Stock market prediction based on machine learning and social sentiment analysis

Mustansar Ali Ghazanfar[1], Madiha Anwar[1], Sin Wee Lee[2], Nadeem Qazi[1], Amin Karimi[1], N.Z Jhanjhi[3,] Ali Javed[4]

[1] School of Architecture, Computing and Engineering, University of East London UK
mghazanfar@uel.ac.uk, u2047200@uel.ac.uk, n.qazi@uel.ac.uk, a.karimi@uel.ac.uk
[2] Deputy Head of School of Computing, Arden University, slee@arden.ac.uk
[3] School of Computer Science SCS, Taylor's University, Malaysia, noorzaman.jhanjhi@taylors.edu.my
[4] Department of Software Engineering, UET Taxila, Pakistan ali.javed@uettaxila.edu.pk

***Abstract* — Precise stock market prediction is crucial for investors, but the volatility of the stock market is influenced by multiple factors such as public sentiments, business news, and related product volatility. While several algorithms have been proposed to predict the stock exchange index based on historical data, they are not ideal as external factors play a critical role in market volatility. To address this issue, we proposed a machine learning model that incorporates historical data with external factors such as social media sentiments, oil and gold trends, and financial news data to enhance prediction accuracy. Our study used HPQ, IBM, ORCL, and MSFT stock market datasets to validate the effectiveness of the proposed model, including an analysis of the impact of Covid19 on companies. Our experimental results showed the highest accuracy of 87.2% using oil and sentiment datasets. Additionally, we identified that social media significantly affects IBM stocks, and the GBM (Gradient Boosting Classifier) classifier produced consistent results.**

Keywords —Stock Market Prediction, Deep learning, Covid19 Impact, Feature selection, Oil/Gold price impact, Sentiment analysis,

1. INTRODUCTION

The stock market plays a vital role in the economies of all countries. It provides companies with the opportunity to raise funds by selling stocks to investors, while also allowing investors to participate in the financial success of these firms, earning profits through capital gains and dividends, despite the risk of potential losses.

Stock investors rely on market trend predictions to make informed decisions about when to buy or sell stocks. To maximize profits, investors seek to purchase stocks that are projected to increase in value and sell those that are expected to decrease in value. Accurate prediction of the stock market is not a simple task, as it is influenced by numerous variables, including the impact of social media and commodity prices, such as oil. The impact of these factors can be either positive or negative, thus necessitating their consideration in accurately predicting the stock market.

Investing in the stock market is inherently risky; however, it can also offer considerable profits if done properly. To mitigate the risk of buying volatile stocks, investors typically conduct assessments of a company's performance, examining factors such as its social media presence, financial news, and the performance of linked goods or companies. However, due to the sheer volume of data from social media and financial news sources, investors cannot fully analyze all of this information manually. An automated decision-making assistance system is necessary to process and evaluate stock trends using large amounts of data. Machine learning techniques can be employed to create such a system, and identifying algorithms that best utilize external data, such as social media information and oil prices, to predict stock market trends is critical. Machine learning researchers have shown significant interest in this area, as accurate stock predictions based on external factors can enhance investor profits.

Prior research on stock market prediction has utilized historical data (Shen et al., 2012; Hegazy et al., 2014; Chen et al., 2018) or social media data (Khatri and Srivastava, 2016; Zhou et al., 2016; Urolagin, 2017) for prediction purposes using machine learning algorithms. Numerous prediction systems utilizing diverse types of data have been proposed, providing valuable insights to investors to help them make informed decisions regarding buying or selling a particular company. However, relying solely on a single type of data may not necessarily lead to improved accuracy in stock market prediction.

In the past, historical data has been analyzed using mathematical techniques for technical analysis to predict future stock market trends (Dang et al., 2018). Using historical stock pricing data, researchers have employed various machine learning methods, including regression analysis (Jeon et al., 2018) and deep learning (Li et al., 2014a, b, c). However, it is also necessary to consider external variables, as unforeseen events mentioned on social media can have a significant impact on stock prices.

