Stock Index.

SVM, or support vector machine, was utilized by Ren, Wu, and Liu (2019) to examine the influence of sentiment on SSE 50 index. They also took into account the day-of-week impact to boost the sentiment index's authenticity, as returns on Mondays are way lower compared to the ones of the rest of the week. The findings showed that merging stojhöck market data with features extracted from sentiment analysis yields an accuracy of 89.93 percent than the one retrieved from the existing methods which gave an average of 71.33 percent, implying that sentiment features are critical for predicting the stock market movement and can assist investors in managing their risks during trading.

By adopting an upgraded version of the asymmetric GARCH model of conditional volatility, Paramanik and Singhal (2020) conducted a study on the Indian Stock Exchange between 2007 and 2020. Three (3) distinct GARCH models were used to determine the impact of these sentiments on

the ISE. Paramanik et al (2020, p. 340) concluded that the impact of the noise traders' negative moods is more precise and explicit than the positive sentiments.

Ghanavati M et al. (2016) established a framework to try to forecast the Hong Kong stock market accumulated over a period of one (1) year. The framework allowed its users to select their machine learning technique and dataset of their preference. The authors established that metric learning methods can better the prediction accuracy results.

Bing L et al. (2014) used sentiment analysis and a machine learning algorithm to predict NASDAQ as well as NYSE stock movements using a dataset of over 15 million. The machine learning algorithm applied Natural Language Processing for classification of tweets. The recommended technique explains the hidden relationships within social networking sites as a graph with multiple levels, including top, middle, and bottom layer characterized to demonstrate the relationship. The findings indicated a stronger correlation between the public's sentiment and the stock market movement three (3) days later.

Chen et al. (2014) conducted another research study using social media sentiments to determine the prediction of the stock market movement using Factorization Machine. The research study included data collected from 361 days of trading and data collected from a social media platform called Sina Weibo. His study included three (3) phases which were analysis of the correlation between factorization machine to the SVM and linear model, usage of social media sentiment to forecast the stock market, and lastly to establish the pros using this method in machine learning and prediction. The Factorization Machine recorded an accuracy of around 82% meaning it provided the better precision than other models.

Chouliaras (2015) conducted a research study to determine where the news articles based on political events had an effect on the stock market prices and if so, could then be trained to forecast the stock market. He conducted sentiment analysis on the articles using a text analysis of news related an economic crisis in Europe. He later concluded that negative news of the political events resulted in either volatility or a decrease of the stock prices. Taimur & Khan (2015) combined both disastrous events and political events to study on how both of these two affected the stock market

prices. The authors conducted this study by an interval of 5 days to see when the news caused the most impact on the Stock market. From their research study, they concluded that all the positive news articles impacted the stock market in the period of 3-5 days later whereas negative impacts showed sudden impact on the market either by depicting volatility or a decrease in the stock's price.

Shakhla et al. (2018) conducted another study on prediction of stock market movement using linear regression. During their research, a model using RMSE was established through studying AAPL stock market prices. Moreover, they further used multiple linear regression during prediction of the given stock's following day's opening price. Oliveira et al. (2013) applied the linear regression method using the datasets acquired from Stockwits, a well-known community for traders. The authors tried to study the prediction of absolute stock return, trading volume as well as the stock's volatility. The authors concluded that it was impossible to forecast the stock returns using the indicators of the given sentiments.

Another familiar research study conducted by Bollen (2011) also evaluated whether the public's emotional sentiments derived from their tweets were correlated to the DJIA. To evaluate their prediction of the variables, Fuzzy neural network model was used. Their findings reveal that public sentiment on Twitter is strongly correlated with the Dow Jones Industrial Index. After considering further case studies, Porshnev (2013) decided to use sentiment analysis as an additional variable towards the improvement of the stock market prediction accuracy. Psychological analysis on Twitter's users was thoroughly done and the information extracted was used in prediction of the following day's stock market price movement. Porshnev further concluded that additional of sentiment analysis had increased the models accuracy to 64% hence establishing that sentiment analysis had no major impact in stock market movement.

Nova and Hinz (2015) are among other scholars who conducted the study on the impact of public mood on the stock market movement. The scholars extracted more than 100 million tweets from Germany users to conduct their research study. During their research, two phases were analyzed with different indicators to see whether both of them would have positive correlation to the stock market movement. During their first phase of analysis, the scholars concluded that there was no correlation between Twitter sentiment and the stock market movement before finding positive

correlation in their second phase. After both analysis, the scholars concluded a correlation could be found after taking factors such as the size of followers, main tweets etc into account.

2.3 ARCH MODEL

Despite the existence of a large number of non-linear forecasting models, ARCH and GARCH models still seem to be the most utilized models when it comes to forecasting especially in financial aspects. One of the main characteristics that comes with autoregressive models is that there is a linear dependency between a conditional variance at any given time in a time series to its past values. However this does not sit well especially with the dynamics of the stock market given it overlooks the recent impacts while overestimating those impacts of far in time. This is model is non-linear therefore does not carry the assumption that the variance is constant in a given time series.

There are several reasons as to why the ARCH model became so popular with precision and forecasting of a given variable in a time series. Among those reasons include its simplicity and only needs one input (return series) to estimate conditional variance and to generate the lagged residuals. Another feature that brought about ARCH model's popularity was leptokurtosis, where by the model's output displays heavier tails compared to the ones given by the normal distribution. Furthermore ARCH models is very popular when it comes to provision of results in terms of volatility clusters. Volatility clustering simply depicts the aptness of major changes in a given assets follow major changes whereas small changes in a given asset as well tend to also follow small changes.

Like discussed before, Brooks (2008) also indicate that through the allowance of the conditional variance of the error term to depend on its past values of the squared error, the autocorrelation in volatility can be modelled with the ARCH(1) model assuming the following equation (1). However the equation of the ARCH model may change to the following form if the error variance is dependent on the lags of squared errors given as n in our equation as depicted in equation (2):

$$\varphi_t^2 = \varphi_0 + \varphi_1 u_{t-1}^2 \text{-------- (1)}$$

$$\varphi_t^2 = \varphi_0 + \varphi_1 u_{t-1}^2 + \varphi_1 u_{t-n}^2 \text{ -------- (2)}$$

The following are among few of the disadvantages of the ARCH model discussed by Tsay (2010) hence enhancing the need for extension to the model:

- The ARCH models tend to assume both the positive and negative shocks could have the same impact on the stock volatility given the dependency of variance on the square of past shocks
- The ARCH model has a very slow reaction to the isolated shocks in a given return time series.
- The ARCH models comes with several restrictive limitations to the parameters used in weighing.

2.4 GARCH MODEL

The GARCH model, an extension to the ARCH model is widely known for forecasting of volatility was designed by Bollerslev (1986) as an extension and generalization of the ARCH model. Bollerslev (1986) indicated in his study that volatility at any given point in a time series is the result of past volatility values and lagged return values. The sole advantage that GARCH possess over ARCH models is that despite the need of fewer parameters for analysis, it still yields better results over the ARCH model.

Among recent studies regarding GARCH models includes that of Panait and Slăvescu (2012), where they indicated that the GARCH in mean model failed to support the hypothesis that the volatility exhibits a positive relationship and future return. Another related study includes the analysis conducted by Harris and Pisedtasalasai (2006), with the aim of analyzing mean and shock transmission as well as the relationship return and volatility of both bigger and smaller companies in the United Kingdom. The study came out with the conclusion that the flow of volatility is always going from the larger markets to the smaller equity markets in the United Kingdom.

Despite further extensions to solve problems associated with ARCH models, further studies and analysis have called on the inclusion of further features to improve the model. The incorporation

of leverage effect still seem to be an essential issue. This has called for further improvement, with Engel et al. (1990) suggesting the adjustment of error term in variance to account for the leverage effect. Furthermore, a non-linear asymmetric GARCH or popularly known as N-GARCH was also suggested by Higgins and Berra (1992).

Beirne et al. (2013) also managed to analyze the flow of shock within the developed and currently developing markets. In their research study, they were able to come up with the conclusion that there was a clear transmission of shock from the developed markets to the currently new / developing markets. The analysis in this study was achieved using a tri-variate GARCH model.

Similar related study was conducted by Wang et al. (2005) where they analyzed the flow of shock within the developed and recently emerging markets. Using EGARCH model, they came out with the conclusion that, an existence of shock transmission and returns from developed markets to the newly emerging markets. Furthermore, Sakthivel et al. (2012) also did an analysis on the movement of shock transmission and stock returns between markets of several countries. Their research study confirmed the existence of a two directional movement of shocks from the Indian market to the US market.

Erdem et al. (2011) in their study decided to try of ten different ARMA-GARCH and ARMA-GARCH-M models on wind speed data in a window of every hour. The analysis pointed out that the further mentioned model is able to easily catch the change of trend of both wind speed's volatility and mean. In their study, they further concluded that symmetric ARMA-GARCH-M were found to be fast, the asymmetric model for the same inputs was found to be more demanding.

3. DATA AND METHODOLOGY

Methodology in research can be basically understood as a collection of guidelines and methods for conducting research. Another simply put definition of research methodology is the author's approach towards answering a given hypothesis (Saunders et. al, 2007). There are three (3) categories of research methodologies (Hair et. al, 2007):

i. Exploratory Research methodology. This is a type of methodology where a researcher has very little knowledge on the given hypothesis hence the main goal in this research is to mainly get a snippet of idea on the given hypothesis.
ii. Casual Research methodology. In this type of methodology, a researcher runs an experiment to determine where there is any relationship between the given variables.
iii. Descriptive Research methodology. This is a type of research methodology which involves a case study. The researcher has to correlate his results from his methodology to the ones obtained from the case study.

This section will elaborate the research methodologies used to test our given hypothesis. The section will cover data collection from both Twitter and Yahoo finance for stock market data as well as all the methodologies we utilized towards achieving our objectives.

Our conceptual framework deputizes on how we will be approaching our dissertation. As described below, data collection will of both twitter sentiments and stock returns will be collected first before processing. From then, we will then check for stationarity test followed by autocorrelation between the variables using Durbin-Watson tests. We will then analyze the best fitting model before using GARCH model predict the stock volatility.

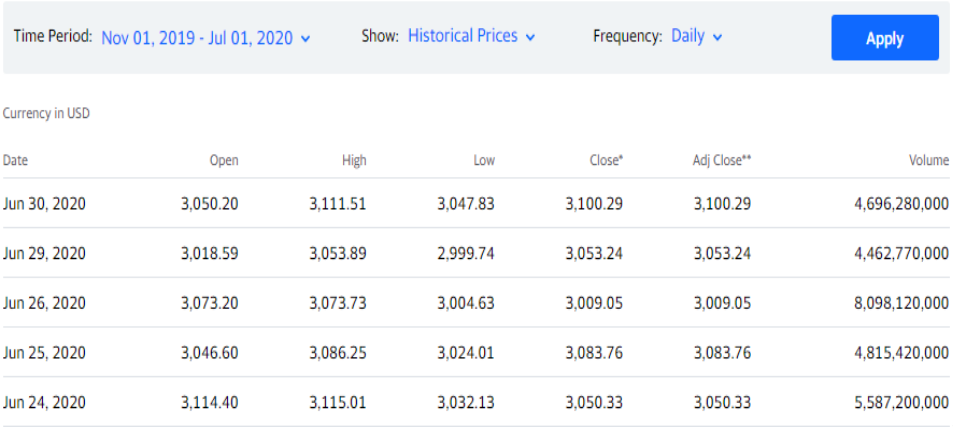
3.1 DATA COLLECTION

To undertake our research study, we were required to collect original datasets of both the stock exchange market as well public mood in tweets from Twitter. Given we decided to go with indices, we used the S&P500 indices stock market data between November 2019 and July 2020.

3.1.1 STOCK MARKET DATA

In this study, the data series of the stock return was collected in a span of 8 months between November 2019 and June 2020. The data was collected from a well trusted website in Yahoo Finance which overall seem to be providing real time data over a given span of time. Furthermore, some of the other reasons we chose to use Yahoo Finance as our source over other platforms include its easiness towards extraction of data into different formats as well as loading speed time of the website during extraction of big data

Figure 3.1: S&P 500 Index Stock Market Prices (Yahoo Finance)



Time Period: Nov 01, 2019 - Jul 01, 2020 Show: Historical Prices Frequency: Daily Apply

Currency in USD

Date	Open	High	Low	Close*	Adj Close**	Volume
Jun 30, 2020	3,050.20	3,111.51	3,047.83	3,100.29	3,100.29	4,696,280,000
Jun 29, 2020	3,018.59	3,053.89	2,999.74	3,053.24	3,053.24	4,462,770,000
Jun 26, 2020	3,073.20	3,073.73	3,004.63	3,009.05	3,009.05	8,098,120,000
Jun 25, 2020	3,046.60	3,086.25	3,024.01	3,083.76	3,083.76	4,815,420,000
Jun 24, 2020	3,114.40	3,115.01	3,032.13	3,050.33	3,050.33	5,587,200,000

3.1.2 TWITTER DATA

Collection of data using a Twitter API such as Tweepy can be very complex and slow given that the API can only allow a maximum of 3200 tweets to be extracted daily. To avoid this and make sure we had a simple and faster way to mine the tweets, we used a python library called TWINT.

As mentioned earlier, most of the investment traders use the Twitter platform as a means to communicate or express their emotions regarding the market. Traders (users) tend to mostly add a USD sign ($) (commonly known as ‘cash tag’) before the discussed indices (S&P500). Using Twint, we were able to extract a dataset of over 120,000 tweets related to S&P500 over the same period was also collected for analysis at the stock market level.

3.2 DATA PRE-PROCESSING

3.2.1 Sentiments

Kaastra and Boyd (1996) are among few scholars who have established a need for a vigorous cleansing of the data before used for stock prediction. A lot of sentiments collected especially on social media platforms need to be cleaned before analyzed given that most of the sentiments may either consist of unnecessary words and as well emoji which could be hard to analyze one's emotional response to a given issue. During our research study, we implemented the following filtering during the pre-processing of tweets:

- Data Normalization: This process includes mostly removal of unnecessary words that do not depict any mood or emotion. Example off these words include a, the, an etc. Among the methods of normalization in our study included removal of Stop Words and converting the string into lower cases (in the corpus).
- Tokenization: In this process, Tweets are broken down into individual words based on the space, and unnecessary symbols such as Emoji get eliminated. For each tweet, we create a list of these words to represent them.

3.2.2 Log Return

Unlike Sentiment scores obtained from social which are basically present every single day, it is hard to achieve the same for stock returns given the stock market trading is closed during holidays, weekends etc. Hence in our study, for everyday the stock market trading was closed, we attributed a value of 0% to the log return on that given day.

3.3 SENTIMENT ANALYSIS

In our research study, we used Neural N Bayes classification technique to classify our tweets into positive, neutral and negative. Values -1, 0 and 1 were assigned to negative, neutral and positive tweets respectively. A model in machine learning was also trained to classify and drop tweets which seemed to be from robots. Naïve Bayes theorem is defined by the formula below:

$$P(x|y) = \frac{P(y|x).P(x)}{P(y)}$$

Where: $P(x|y)$ is a probability of X happening provided Y happens, $P(y|x)$ is a probability of Y happening provided X happens, $P(x)$ is a probability that X happens and $P(y)$ is a probability that Y happens. Assuming our set of tweets denoted as set C, to establish the conditional probability of a class, we combine all these assumptions to the Naïve Bayes Theorem formula to finally come with the below formula:

$$\hat{C} = argmax\ P(c|d)\ \ c \in C$$

Where c is among the classes taken from the Set C, d denotes the document provided for classification and $P(c|d)$ denotes the probability of class C if document d is given.

3.4 FEATURES EXTRACTION FOR SENTIMENT ANALYSIS

Using Natural Language Tools Kit, we established our training data as a corpus of sample tweets containing both positive and negative tweets. These tweets corpus is combined of two 3 sets; a sample of both positive and negative tweets combined as well as two sets of positive and negative tweets separately. After achieving the score for every sentiment that was collected. A total sentiment score was calculated for that given day by adding the sentiment scores from 60 tweets relating to the S&P500 indices. Positive results indicated positive sentiments where as a negative score represented a negative sentiment.

3.5 MODEL OF OUR STUDY

Previous research studies have out found multiple regression techniques to produce a sufficient accuracy between the given variables. Kamley et al. (2013) among the many, established an accuracy of about 88.63% on the prediction of stock market movement. But considering the assumption that all the values in a time series have a constant variance over time, OLS regression was not deemed to be a proper method to analyze and forecast the stock volatility given the inability

of a linear regression model to effectively elaborate the data included in a non-linear relationship. Instead, we went on with ARCH and GARCH models to find the best fitting model to the data so we could try forecasting the stock volatility using the sentiment scores given the existence of heteroskedasticity between the values in the time series.

First and foremost, we tried using the Ordinary Least Squares Regression to determine whether there was significant relationship between the two variables. Moreover, Ljung-Box test was also conducted to check whether there was autocorrelation of the variables at multiple lags jointly. Like stated above, most studies have assumed of the correlation between variables to be linear. A scatter plot between the two variables was also plotted to see whether there is linearity or not.

If there is existence of autocorrelation between the variables, ar and ma terms to check for autoregressive effect of log returns and presence of moving average effect respectively so that a model with best fitting to the data could be used for precision and forecast of the stock return and volatility.

In general most of the data provided in the financial time series tend not to be stationary given the presence of volatility which may be caused by different reasons. Hence to make sure we converted our data to stationary values, we used closing prices to compute logarithmic returns using the following method where $Closing\ Price_t$ represents the current closing price and $Closing\ Price_{t-1}$ represents the closing price of the indices a day before.

$$Log\ Return_t = Log\ (^{Closing\ Price_t}/_{Closing\ Price_{t-1}})$$

Furthermore, a test was undertaken on the sample data to check whether the values of stock returns are stationary. This was done using the Unit Root Hypothesis and Dickey-Fuller test, which according to the rule states the variable would be considered not stationary if there is an existence of a unit root test in a time series. To achieve a better fitting and precision for our variables, the data in a given time series must be stationary. The equation for Dickey and Fuller test is provided below:

$$\nabla Z_t = \alpha_0 + \alpha_1 z_{t-1} + \sum_{i=1}^{nl} \alpha \nabla Z y_t + \varepsilon_t$$

Where ∇: the first difference,

Z: represents the time series

α_0: represents the constant

ε_t : represents the white noise

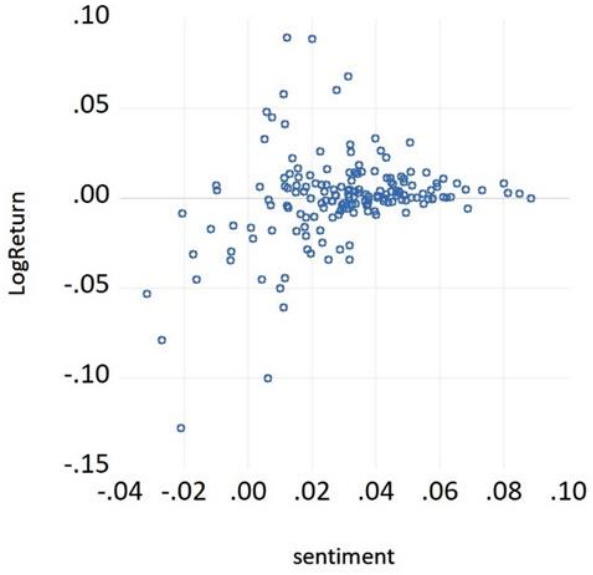
The best model derived from the above, will then be tested for ARCH effect for Heteroscedasticity, and if there is a presence of arch effect on that particular model, sentiment scores will be added to the GARCH model to analyze and forecast the impact of sentiment scores on the stock volatility. The forecasting model is then formed by using the combination of the following input parameters; AR-lag and MR- lag terms or either one of them depending on the significant levels at the given orders, the GARCH model and the return distribution which provides the lowest Akaike Information Criterion.