Fig. 1. Showing how social media affects the stock market trends

Social media is a comparatively recent kind of internet content. The quick availability of new information and the rapid interaction between the users is one of its main qualities. This kind of engagement may be interpreted as a measure of consumers' interest in a large variety of issues, which includes the stock market. However, just social media has little effect on the behaviour of stock traders and, as a result, stock markets. People seeking to invest in the stock market are frequently uninformed of the market's behaviour. Thus they aren't able to predict the stocks that need to be purchased or sold for maximizing their earnings. These investors understand that the growth of the stock market is dependent on relevant news and other goods or company stocks as well. Hence, they require accurate and fast information regarding stock market listings for making trading decisions based on timely and correct information. This information could be gained from any social media platform, but most of the social media platforms do have a large amount of spam or unstructured data. Twitter of all the other social media platforms has evolved for becoming a valuable source of information, which assists traders. However, as a trading technique, investors' expectations based just on financial news may be insufficient (Brown and Clif 2004). For stock prediction, existing predictive algorithms employed social media posts or oil or gold prices in addition to stock market data. No predictive method, in my opinion, has incorporated both sorts of data for stock prediction. The use of only one type of data might not provide the best forecast accuracy. Both data sources, which include social media and oil or gold price, can alter and impact the decisions of traders, thus when designing a prediction system for stock markets, both of these sources must be included. Taking into account both sources of prediction would improve the accuracy of the suggested prediction system. As a consequence, utilizing social media and the oil prices will result in an increased amount of accurate stock trend predictions. Figure 1 depicts a general illustration of the ways in which social media affects stock market trends.

Since external data sources for our proposed machine learning model, social media and oil prices would supply raw textual data as historical data and tweets. The raw data from social media is incomprehensible to machine learning algorithms. Pre-processing is a must to do on the data before using it for input to any machine learning algorithm. Natural language processing (NLP) is employed in the basic analysis method for analysing social media for identifying neutral, positive, or negative attitudes on the basis of the documentation content. Later, machine learning algorithms may be utilized for learning the relationship between text content sentiment and stock market trends.

For stock investors, an effective forecasting system is critical. Investors seek algorithms that could properly utilize enormous amounts of data. The stock data for predicting trends comprises a combination of textual and stock price data characteristics, a few of which are more significant than others for generating predictions. However, we aren't aware of the features that must be removed or selected. Hence, feature selection on the final datasets must come first in this process.

A high-quality prediction system that generates high-quality outcomes is extremely valuable to stock investors. Investors are looking for prediction algorithms that are accurate and capable of detecting spam data. Because of the growing usage of social media, spammers have begun to target it as well. Spammers utilize several Twitter accounts to promote their services and products by posting duplicate tweets (Sedhai and Sun 2018). Since they are very engaged in spam messaging on social media, spam tweets must be eliminated from the social media dataset. A prediction system that integrates past data as well as oil prices and social media for predictions and good performance after selection of features and decrease of spam tweets will benefit for accuracy improvement. The stock markets usually operate in different ways from one another. Certain people have different behaviour as a result of stock volatility; forecasting such stock markets is challenging. Identifying these stock markets can also help stock investors make decisions regarding trading. Moreover, stock investors are concerned about popular and common stock markets. More often, social media discusses the most popular stock markets, where investors are aimed at gaining knowledge about these stocks. Traders of stocks look for stock exchange that attract the interest of investors and talk about them on social media sites. Thus, stock traders must be able to identify these kinds of stock markets as well.

Deep neural networks have achieved several achievements in various domains recently, which include computer vision (He et al. 2016) and voice recognition (Noda et al. 2015). The idea of deep learning can also be applied in stock prediction because of its efficacy on large datasets (Li et al. 2014a, b, c).

Combination of classifiers is a common method in machine learning that has been proved to be higher in performance in comparison to the use one classifier (Tsai et al. 2011). With the use of ensemble methods in machine learning, various classifiers may be merged to improve the accuracy of the various classifiers.

Based on the identified gaps in existing research, we propose an innovative machine learning technique for stock market forecasting that incorporates social media posts and oil prices as external inputs. For social media data, we have chosen Twitter as the information hotspot for this study due to its succinctness (Tayal and Komaragiri, 2009). Historical data has been collected from Yahoo Finance for this purpose.