4. FINDINGS

Ordinary Least squares Regression

A Scatter plot of log Returns against sentiment score was first used to visualize the relationship between the two variables. The plot shows an existence of a non-linear relationship between the two variables. Similarly, there is heteroskedasticity of error variance where by variation in log returns decline as sentiment score increases, i.e. as we get more positive sentiments, returns tend be consistent. The figure below displays the scatter plot that was modelled to determine the linearity between the variables.

Figure 4.1: Scatter plot between Log Returns v/s Sentiment



Intuition behind the non-linear relationship.

The intuition behind the non-linearity function between the sentiments and stock return could be well associated with “Noisy Traders”, a theory also supported by Cliff et al. (2004) who stated that nose traders’ sentiments had a negative impact on returns. An article in Investopedia classified the traders into the following types:

- Fundamental Traders: These are type of traders who mainly trade by analyzing past events of the given company.
- Sentiment Traders: These are traders who mainly tend to identify and participate in trends. Example of these traders include Swing traders who mainly trade by trying to catch a significant movement of prices.
- *Noise Traders*: These are traders influenced by mostly outside noises with neither professional help nor technical ability to trade.
- Market Timers: These are traders who mainly use either technical or economic indicators to try forecasting the price change.
- Arbitrage Traders: These are traders who concurrently buy and sell their stocks in attempt to achieve profits.

Looking at our plot displaying non-linearity between the variables, we can assess from our intuition that *Noisy Traders* may be the reason behind the non-linearity. As more negative sentiments increased, noisy traders will mostly away sell their stocks rather than buying. At some degree of negative sentiment, only traders investing basing on fundamentals and technical analysis will keep on holding the stock, hence depicting consistency in returns. Returns are less affected by negative sentiment.

Figure 4.2: Ordinary Least Squares Regression Output.

Dependent Variable: LOGRETURN
Method: Least Squares
Date: 09/07/21 Time: 12:44
Sample: 11/01/2019 6/30/2020
Included observations: 166

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.014728	0.003217	-4.577753	0.0000
SENTIMENT	0.974267	0.166564	5.849207	0.0000
S2	-10.61753	2.507203	-4.234812	0.0000

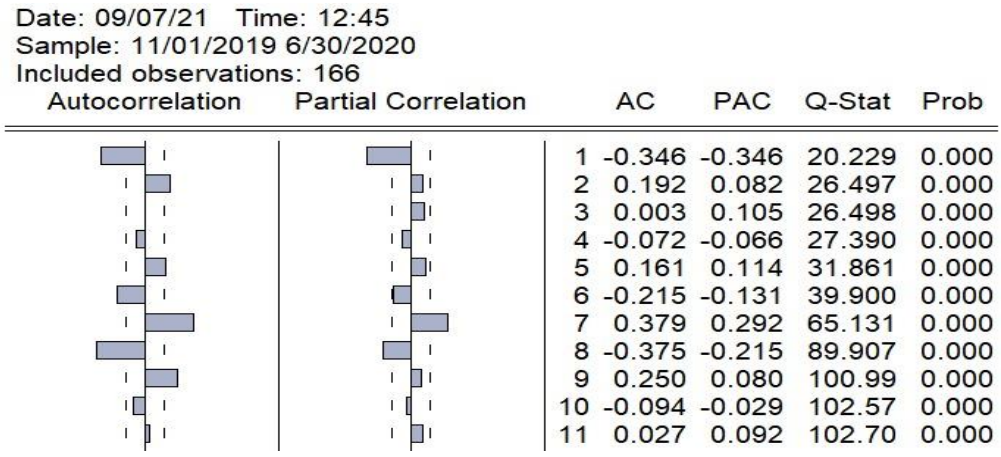
R-squared	0.184897	Mean dependent var	6.52E-05
Adjusted R-squared	0.174896	S.D. dependent var	0.025462
S.E. of regression	0.023129	Akaike info criterion	-4.677588
Sum squared resid	0.087194	Schwarz criterion	-4.621347
Log likelihood	391.2398	Hannan-Quinn criter.	-4.654759
F-statistic	18.48734	Durbin-Watson stat	2.690515
Prob(F-statistic)	0.000000		

Autocorrelation

A non-linear coefficient for sentiment score was included on the model to account for the non-linearity. The F test for the model is significant at 5% level ($F=18.48$, $p < 0.001$) this means that at least one of the coefficient is significant can also be interpreted to mean that sentiment influences log returns either in its linear or quadratic form. Both the linear and quadratic coefficient. When sentiment is 0 (neutral sentiment) the slope is such that Log Return would increase by 0.974267 on average for every unit increase in sentiment, but this slope would keep declining by 10.62 on average for every unit increase in sentiment score.

The value of 4-DW = 4-2.690515=1.309485, This value falls below the tabulated 1% alpha critical values for Durbin Watson test (1.584,1.665) meaning there is evidence of negative autocorrelation at 1% level. A residual correlogram for shows existence of autocorrelation. Conversely, Ljung-Box Q is significant for all the 11 lag orders tried.

Figure 4.3: Residual Autocorrelation Test



Stationarity Test

A graph of log returns shows no trend, similarly, with Argument Dick fuller test we see no evidence of trend. The test is significant which means the time series is stationary.

Figure 4.4: Graph of Log Returns Volatility

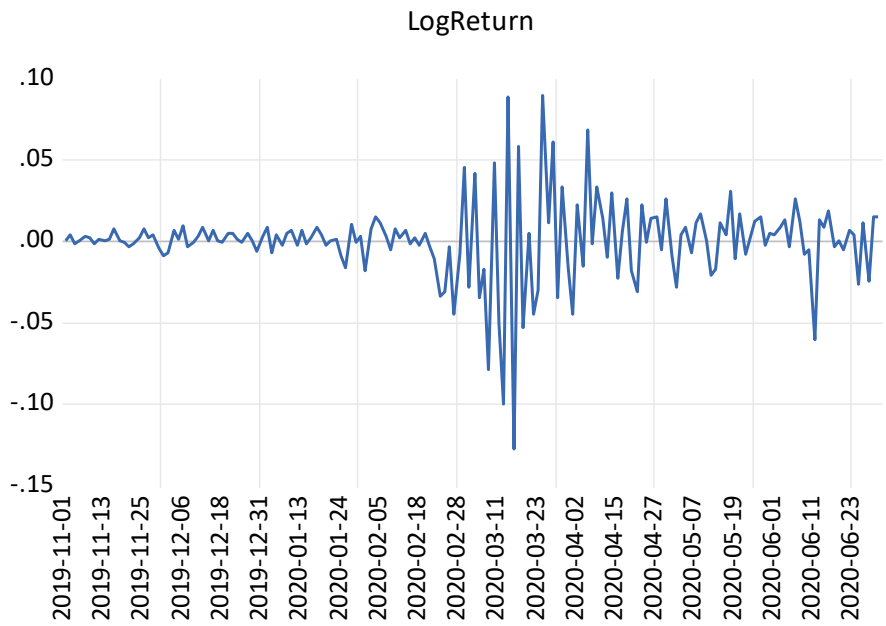


Table 4.1: Unit Root Hypothesis and Dickey Fuller Test

Null Hypothesis: LOGRETURN has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 0 (Automatic - based on SIC, maxlag=11)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-18.85535	0.0000
Test critical values:	1% level	-4.014635	
	5% level	-3.437289	
	10% level	-3.142837	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(LOGRETURN)
Method: Least Squares
Date: 09/18/21 Time: 20:53
Sample (adjusted): 2 166
Included observations: 165 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.

LOGRETURN(-1)	-1.374943	0.072921	-18.85535	0.0000
C	-0.000913	0.003726	-0.244961	0.8068
@TREND("1")	1.17E-05	3.89E-05	0.299586	0.7649

R-squared	0.686976	Mean dependent var	9.27E-05
Adjusted R-squared	0.683111	S.D. dependent var	0.042319
S.E. of regression	0.023822	Akaike info criterion	-4.618369
Sum squared resid	0.091936	Schwarz criterion	-4.561897
Log likelihood	384.0154	Hannan-Quinn criter.	-4.595445
F-statistic	177.7659	Durbin-Watson stat	1.867109
Prob(F-statistic)	0.000000		

Adding AR(1) and MA(1) Model

Given that we found autocorrelation in the residuals, three (3) models were included at this stage ar and ma terms were added to the model to select the one which provided best fitting to the data. *Model 1* was added with ar(1) term only, *model 2* with ma(1) term only whereas both ar(1) and ma(1) terms were added to *model 3*.

Using the models we included with ar and ma terms, From the illustrated table below we could be able to conclude that the ar(1) coefficient in *model 1* is significant, at 5% level (B = -0.383721, $p < 0.001$) of significance. Which means there is autoregressive effect on log returns.