Following are the key contributions of this study:

- Advising on the use of a mix of oil prices trends and social media data to forecast stock market developments.
- Making recommendations for selecting various features to improve prediction performance.
- Proposing a reduction in spam tweets to increase algorithm prediction performance.
- Developing an algorithm that produces consistent outcomes.
- Advising on stock exchanges that are increasingly impacted by social media.
- Suggesting deep learning for stock market forecasting.

- Proposing a system for stock market prediction based on hybrid algorithms.
- Proposing a Covid19 impact analysis system on stock markets.

The rest of this research document is organized as follows. In Section two summarizes existing work in stock trend prediction. In Section three, we discuss study methods and detail each process. In Section four, we provide the suggested system's implementation specifics. Section five contains the experimental results as well as a commentary. In Section six, we provide our findings and recommendations for further research on this subject.

2. RELATED WORK

2.1 Sentiment Analysis

Popularity of sentiment analysis have grown over the past ten years as a result of the vast amounts of text based information accessible via social media sites. Such data can be extracted for several application areas to acquire users' perspectives. For sentiment analysis on this huge volume of textual data, both data extraction and machine learning are very crucial. Hence, machine learning experts have therefore carried out study on the basis of the views on using these social media sites.

Based on its content, tweets can be classified into numerous categories. Yuan (2016) investigated sentiment classification approaches based on rules, lexicons, and machine learning. Analysing the sentiments regarding the opinions of the users is done for various application areas. A team of data-scientist, Joshi and Tekchan-dani (2016) conducted a comparison analysis of different machine learning algorithms for classifying data regarding movie reviews from Twitter with the help of bigram, anagram and also by combining both features. Algorithms that were used in the study were NB, ME and SVM, and their study showed that SVM outperformed NB and ME. Moreover, a study by Qasem et al. (2015) was performed to assess neural network and logistic regression (LR). Neural network and logistic regression were assessed for tweets associated with technology stocks on the basis of two schemes, i.e., inverse document, bigram and unigram term frequency (TF). Stock markets for which this study was performed were Facebook , Twitter, Google, and Tesla. Study results showed that all classifiers gave same accuracy for negative, positive and neutral classes.

2.2 Stock market prediction

Below sections explains the purpose to use different types of data in this literature for the prediction of the stock market.

2.2.1 Price data

For building prediction models for the stock market, researchers have used a variety of machine learning approaches to extract social media and historical data. Prior to the widespread usage of social networking sites, stock price data was commonly utilized to forecast the stock market. Hegazy. et al. (2014) offered a solution of machine learning for price predictions of S&P 500 index. The model made use of price data and technical indicators from several stock markets represented by the S&P 500 index. Suggested algorithm combined the particle swarm optimization and least square SVM. The particle swarm optimization method is utilized for optimizing least square SVM after identifying the optimum parameter combination for forecasting daily stock prices. Shen. et al. (2012) developed a new prediction method that exploited temporal connections between global stock markets and other financial products for anticipating future trends in stock markets with the use of SVM, and the input parameter fed to it was stock price data. Additionally, Chen. et al. (2018) employed data of stock prices using CSI 300 index for stock market of China, on comparing the performance of traditional neural networks for price prediction and deep learning it revealed that the performance of deep learning outperformed standard neural nets. Yetis et al. (2014) used a feed-forward artificial neural network (FFANN) to forecast stock price data for stock markets and discovered that the performance of ANN performed well for NASDAQ to forecast stock value. Ou and Wang (2009) employed ten machine learning algorithms for stock market closing price for forecasting price index trends in the Hong Kong market. Lease square SVM and SVM were shown to outperform the other forecasting models in terms of predictive performance.

2.2.2 Social media data

For informing predictions of stock market, many machine learning researchers evaluated investor comments accessible on social media sites. These platforms include a wealth of information on businesses such that services and products they provide. Urolagin (2017) investigated link between a company's emotions from social media text and stock value from Yahoo Finance. He performed sentiment categorization using NB and SVM classifiers. N-gram feature vectors were created using most important terms from tweets text for this categorization. Moreover, the pattern of association between quantity of negative and positive tweets along with the stock value were investigated as well. He discovered a correlation between tweet characteristics such as the amount of positive, neutral, and negative tweets, as well as the overall number of tweets while using SVM to anticipate the stock market state.