Table 4.2: Model 1 adding AR(1) terms only

Dependent Variable: LOGRETURN
Method: ARMA Maximum Likelihood (OPG - BHHH)
Date: 09/18/21 Time: 12:48
Sample: 1 166
Included observations: 166
Convergence achieved after 38 iterations
Coefficient covariance computed using outer product of gradients

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.011474	0.001708	-6.717776	0.0000
SENTIMENT	0.729681	0.127722	5.713030	0.0000
S2	-7.622292	3.203305	-2.379509	0.0185
AR(1)	-0.383721	0.046635	-8.228261	0.0000
SIGMASQ	0.000453	2.54E-05	17.83816	0.0000

R-squared	0.296901	Mean dependent var	6.52E-05
Adjusted R-squared	0.279433	S.D. dependent var	0.025462

S.E. of regression	0.021614	Akaike info criterion	-4.800350
Sum squared resid	0.075212	Schwarz criterion	-4.706615
Log likelihood	403.4290	Hannan-Quinn criter.	-4.762302
F-statistic	16.99660	Durbin-Watson stat	1.942543
Prob(F-statistic)	0.000000		
Inverted AR Roots	-.38		

Similarly to model 1, the ma(1) coefficient in model 2, was found to be significant at 5% level (B = -0.296015, p < 0.001) which gave the indication of existence of Moving average effect.

Table 4.3: Model 2 adding MA(1) terms only

Dependent Variable: LOGRETURN
Method: ARMA Maximum Likelihood (BFGS)
Date: 09/18/21 Time: 12:49
Sample: 1 166
Included observations: 166
Convergence achieved after 4 iterations
Coefficient covariance computed using outer product of gradients

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.011956	0.001727	-6.924276	0.0000
SENTIMENT	0.785863	0.123146	6.381556	0.0000
S2	-8.504131	3.002909	-2.831965	0.0052
MA(1)	-0.296015	0.050253	-5.890502	0.0000
SIGMASQ	0.000471	2.83E-05	16.62262	0.0000

R-squared	0.269764	Mean dependent var	6.52E-05
Adjusted R-squared	0.251621	S.D. dependent var	0.025462
S.E. of regression	0.022027	Akaike info criterion	-4.762886
Sum squared resid	0.078115	Schwarz criterion	-4.669151
Log likelihood	400.3195	Hannan-Quinn criter.	-4.724839
F-statistic	14.86916	Durbin-Watson stat	2.126506
Prob(F-statistic)	0.000000		
Inverted MA Roots	.30		

In model 3 *(with both terms)* only ar(1) term is significant at 5% level (B = -4.345881, p < 0.001). Which suggests that we only need ar(1) term. On comparing the three models model 1 using only

ar(1) terms found to have the least AIC (-4.800350) which means it provides the best fits to the data. It fits the data even better than the original OLS regression.

Table 4.4: Model 3 adding both AR(1) and MA(1) Terms Together

Dependent Variable: LOGRETURN
Method: ARMA Maximum Likelihood (BFGS)
Date: 09/18/21 Time: 12:50
Sample: 1 166
Included observations: 166
Convergence achieved after 10 iterations
Coefficient covariance computed using outer product of gradients

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.011500	0.001819	-6.320539	0.0000
SENTIMENT	0.726342	0.136493	5.321474	0.0000
S2	-7.527843	3.270792	-2.301535	0.0227
AR(1)	-0.485370	0.111685	-4.345881	0.0000
MA(1)	0.117644	0.123104	0.955644	0.3407
SIGMASQ	0.000451	2.79E-05	16.16516	0.0000

R-squared	0.299452	Mean dependent var	6.52E-05
Adjusted R-squared	0.277560	S.D. dependent var	0.025462
S.E. of regression	0.021642	Akaike info criterion	-4.791901
Sum squared resid	0.074940	Schwarz criterion	-4.679420
Log likelihood	403.7278	Hannan-Quinn criter.	-4.746244
F-statistic	13.67851	Durbin-Watson stat	1.984641
Prob(F-statistic)	0.000000		

Inverted AR Roots	-.49
Inverted MA Roots	-.12

Next we tried the ar (2) model where we determined that the coefficient is significant at 5% level ($B = 4.185212$, $p < 0.001$) which means that there is significant autoregressive effect of order 2. The model AIC is although not better(less) than that of ar(1) meaning that ar(1) still provides a better fit to the data.

Table 4.5: Autoregressive AR (2) Model

Dependent Variable: LOGRETURN
Method: ARMA Maximum Likelihood (BFGS)
Date: 09/19/21 Time: 12:26
Sample: 1 166
Included observations: 166
Convergence achieved after 3 iterations
Coefficient covariance computed using outer product of gradients

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.014174	0.002752	-5.150909	0.0000
SENTIMENT	0.884800	0.172050	5.142689	0.0000
S2	-9.044562	3.748153	-2.413072	0.0169
AR(2)	0.201276	0.048092	4.185212	0.0000
SIGMASQ	0.000505	3.35E-05	15.06810	0.0000

R-squared	0.216755	Mean dependent var	6.52E-05
Adjusted R-squared	0.197296	S.D. dependent var	0.025462
S.E. of regression	0.022812	Akaike info criterion	-4.692863
Sum squared resid	0.083786	Schwarz criterion	-4.599128
Log likelihood	394.5076	Hannan-Quinn criter.	-4.654816
F-statistic	11.13880	Durbin-Watson stat	2.627241
Prob(F-statistic)	0.000000		

Inverted AR Roots	.45	-.45	

Furthermore, from the autoregressive AR(3) model for LOGRETURN, we found that the Ar(3) coefficient was not significant at 5% level (B = 0.002862, p=0.9555), and also the AIC is higher than that of ar(1). Hence, we can still conclude that Ar(1) scores provides the best option so far.

Table 4.6: Autoregressive AR (3) Model

Dependent Variable: LOGRETURN
Method: ARMA Maximum Likelihood (BFGS)
Date: 09/19/21 Time: 12:31
Sample: 1 166
Included observations: 166
Convergence achieved after 4 iterations
Coefficient covariance computed using outer product of gradients

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.014749	0.002293	-6.433186	0.0000
SENTIMENT	0.975308	0.154314	6.320262	0.0000
S2	-10.62508	3.376487	-3.146785	0.0020
AR(3)	0.002862	0.051213	0.055891	0.9555
SIGMASQ	0.000525	3.65E-05	14.39633	0.0000

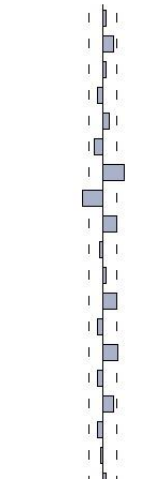
R-squared	0.184903	Mean dependent var	6.52E-05
Adjusted R-squared	0.164652	S.D. dependent var	0.025462
S.E. of regression	0.023272	Akaike info criterion	-4.653499
Sum squared resid	0.087193	Schwarz criterion	-4.559765
Log likelihood	391.2405	Hannan-Quinn criter.	-4.615452
F-statistic	9.130645	Durbin-Watson stat	2.690763
Prob(F-statistic)	0.000001		

Inverted AR Roots	.14	-.07-.12i	-.07+.12i

We also notice that from *the residual autocorrelations plot* below, autocorrelation has also been dealt with since there is no significant autocorrelation for 1 lag order. Hence after comparison of the autoregressive orders, we came up with the conclusion that ar(1) provides the best fit.

Figure 4.5: Residual autocorrelation for AR(1)

Date: 09/07/21 Time: 12:50
Sample: 11/01/2019 6/30/2020
Q-statistic probabilities adjusted for 1 ARMA term



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob*
		1	0.030	0.030	0.1507	
		2	0.108	0.107	2.1380	0.144
		3	0.035	0.029	2.3417	0.310
		4	-0.061	-0.075	2.9755	0.395
		5	0.067	0.065	3.7572	0.440
		6	-0.093	-0.085	5.2807	0.383
		7	0.233	0.236	14.835	0.022
		8	-0.215	-0.246	22.991	0.002
		9	0.147	0.177	26.810	0.001
		10	-0.025	-0.074	26.919	0.001
		11	0.039	0.124	27.189	0.002
		12	0.147	0.038	31.116	0.001
		13	-0.058	0.023	31.722	0.002
		14	0.160	0.026	36.446	0.001
		15	-0.054	0.079	36.977	0.001
		16	0.116	-0.022	39.477	0.001
		17	-0.065	-0.000	40.264	0.001
		18	-0.022	-0.071	40.359	0.001

Heteroskedasticity Test: ARCH for Order (1)

From the Arch effect test below, we observed that the heteroskedasticity test for arch effect was significant at 5% level of significance ($\chi 2(1) = 8.392256$, p-value 0.0038 which is less than 0.05) concluding the existence of significant arch effect.

Table 4.7: Heteroskedasticity ARCH Test for Order (1)

Heteroskedasticity Test: ARCH

F-statistic	8.734802	Prob. F(1,163)	0.0036
Obs*R-squared	8.392256	Prob. Chi-Square(1)	0.0038

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 09/18/21 Time: 13:07
Sample (adjusted): 2 166

Included observations: 165 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.000353	0.000105	3.347366	0.0010
RESID^2(-1)	0.225484	0.076294	2.955470	0.0036

R-squared	0.050862	Mean dependent var	0.000456
Adjusted R-squared	0.045039	S.D. dependent var	0.001310
S.E. of regression	0.001280	Akaike info criterion	-10.47236
Sum squared resid	0.000267	Schwarz criterion	-10.43472
Log likelihood	865.9700	Hannan-Quinn criter.	-10.45708
F-statistic	8.734802	Durbin-Watson stat	2.144674
Prob(F-statistic)	0.003585		

Looking at a correlogram of squared residuals we can confirm persistence in lags and this is why the above test detected arch effects. To put it in another words, we can see autocorrelation on squared residuals meaning there is ARCH effect from the figure below:

Figure 4.6: Autocorrelation on Squared Residuals

Date: 09/18/21 Time: 14:43
Sample: 1 166
Included observations: 166

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.225	0.225	8.5889	0.003
		2	0.354	0.319	29.883	0.000
		3	0.101	-0.029	31.639	0.000
		4	0.049	-0.089	32.050	0.000
		5	0.060	0.050	32.682	0.000
		6	0.171	0.202	37.768	0.000
		7	0.148	0.081	41.584	0.000
		8	0.524	0.455	90.058	0.000
		9	0.222	0.035	98.797	0.000
		10	0.273	-0.076	112.09	0.000
		11	0.060	-0.102	112.74	0.000

Heteroskedasticity Test: ARCH for higher Autoregressive orders

Trying higher Autoregressive orders, the heteroskedasticity test for arch effect was significant at 5% level of significance ($\chi 2(1) = 2.684926$, $p=0.0080$) concluding that there is also a significant arch effect at order (2) just like there was at order(1).

Table 4.8: Heteroskedasticity Test: ARCH for Order (2)

Heteroskedasticity Test: ARCH

F-statistic	24.27539	Prob. F(2,161)	0.0000
Obs*R-squared	37.99713	Prob. Chi-Square(2)	0.0000

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 09/19/21 Time: 14:12
Sample (adjusted): 3 166
Included observations: 164 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.000226	0.000111	2.025159	0.0445
RESID^2(-1)	0.352109	0.077094	4.567297	0.0000
RESID^2(-2)	0.206991	0.077094	2.684926	0.0080

R-squared	0.231690	Mean dependent var	0.000511
Adjusted R-squared	0.222146	S.D. dependent var	0.001503
S.E. of regression	0.001325	Akaike info criterion	-10.39600
Sum squared resid	0.000283	Schwarz criterion	-10.33929
Log likelihood	855.4718	Hannan-Quinn criter.	-10.37298
F-statistic	24.27539	Durbin-Watson stat	1.975697
Prob(F-statistic)	0.000000		

By analyzing the results of the arch effect at order (3), we find out that the arch effect isn't significant at 5% significance with a value of 0.4559, on the other hand we can also observe that the variance coefficient is negative. The figure below depicts the results to show that the arch effect is not significant at this order.

Table 4.9: Heteroskedasticity Test: ARCH for Order (3)

Heteroskedasticity Test: ARCH

F-statistic	16.19363	Prob. F(3,159)	0.0000
Obs*R-squared	38.14747	Prob. Chi-Square(3)	0.0000

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 09/19/21 Time: 14:14
Sample (adjusted): 4 166
Included observations: 163 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.

C	0.000240	0.000114	2.112440	0.0362
RESID^2(-1)	0.364172	0.079164	4.600239	0.0000
RESID^2(-2)	0.227698	0.082316	2.766132	0.0063
RESID^2(-3)	-0.059167	0.079153	-0.747504	0.4559

R-squared	0.234034	Mean dependent var	0.000513
Adjusted R-squared	0.219581	S.D. dependent var	0.001507
S.E. of regression	0.001331	Akaike info criterion	-10.38102
Sum squared resid	0.000282	Schwarz criterion	-10.30510
Log likelihood	850.0529	Hannan-Quinn criter.	-10.35019
F-statistic	16.19363	Durbin-Watson stat	2.004927
Prob(F-statistic)	0.000000		

Adding Arch and Garch terms

Next both Arch and Garch coefficient were estimated. From the arch model, we see that both the Arch (alpha) and Garch (beta) coefficients are largely significant. The significance of alpha tells us that arch effect does exist. The asymmetric term RESID (-1)^2*(RESID(-1)<0), also known as gamma is positive as expected and statistically significant (indicating presence of *leverage* effect) in other words negative sentiment has a higher impact than positive sentiments. The sum of the alpha and gamma coefficients is close to 1, which means the shock to the conditional variance is highly persistent. Because the Garch coefficient is significant, large values of Log return are expected to yield extreme forecasts for prolonged period of time. Negative shocks (resid), on the other hand increases the variance more than the positive shocks.

Table 4.10: GARCH Model output

Dependent Variable: LOGRETURN
Method: ML ARCH - Normal distribution (BFGS / Marquardt steps)
Date: 09/19/21 Time: 18:34
Sample (adjusted): 2 166
Included observations: 165 after adjustments
Convergence achieved after 55 iterations
Coefficient covariance computed using outer product of gradients
Presample variance: backcast (parameter = 0.7)
GARCH = C(5) + C(6)*RESID(-1)^2 + C(7)*RESID(-1)^2*(RESID(-1)<0) +
C(8)*RESID(-2)^2 + C(9)*GARCH(-1)

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-0.006393	0.002181	-2.931878	0.0034
SENTIMENT	0.335064	0.115850	2.892224	0.0038

S2	-3.123296	1.369190	-2.281126	0.0225
AR(1)	-0.121400	0.099129	-1.224662	0.2207

Variance Equation

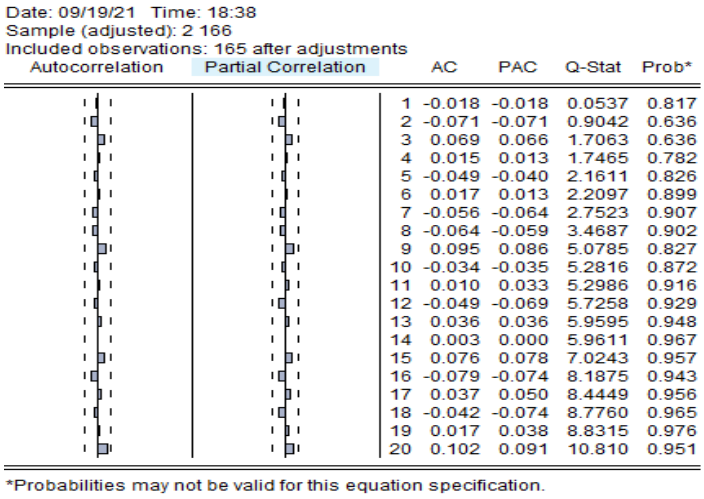
C	4.30E-06	3.82E-06	1.126823	0.2598
RESID(-1)^2	-0.012867	0.058028	-0.221731	0.0845
RESID(-1)^2*(RESID(-1)<0)	0.763978	0.194645	3.924977	0.0001
RESID(-2)^2	0.471773	0.192805	2.446899	0.0144
GARCH(-1)	0.465340	0.106057	4.387652	0.0000

R-squared	0.182662	Mean dependent var	6.56E-05
Adjusted R-squared	0.167432	S.D. dependent var	0.025540
S.E. of regression	0.023304	Akaike info criterion	-5.715198
Sum squared resid	0.087433	Schwarz criterion	-5.545783
Log likelihood	480.5038	Hannan-Quinn criter.	-5.646426
Durbin-Watson stat	2.558679		

Inverted AR Roots	-.12

The residual autocorrelation plot shows no significant autocorrelation on this final model (refer to the figure below). Just as expected (similar to what we saw on the OLS model) Both the linear and quadratic coefficient of sentiment are significant at 5% level after these adjustments, this means When sentiment is 0 (neutral sentiment) the slope is such that Log Return would increase by 0.335064 on average for every unit increase in sentiment, but this slope would keep declining by 3.12 on average for every unit increase in sentiment score. In other words, as seen on figure 1 the slope of sentiment keeps decreasing as sentiment score increase hence the *negative S2.*

Figure 4.7: Autocorrelation plots after the adjustments



	AC	PAC	Q-Stat	Prob*
1	-0.018	-0.018	0.0537	0.817
2	-0.071	-0.071	0.9042	0.636
3	0.069	0.066	1.7063	0.636
4	0.015	0.013	1.7465	0.782
5	-0.049	-0.040	2.1611	0.826
6	0.017	0.013	2.2097	0.899
7	-0.056	-0.064	2.7523	0.907
8	-0.064	-0.059	3.4687	0.902
9	0.095	0.086	5.0785	0.827
10	-0.034	-0.035	5.2816	0.872
11	0.010	0.033	5.2986	0.916
12	-0.049	-0.069	5.7258	0.929
13	0.036	0.036	5.9595	0.948
14	0.003	0.000	5.9611	0.967
15	0.076	0.078	7.0243	0.957
16	-0.079	-0.074	8.1875	0.943
17	0.037	0.050	8.4449	0.956
18	-0.042	-0.074	8.7760	0.965
19	0.017	0.038	8.8315	0.976
20	0.102	0.091	10.810	0.951

*Probabilities may not be valid for this equation specification.

Adding Sentiment as a Regressor to the GARCH Equation

With ARCH of order 1 we see that sentiment is not significant predictor of the conditional variance, which means it's not worth adding to the variance equation. The AIC is also larger than we had on the previous model.