Opinions given over social media tend to reflect the emotional condition of many of these sites' users. In tweets or comments, these individuals share their feelings about a firm or its products. The withdrawal of emotions from these tweets or remarks is used for detecting the user's opinion of a certain firm or product. For instance, Chakraborty et al. (2017) gathered tweets including the terms “AAPL”, “StockTwits”, and “stock market”, “AAPL” term tweets were utilised for predicting stock index trend of Apple Inc. stock market, whilst “StockTwits” term tweets along with the phrases “stock market” were used for predicting the trend of stock market. SVM classifier was used to classify sentiments, and boosted regression tree algorithm for predicting next day stocks difference. Khatri and Srivastava (2016) gathered comments and tweets from the Stock Twits3 website, whereas market data for Apple, Facebook, Oracle,

Google, and Microsoft was obtained from Yahoo Finance. The comments and tweets were divided into four categories, i.e., happy, rejected, up, and down. This stock data and polarity index were fed into an ANN for forecasting stock closing prices.

Past studies have found a link amongst the popular emotion represented in textual data of tweets and the stock prices. Yan et al. (2016), for example, presented the Chinese Profile of Mode States (C-POMS) model to evaluate attitudes shown in web-feeds of microblog. To identify the link they ran the test called Granger causality. This revealed a link amongst stock price series and C- POMS analysis. Predictions were made using SVM and a probabilistic neural network (PNN). From above tests, the experimental findings revealed that SVM outperformed PNN in forecasting stock market fluctuations.

For certain equities, financial experts communicate their analyses via Twitter. To uncover these insights from tweets, data mining methods and natural language processing (NLP) approaches might be used. Zhou et al. (2016), for instance, conducted a study to identify relation from 10 million tweets related to stock from Sina Weibo. It was discovered that distinct emotions exhibited in tweets, like fear, delight, dissatisfaction, and disgust, can forecast five aspects of Chinese stocks. Aspects of the Chinese stocks that were predicted include intra-day highest index, intra-day lowest index, trading volume, closing, and opening. In forecasting these characteristics of the stock market of China, their model beat baseline models. They predicted these characteristics using K-means discretization.

Various stock market-related events are shared on social media channels, which might have an impact on stock market results. These occurrences give meaningful categories for categorizing events, like loss or profits. Makrehchi et al. (2013) devised another approach to estimate feelings using stock market occurrences. An effective classifier was created to assess the emotions of tweets so that the information can be utilized to develop an efficient trading strategy.

2.3 Feature selection

Researchers have utilized a variety of approaches to pick features from datasets having numerous characteristics. Currently, researchers from machine learning recognize how significant feature selection is, when evaluating data of very high-dimensions it affects learning models, as well as increases computing time and also results in poor information (Cao et al. 2016; Cheng et al. 2013). Furthermore, because of these high numbers of characteristics, data-scientists confront the dimensionality curse that asserts that data becomes sparse within high-dimensional spaces (Cheng et al. 2013).

Researchers utilize two techniques to tackle high-dimensional data problems, i.e., feature selection and extraction. Here, for 1st method, we form another feature space of low-dimensions, whereas for 2nd approach, characteristics that are unnecessary and redundant are eliminated. From all the features we choose only a limited subset of most important features.

Swarm intelligence (SI) algorithms are increasingly being used for feature selection (Blum and Li 2008; Hassanien and Emary 2016). Reason behind its popularity is that it is widely used to solve optimization issues, and identifying optimum features is unquestionably one of them. SI algorithms have grown in popularity in past few years (Blum and Li 2008), at the present, 2 of its common algorithms include colony optimization (ACO) (Dorigo 1992) and PSO (Eberhart and Kennedy 1995).

Brezocnik et al. (2018) performed very detailed study on SI algorithm for feature selection in order to compare them. They discussed numerous application areas, strategies, methodologies, and settings for different elements. They discovered that SI methods were often tested for small datasets containing 150 or less characteristics. They discovered that SI algorithms were most commonly utilized for Bio-Medical Engineering field with a percentage of 60.53 and Computer Engineering field with a percentage of 28.95.