Table 4.11: GARCH Equation with ARCH of Order (1) with Sentiment (Regressor)

Dependent Variable: LOGRETURN
Method: ML ARCH - Normal distribution (BFGS / Marquardt steps)
Date: 09/22/21 Time: 09:14
Sample (adjusted): 11/04/2019 6/30/2020
Included observations: 165 after adjustments
Convergence achieved after 58 iterations
Coefficient covariance computed using outer product of gradients
Presample variance: backcast (parameter = 0.7)
GARCH = C(5) + C(6)*RESID(-1)^2 + C(7)*GARCH(-1) + C(8)*SENTIMENT

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-0.003802	0.002185	-1.740205	0.0818
SENTIMENT	0.214214	0.114449	1.871698	0.0612
S2	-1.824767	1.254889	-1.454126	0.1459
AR(1)	-0.171129	0.099359	-1.722320	0.0850
Variance Equation				
C	7.59E-06	5.52E-06	1.375427	0.1690
RESID(-1)^2	0.279544	0.102704	2.721828	0.0065
GARCH(-1)	0.773413	0.072480	10.67076	0.0000
SENTIMENT	-0.000102	9.20E-05	-1.113028	0.2657

R-squared	0.175976	Mean dependent var	6.56E-05
Adjusted R-squared	0.160621	S.D. dependent var	0.025540
S.E. of regression	0.023399	Akaike info criterion	-5.666394
Sum squared resid	0.088148	Schwarz criterion	-5.515803
Log likelihood	475.4775	Hannan-Quinn criter.	-5.605264
Durbin-Watson stat	2.432422		
Inverted AR Roots	-.17		

The same can also notice also using ARCH of order (2) with GARCH of order (1) Garch to fit the model, but the model fit achieved is still less than we achieved without including sentiment score as a Regressor.

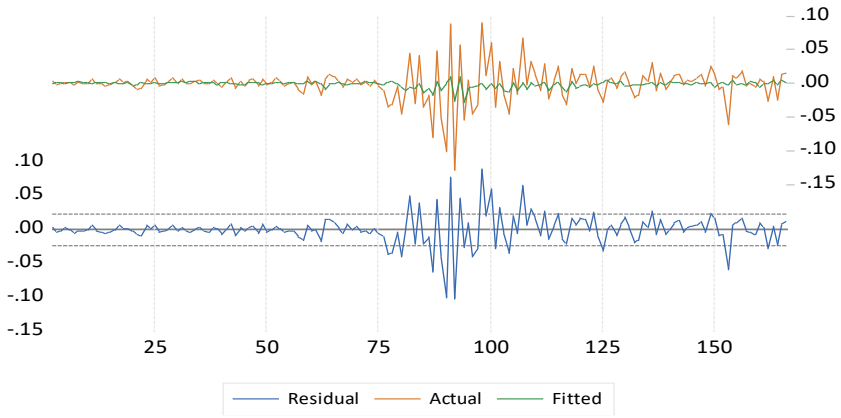
Table 4.12: GARCH Equation with ARCH of Order (2) and GARCH (1) with Sentiment

Dependent Variable: LOGRETURN
Method: ML ARCH - Normal distribution (BFGS / Marquardt steps)
Date: 09/22/21 Time: 09:24
Sample (adjusted): 11/04/2019 6/30/2020
Included observations: 165 after adjustments
Failure to improve likelihood (non-zero gradients) after 32 iterations
Coefficient covariance computed using outer product of gradients
WARNING: Singular covariance - coefficients are not unique
Presample variance: backcast (parameter = 0.7)
GARCH = C(5) + C(6)*RESID(-1)^2 + C(7)*RESID(-1)^2*(RESID(-1)<0) +
C(8)*RESID(-2)^2 + C(9)*GARCH(-1) + C(10)*SENTIMENT

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-0.018395	NA	NA	NA
SENTIMENT	0.947310	NA	NA	NA
S2	-8.385617	NA	NA	NA
AR(1)	-0.102474	NA	NA	NA
Variance Equation				
C	0.000322	NA	NA	NA
RESID(-1)^2	-0.013283	NA	NA	NA
RESID(-1)^2*(RESID(-1)<0)	0.132669	NA	NA	NA
RESID(-2)^2	0.344791	NA	NA	NA
GARCH(-1)	0.117426	NA	NA	NA
SENTIMENT	-0.004097	NA	NA	NA
R-squared	0.217663	Mean dependent var		6.56E-05
Adjusted R-squared	0.203086	S.D. dependent var		0.025540
S.E. of regression	0.022799	Akaike info criterion		-5.396879
Sum squared resid	0.083689	Schwarz criterion		-5.208640
Log likelihood	455.2426	Hannan-Quinn criter.		-5.320467
Durbin-Watson stat	2.434582			
Inverted AR Roots	-.10			

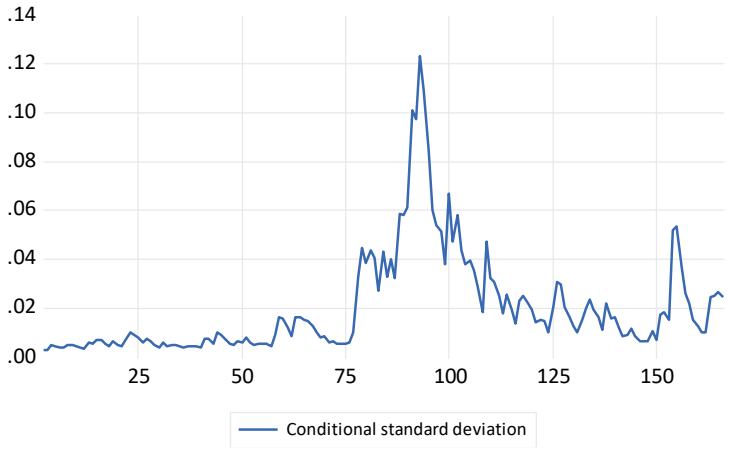
A look on a graph comparing predicted values (fitted) to the actual (observed values) shows that, the trend of log returns was captured. Despite that we still manage to observe a high volatility past 75th observation which was not well captured. There is no reason we can attribute to that, we only know that there is volatility clustering at that point we can't determine why. Its cause can be external factors i.e. economic issue, geographical issue etc.

Figure 4.8: Predicted Values Vs Actual Values



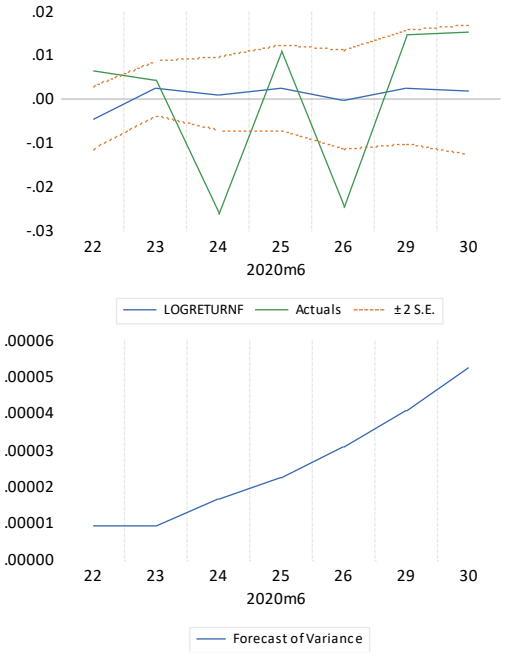
Similarly a graph of conditional standard deviation shows a spike at this section was also modelled in the figure below:

Figure 4.9: Conditional Standard Deviation



The graph below shows forecasts of the last 10 days of the observation period. The Mean Squared Error (MSE) for the prediction was found to be 0.0162, whereas the Mean Absolute Error MAE was found to be 0.014. The forecasted volatility values for the last 10 days increases gradually

Figure 4.10: The Forecasted Volatility values for the Last 10 days



Forecast: LOGRETURNF	
Actual: LOGRETURN	
Forecast sample: 6/22/2020 6/30/2020	
Included observations: 7	
Root Mean Squared Error	0.016238
Mean Absolute Error	0.014006
Mean Abs. Percent Error	94.63600
Theil Inequality Coef.	0.854023
Bias Proportion	0.001932
Variance Proportion	0.764101
Covariance Proportion	0.233967
Theil U2 Coefficient	0.842423
Symmetric MAPE	153.2459

5. DISCUSSION AND CONCLUSION

5.1 Analysis of the Study Results

The study of volatility in the stock market seem to be the most important research currently given most of the financial institutions, traders and investors tend to use it with a projection of risk management as well as financial gain. An ability to determine the stock volatility in a time series will not only help the investor know more about the stock movement, but can also help one towards forecasting of the upcoming stock returns. The impact of public sentiment can be very crucial towards decision making for most of the people in different aspects e.g. personal matters, economic, social and even political aspects.

From the findings, the asymmetric term RESID (-1)^2*(RESID(-1)<0), also known as gamma is positive as expected and statistically significant (indicating presence of *leverage* effect) in other words negative sentiment has a higher impact than positive sentiments. The sum of the alpha and gamma coefficients is close to 1, which means the shock to the conditional variance (volatility) is highly persistent. Moreover, Negative shocks (resid), on the other hand increases the variance more than the positive shocks.

5.2 Suggestions for Future Work.

This section covers several ideas we have come up with which we think could help improve the performance of forecasting for ARCH and GARCH models.

A long window is very important if we will need to improve the performance of our model's precision. Given the complexity computation and non-linearity issue with the data in a time series, a period of one year of data is deemed not to be sufficient at all. Testing our models over a bigger

period of time will not only give us more precise results in forecasting but will also help us understand how the stock market behaves during different situations.

Although Both ARCH and GARCH models have the ability of capturing volatility and it's clustering, these models are unable to model the leverage effect given their symmetrical distribution. Nelson (1991), suggested EGARCH (Exponential GARCH), a further extension to the known GARCH model to address the non-symmetric distribution problem especially the leverage effect.

The size of the sample for analysis is also another big factor that should be considered when running forecasting tests on a given set of data. Having more observational and fitting data will definitely improve tour forecasting performance.