For instance, if we look into Bio-Medical Engineering field, Jayaraman and Sultana (2019) employed the gravitational cuckoo search method (Yang and Deb 2009) to select features for a heart disease classification problem. For Computer Engineering field, Wang and Dai (2013) utilised the artificial fish swarm method (Li 2003), Enache et al. (2015) used the bat algorithm (Yang 2010), while Seth and Chandra (2016) employed the grey wolf optimizer (Mirjalili et al. 2014) for feature selection in intrusion detection issues. Meanwhile, Mohammadi and Abadeh (2014) utilised new feature selection approach called IFAB which is based on the artificial bee colony algorithm (Karaboga 2005) for image steganalysis, another study done by Chhikara et al. (2018) employed the firefly algorithm (Yang 2008) for same problem. Other issues which are dealt inside this area include fault diagnosis in complex structures using the bacterial foraging optimization (Passino 2002) algorithm (Wang et al. 2016), enhanced facial identification using adaptive binary PSO (BPSO) (Sattiraju et al. 2013), recognition of web domains that looks malicious using BPSO (Hu et al. 2016), and categorizing web-pages with the help of ACO (Saraç and Zel 2014). Brezocnik et al. (2018) found from a study of several SI algorithms for feature selection that there isn't any one SI method that has the highest efficacy for feature selection.

A few researchers combined distinct strategies to create hybrid ways for feature selection. Ibrahim et al. (2019), for instance, coupled the method called slap swarm (Mirjalili et al. 2017) along with the PSO algorithm and discovered some performance and accuracy improvements. Likewise, Moslehi and Haeri (2019) presented another hybrid technique that included some genetic algorithm with PSO, from their study they discovered that it was capable of obtaining correct classification. Zhong and Enke (2016) investigated 3 approaches for feature selection, including fuzzy robust PCA, PCA, and kernel-based PCA, for minimising 60 characteristics of economic and financial data to estimate the S&P 500 prices. According to the findings of this study, PCA performed marginally better in terms of accuracy than the other algorithms.

2.4 Reduction of spam tweets

Social media is getting a lot of spam material which has piqued the interest of experts as well, owing to the difficulties that it causes. Several studies have been conducted on the

detection of spam tweets for a variety of applications. Sedhai and Sun (2018), for instance, presented a semi- supervised system for spam tweet identification. NB, LR, and RF classifiers were evaluated using the HSpam14 dataset (Sedhai and Sun 2015) for categorizing tweets text data as spam and ham. Chen et al. (2017a, b) developed a list of terms as blacklist for improving the performance of spam and low-quality material detection using their algorithm. For tweet classification they picked RF as a classifier.

Spam profiles, like spam tweets, are a source of unnecessary ads, posing a potential threat to social media site users. Al-Zoubi and Faris (2017) used information gain and ReliefF for extracting characteristics from Twitter spam accounts. Using the NB classifiers, K nearest neighbour (KNN), decision tree (DT), and multilayer perceptron (MLP), collected features were utilized to categorize and identify the spam profiles of Twitter users.

2.5 Oil/Gold price trends

The relationship between oil prices and macroeconomic factors has been studied extensively. Changes in oil prices could, in theory, have a range of effects upon economic growth. For example, oil price fluctuations may impact the availability of basic production inputs and investment costs (side-effects of supply), company production structures and unemployment, trade relations and redistribution of wealth from oil customers to production companies, financial policies, inflation and interest rates, costs, spending possibilities, and customer expectations and sentiments (side-effects of demand) [Hamilton (1983), Jones et al. (2004)]. Several empirical researches have indicated that rising oil prices get a significant negative influence on economic development in a variety of developed and developing countries [Cunado and Perez de Garcia (2005), Balaz and Londarev (2006), Gronwald (2008), and Cologni and Manera (2008)]. Nonetheless, in recent years, there appears to have been some signs of reverse causation among macroeconomic factors and oil prices. Due to rising demand, economic expansion would indeed be connected to an increase in oil prices [Barsky and Killian (2004)].