All in all, there is a belief that with addition of further extensions to our already present models, the forecasting performance of our model could be improved. Just like GARCH was an improvement on ARCH, we believe that further addition to combat problems such as the leverage effect etc. could also help us yield better results and performances.

5. REFERENCES

Ahuja R, Rastogi H, Choudhuri A, Garg B (2015) Stock market forecast using sentiment analysis. In: IEEE 2nd international conference on computers for sustainable global development, pp 1008–1010

Al-Augby, S.H. (2015). Text mining method in evaluation of media's impact on market value.

Beirne, J.; Caporale, G. M.; Ghattas, M. S.; Spagnolo, N. 2013. Volatility spillovers and contagion from mature to emerging stock markets, Review of International Economics 21(5): 1060–1075. https://doi.org/10.1111/roie.12091

Bing L, Chan KC, Ou C. Public sentiment analysis in Twitter data for prediction of a company's stock price movements. In: 2014 IEEE 11th International Conference on e-Business Engineering. IEEE; 2014. pp. 232-239

Bo Pang, Lillian Lee and Vaithyananthan (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Vol. 10, pp 79-86.

Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Compu-tational Science, 2(1), 1-8 (2011)

Bollerslev, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return. The review of economics and statistics, 542-547

Cakra, Y.E and B.D. Trisedya (2015). Stock price prediction using linear regression based on sentiment analysis, pp: 149-154.

Cordelia Schmid. Bag of Features for category classification, www.cs.umd.edu/~djacobs/CMSC426/BagofWords.pdf

Dani Yogatama and Noah A. Smith, (2014), Making the Most of Bag of Words: Sentence Regularization with Alternating Direction Method of Multipliers, Proceedings of the 31st International Conference on Machine Learning, Beijing, China, JMLR: W&CP, volume 32

Daniel, M. R. F. Neves and N. Horta (2017). Company event popularity for financial markets using Twitter and sentiment analysis, Journal 71: 120-124.

Ederington, L. H., & Guan, W. (2005). Forecasting Volatility. Journal of Futures Markets, 25(5), 465–490. http://dx.doi.org/10.1002/fut.20146

Engle, Robert F. and Victor Ng. 1993. "Measuring and Testing the Impact of News on Robert Engle 167 Volatility."

Engle, R. F., Ng, V. K., Rothschild, M. (1990). Asset pricing with a factor-ARCH covariance structure: Empirical estimates for treasury bills. Journal of Econometrics, 45(1-2), 213-237

Liu, H., Erdem, E., Shi, J. (2011). Comprehensive evaluation of ARMA–GARCH (-M) approaches for modeling the mean and volatility of wind speed. Applied Energy, 88(3), 724-732

Faten Subhi Alzazah and Xiaochun Cheng (June 1st 2020). Recent Advances in Stock Market Prediction Using Text Mining: A Survey, E-Business - Higher Education and Intelligence Applications, Robert M.X. Wu and Marinela Mircea, IntechOpen, DOI: 10.5772/intechopen.92253

Fung G.P.C., Yu. J.X and Lam.W (2002). News Sensitive Stock Trend Prediction. In Proceedings of the 6th Pacific-Asia Conference(PAKDD) on Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science, Springer,Vol.2336, pp481- 493.

Harris, R. D. F.; Pisedtasalasai, A. 2006. Return and volatility spillovers between large and small stocks in the UK, Journal of Business Finance & Accounting 33(9–10): 1556–1571. https://doi.org/10.1111/j.1468-5957.2006.00635.x

Ghanavati M, Wong RK, Chen F, Wang Y, Fong S. A generic service framework for stock market prediction. In: 2016 IEEE International Conference on Services Computing (SCC). IEEE; 2016. pp. 283-290

G. W. Brown and M. T. Cliff, "Investor sentiment and the near-term stock market," Journal of Empirical Finance, vol. 11, no. 1, pp. 1–27, 2004.

Hamed, A.-R., Qiu, R., & Li, D. (2015). Analysis of the relationship between Saudi twitter posts and the Saudi stock market. Paper presented at the 2015 IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS).

Hassan, H. G., Bakr, H. M. A., & Ziedan, I. E. (2018). A Framework for Arabic Concept-Level Sentiment Analysis using SenticNet. *International Journal of Electrical and Computer Engineering (IJECE)*, *8*(5), 4015. https://doi.org/10.11591/ijece.v8i5.pp4015-4022

Higgins, M. L., Bera, A. K. (1992). A class of nonlinear ARCH models. International Economic Review, 137-158

Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment Analysis of Short Informal Texts. *Journal of Artificial Intelligence Research*, *50*, 723–762. https://doi.org/10.1613/jair.4272

Keshavarz, H., & Abadeh, M. S. (2017). ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysis of microblogs. *Knowledge-Based Systems*, *122*, 1–16. https://doi.org/10.1016/j.knosys.2017.01.028

Nayak, A., M. M. Pai and R. M. Pai (2016). Prediction models for Indian stock market Procedia Comput. Sci., 89: 444-452

Nguyen TH, Shirai K, Velcin J (2015) Sentiment analysis on social media for stock movement prediction. Exp Syst Appl 42(24):9603–9611

Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2020). Predicting Stock Market Price Movement Using Sentiment Analysis: Evidence from Ghana. *Applied Computer Systems*, *25*(1), 33–42. https://doi.org/10.2478/acss-2020-0004

Oliveira N, Cortez P, Areal N (2013) On the predictability of stock market behavior using stocktwits sentiment and posting volume. Portuguese conference on artificial intelligence. Springer, Berlin, pp 355–365

Panait, I., & Slăvescu, F. O. (2012). Using GARCH-in-mean model to investigate volatility and persistence at different frequencies for Bucharest Stock Exchange during 1997 - 2012. Theoretical and Applied Economics, 19(5), 55-76.

Paramanik, R. N., & Singhal, V. (2020). Sentiment Analysis of Indian Stock Market Volatility. *Procedia Computer Science*, *176*, 330–338. https://doi.org/10.1016/j.procs.2020.08.035

A. Porshnev, I. Redkin and A. Shevchenko, "Machine Learning in Prediction of Stock Market Indicators Based on Historical Data and Data from Twitter Sentiment Analysis," 2013 IEEE 13th International Conference on Data Mining Workshops, 2013, pp. 440-444, doi: 10.1109/ICDMW.2013.111.

Qiu, L., H.Rui and A. Whinston, 2013. Social network-embedded prediction markets: The effects of information acquisition and communication on predictions. Decision Support Systems. 55: 978-987

Ren, R., Wu, D. D., & Liu, T. (2019). Forecasting Stock Market Movement Direction Using Sentiment Analysis and Support Vector Machine. *IEEE Systems Journal*, *13*(1), 760–770. https://doi.org/10.1109/jsyst.2018.2794462

Salam, A., Noor, A. and Hussein M. Stock Market prediction using Twitter Sentiment Analysis Based on Social Network? Analytical Study

Sakthivel, P.; Bodkhe, N.; Kamaiah, B. 2012. Correlation and volatility transmission across international stock markets: a bivariate GARCH analysis, International Journal of Economics and Finance 4(3). https://doi.org/10.5539/ijef.v4n3p253

Saunders, M. and M. M. Cornett (2015). Financial institutions management. A Risk Management Approach. Irwin-McGraw-Hill, New York.

Shruti Shakhla, Bhavya Shah, Niket Shah, Vyom Unadkat, and Pratik Kanani (2018). Stock price trend prediction using multiple linear regression.

Şimşek, M. U., & Özdemir, S. (2012). Analysis of the relation between Turkish twitter messages and stock market index. Paper presented at the Application of Information and Communication Technologies (AICT), 2012 6th International Conference on.

Skuza, M and A. Romanowski (2015). Sentiment analysis of Twitter data within big data distributed environment for stock prediction.

Sutheebanjard, P., & Premchaiswadi, W. (2010). Forecasting the Thailand stock market using evolution strategies. Asian Academy of Management Journal of Accounting and Finance, 6(2), 85–114.

Taboada, M. (2016). Sentiment analysis: An overview from linguistics. Ann. Rev. Ling, 2:330-341.

Tsay, R. S. (2010), “Analysis of financial time series”, Wiley, third edition

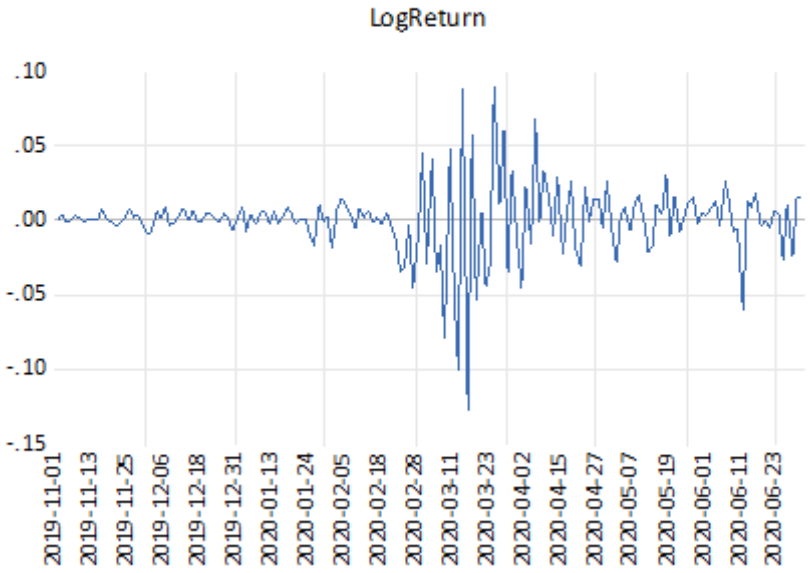
Valle-Cruz, D., Fernandez-Cortez, V., López-Chau, A., & Sandoval-Almazán, R. (2021). Does Twitter Affect Stock Market Decisions? Financial Sentiment Analysis during Pandemics: A Comparative Study of the H1N1 and the COVID-19 Periods. *Cognitive Computation*. https://doi.org/10.1007/s12559-021-09819-8

Wang, Y.; Gunasekarage, A.; Power, D. M. 2005. Return and volatility spillovers from developed to emerging capital markets: the case of South Asia, Contemporary Studies in Economics and Financial Analysis 86: 139–166. https://doi.org/10.1016/S1569-3759(05)86007-3

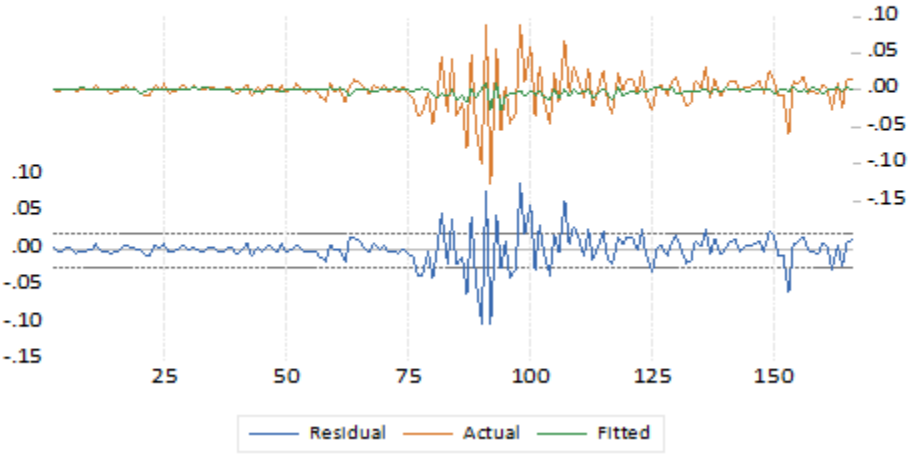
Zhou Z, Zhao J, Xu K (2016) Can online emotions predict the stock market in China? In: International conference on web information systems engineering, pp 328–342

APPENDICES

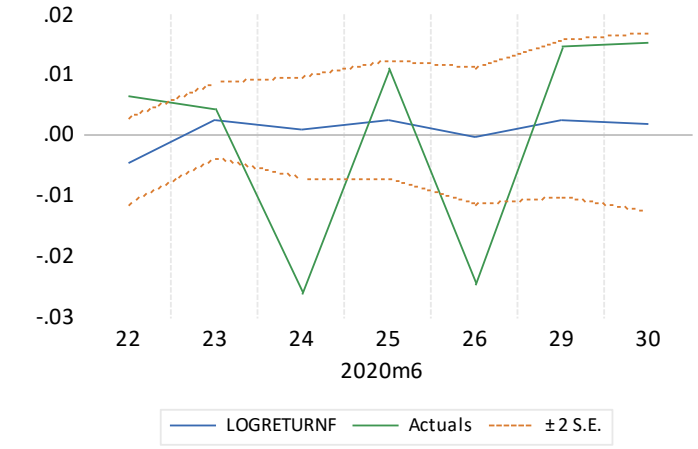
Graph of Log Returns Volatility in Time Series

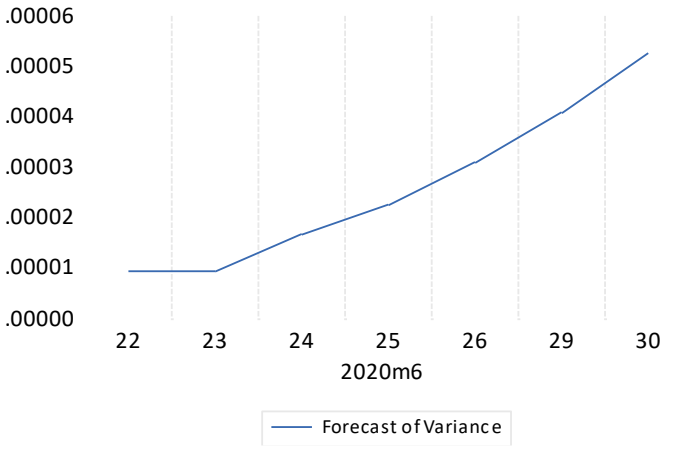


Comparison between Fitted and Observed Values



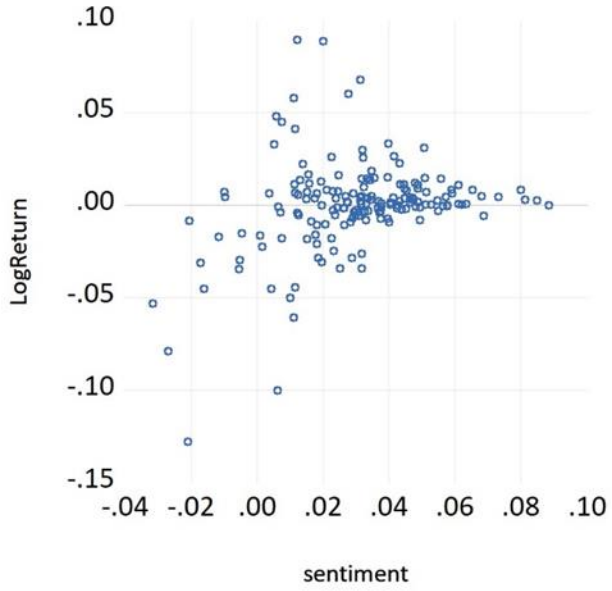
Forecast For the last 10 days of Log Return using sentiment scores





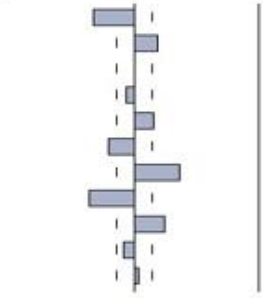
Forecast: LOGRETURNF	
Actual: LOGRETURN	
Forecast sample: 6/22/2020 6/30/2020	
Included observations: 7	
Root Mean Squared Error	0.016238
Mean Absolute Error	0.014006
Mean Abs. Percent Error	94.63600
Theil Inequality Coef.	0.854023
Bias Proportion	0.001932
Variance Proportion	0.764101
Covariance Proportion	0.233967
Theil U2 Coefficient	0.842423
Symmetric MAPE	153.2459

Plot of LogReturn v/s Sentiment scores Linearity Test



Residual Autocorrelation Test

Date: 09/07/21 Time: 12:45
Sample: 11/01/2019 6/30/2020
Included observations: 166



	AC	PAC	Q-Stat	Prob
1	-0.346	-0.346	20.229	0.000
2	0.192	0.082	26.497	0.000
3	0.003	0.105	26.498	0.000
4	-0.072	-0.066	27.390	0.000
5	0.161	0.114	31.861	0.000
6	-0.215	-0.131	39.900	0.000
7	0.379	0.292	65.131	0.000
8	-0.375	-0.215	89.907	0.000
9	0.250	0.080	100.99	0.000
10	-0.094	-0.029	102.57	0.000
11	0.027	0.092	102.70	0.000

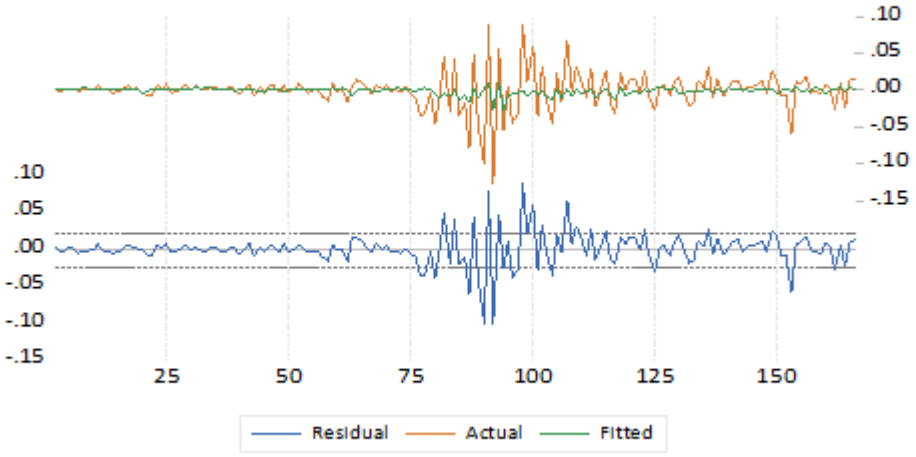
Ordinary Least Squares Regression Output.

Dependent Variable: LOGRETURN
Method: Least Squares
Date: 09/07/21 Time: 12:44
Sample: 11/01/2019 6/30/2020
Included observations: 166

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.014728	0.003217	-4.577753	0.0000
SENTIMENT	0.974267	0.166564	5.849207	0.0000
S2	-10.61753	2.507203	-4.234812	0.0000

R-squared	0.184897	Mean dependent var	6.52E-05
Adjusted R-squared	0.174896	S.D. dependent var	0.025462
S.E. of regression	0.023129	Akaike info criterion	-4.677588
Sum squared resid	0.087194	Schwarz criterion	-4.621347
Log likelihood	391.2398	Hannan-Quinn criter.	-4.654759
F-statistic	18.48734	Durbin-Watson stat	2.690515
Prob(F-statistic)	0.000000		

Predicted Values Vs Actual Values



Graph of Log Returns Volatility