Moreover, certain theories imply that there is relationship between economic growth and oil prices is not linear entirely. The researchers also primarily concentrated on 3 potential reasons for antisymmetric macroeconomic variable reactions to oil price changes, namely, counter-inflationary corporate governance reactions to rising prices, investment uncertainty, and industry shock transmission mechanisms [Hamilton (1988), Mork et al. (1994), Ferderer (1996) and references therein]. Mork (1989) builds on Hamilton's (1983) research by establishing how oil prices and production increases have an uneven relationship. He shows that growth of economy does have a large negative relationship with oil price increases, but a small positive relationship with oil price declines. Haltiwanger and Davis (2001) distinguished among accumulated policy transmission (the effects of rising oil prices on potential outcomes, sticky incomes, and revenue transfer) and distributive policy transmission (the effects of rising oil prices on the closeness of the match between the actual and desired capital and labour levels). Because both rises and falls in oil prices will affect companies' desired employment patterns, the allocative transmission mechanisms must function asymmetrically, but the aggregate channels must be operating symmetrically. Haltiwanger and Davis (2001) used this research to show that the economy's response to rising oil prices are much larger than those to falling oil prices. According to Lee and Ni (2002), power-intensive businesses are more likely to experience price movements as supply shocks (aggregate effects), whereas less power-intensive businesses are more likely to experience demand shocks. Finally, a number of recent empirical research [Hamilton (2003), Zhang (2008), Lardic and Mignon (2006, 2008), and Cologni and Manera (2009)] utilising increasingly rigorous econometric methodologies support this asymmetric relationship among economic growth and oil shocks.

These studies indicate that rising oil prices stifle aggregate economic activity more as compared to falling prices do.

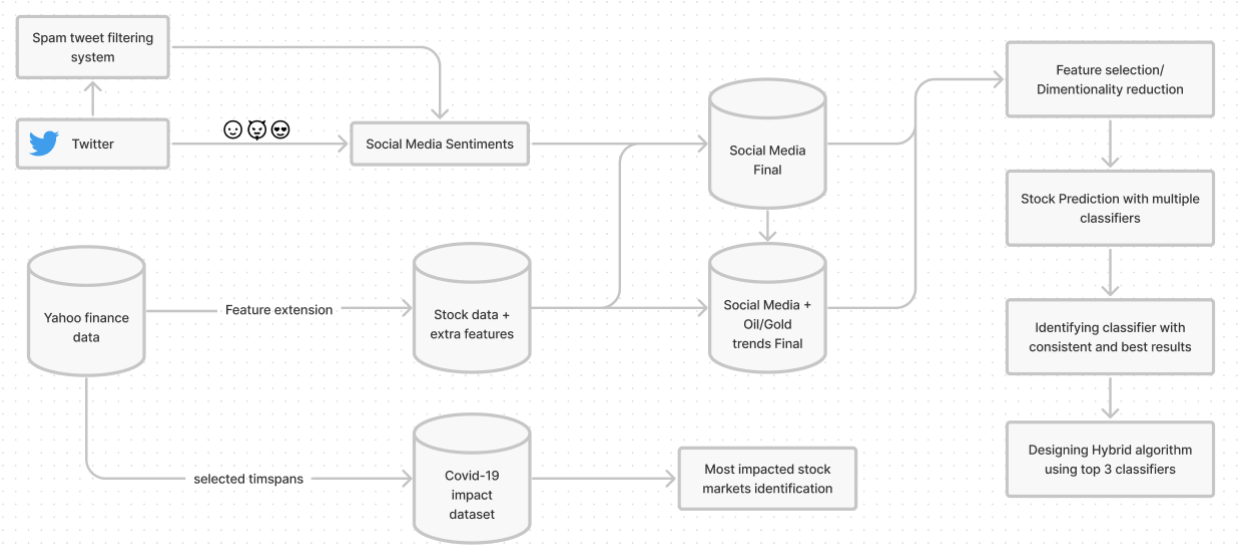

Fig. 2. Flow chart of the methodology steps for stock market prediction using oil/gold price trends and social media explained

Nevertheless, as previously stated, in recent years, economic development has been connected with rises in oil prices due to increased demand and investor mood [Barsky and Killian (2004), and Killian (2008, 2009)].

Jones and Kaul (1996) were the first to use the traditional valuation model for cash flow dividend to assess the sensitivity of international stock markets (UK, US, Japan, and Canada) to oil price changes. This article finds that the impact of oil shocks on cash flows may account for the whole reaction in the Canada and US. In the case of UK and Japan, the results were inconclusive. Using an unconstrained vector autoregressive (VAR) model, Huang et al. (1996) found no evidence of a link among market indexes and oil prices such as the S&P 500. Sadorsky (1999) applies GARCH with an unconstrained VAR to find influences on monthly American dataset and discovers a substantial short-term link between unanticipated oil price changes and returns on the S&P 500. Ciner (2001) employs causality tests using nonlinear approach to demonstrate that shocks in oil prices have a nonlinear impact on stock index returns in the United States, which is consistent with the stated impact of oil on economic activity. Oil prices have a negative impact on stock returns in the 12 European countries and US, according to Park and Ratti (2008), but not for stock markets in Norway, a positive response is observed from oil exporter on price increases. Apergis and Miller (2009) looked at whether structural oil-market shocks have an impact on stock price in 8 developed countries.

oil exporter, respond favourably to price rises. Apergis and Miller (2009) have investigated if structural oil-market shocks impact stock prices in eight industrialised nations. The authors discover, using several econometric approaches that international stock market returns don't provide any significant response to oil price shocks.

2.6 Stock market volatility

Because of the volatility of stock markets, their behaviour can't be reliably forecasted. Stock volatility can be influenced by external events such as financial crises. Various approaches have been used by researchers for identifying volatility of stock market. For example, Kumar and Patil (2015) employed a machine learning and time series approach to predict the fluctuation of the S&P 500. The standard deviation of the stock market was believed as a very good indicator of volatility.

The GARCH models have shown to be accurately forecasting the volatility of stock market. Omer and Halim (2015), for instance, examined 3 GARCH models and discovered that the exponential model from GARCH family beat the remaining models while predicting stock market volatility of Malaysia. A multi-kernal based extreme machine learning model was presented by Wang et al. (2014) for increasing volatility prediction performance utilizing historical market data and news.

No.	Stock Market	Ticker Symbol	Stock Exchange /Country	Raw Tweets count
1	London Stock Exchange Group	LSEG.L	United Kingdom	441
2	Hewlett-Packard Inc.	HPQ	NYSE	3,187

No.	Stock Market	Ticker Symbol	Stock Exchange /Country	Raw Tweets count
3	International Business Machines Corporation	IBM	NYSE	296,680
4	Microsoft Corporation	MSFT	NASDAQ	20,982
5	Oracle Corporation	ORCL	NYSE	35,396
6	Reliq Health Technologies Inc.	RHT.V	NYSE	4,923
7	Twitter, Inc.	TWTR	NYSE	23,014

Table 1: Summary of tweets with stock market ticker symbols

2.7 Deep learning for stock market prediction

There has been a significant growth in deep learning in a variety of disciplines during the last decade. It has been utilized by researchers in a variety of disciplines. Because of its efficacy on big datasets, it may also be used to anticipate stock trends. For example, Dang et al. (2018) built a network of TGRU to forecast S&P 500 stock market movements and discovered that its proposed model outperformed baseline approaches with very promising accuracy rate of 66.32%. Khare et al. (2017) used technical analysis to anticipate short-term price trend of the New York Stock Exchange using RNN and FFANN (NYSE). From the study it was discovered that FFANN outperformed other models for the prediction of short-term stock market. Li et al. (2017) utilized a neural network of long short-term memory for identifying sentiments of investor and price data of stock markets for forecasting CSI300 index value in the Chinese stock prices. For extracting emotions of investors NB algorithm has been used after extracting data from numerous forum posts.

2.8 Hybrid algorithm

When compared to separate classifiers, combining classifiers performed well. In several domains, researchers have utilised various ensemble approaches to improve the accuracies of individual classifiers. Todorovski and Deroski (2003), for instance, introduced meta DTs, a method for the combination of classifiers (MDTs). Stacking and voting ensemble approaches were used to integrate ordinary DTs, KNN, and NB algorithms. They made a comparison of these approaches and discovered that stacking outperformed voting. , NB, IBK, and J4.8 with stacking were examined by Džeroski and Ženko (2004) for constructing heterogeneous classifiers. From this study it was shown that utilizing cross-validation to select the most optimal classifier from the ensemble outperformed using the ensemble.

For enhancing stock return forecast in stock markets ensemble techniques may be utilized. Tsai et al. (2011) improved the prediction performance of LR, MLP, classification, and regression tree classifiers using bagging and majority voting ensemble techniques. They demonstrated that classifier ensembles outperformed individual classifiers regarding the accuracy of prediction. Kim et al. (2003) proved that combining numerous neural network classifiers with majority voting methodologies as well as genetic algorithms to predict client purchasing behaviour outperformed individual classifiers. Their findings also demonstrated that ensemble approaches did not differ significantly. Sun and Li (2012) combined various SVM classifiers using weighted majority voting to forecast financial hardship. Experimental results of their study revealed that the SVM classifier with ensemble outperformed many other SVM classifiers.

Generation of ensemble is a common method used to increase the accuracy of classifier decision making. In general, majority voting is the paradigm that classifiers use to make decisions. With the use of majority voting, Hajdu et al. (2013) created a system based of ensemble that detects the optic disc in retinal images. To create a positive and unlabeled learning system, Liu et al. (2018) combined multiple classifiers with 3 ensemble strategies defined as a majority combination rules, weighted vote, and weighted average. According to the data, it was revealed that individual classifiers performed worse than ensemble techniques based on weighted average and weighted vote.

3. METHODOLOGY

3.1 Data Collection

The section explains process of data collection, describes the source of data obtained, and the data structure. Table 1 shows a list of the stock markets chosen as case studies for this study, as well as the number of tweets they received. The stock exchanges in this table represent the overall stock markets, whilst all remaining stocks represents the stocks of a specific firm. It should be noted that stock market terminology would be emloyed interchangeably for referring to both the stock market of particular firms and the entire market. The chosen stock markets' stock market, social media, and oil or gold prices trend data are collected timespan of 1st July, 2018, to 30th June, 2021 which is in total three years of dataset.

3.1.1 *Yahoo Finance Data*

Yahoo Finance provides historical stock price data. The collection of price data for the specified stock markets is automated with the use of Python and retrieved from yahoo finance APIs as comma seperated file format for the requested timespan. The downloaded data files contain seven features, i.e., Date, Open, Close, Low, High, Volume, and Adjusted Close, which show the stock traded to date, stock lowest trading price, stock maximum trading price, stock open price, stock closing price, number of shares traded, and closing price of a stock when dividends are respectively paid to investors on a specific date.

3.1.2 *Social media Data*

For social media data source we chose Twitter because of its correctness and specifics (Tayal and Komaragiri 2009). API to scrape data from Twitter is developed using Python to download required tweets. Our Twitter application accepts console parameters the start of timespan, the end of timespan, results file location, and a search string that contains cashtag based of the ticker symbol of the specified stock market followed with a $ symbol, such as "$HPQ", "$NYSE", and so on. Cashtags are used as a search term to obtain tweets of stock market since they have been proved to be useful for financial data analysis and providing insights into stocks and companies (Hentschel and Alonso 2014). Tweets from the chosen financial markets are downloaded as json file format. The downloaded raw tweets file provides a complete tweet object with a lot of information about the tweet.

3.2 *Preprocessing*

3.2.1 *Yahoo finance data*

From the stock price data the adjusted close characteristic has been deleted because it plays no function in this stock prediction model.

3.2.2 *Social media data*

The downloaded tweets are in their raw state and must be pre-processed before machine learning techniques can be used. The following steps are conducted to convert the tweets into such a form that machine learning algorithms can understand.

1. Convert tweets text to word tokens – Sequence of letters and digits.
2. Remove author tags (@) and cashtags – these tags provide no information that machine learning algorithms can use to find emotions.
3. Remove URLs
4. Remove stop words.
5. Perform words stemming
6. Remove the re-tweet or cc
7. Remove HTML quotation marks
8. Remove punctuation and words containing just numbers
9. Remove duplicate tweets

Stop words are the terms (is, are, the, and, an) that frequently appear in tweets yet provide little (or no) meaningful information hence it is best practice to remove those. Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally written word form.

3.3 *Sentiment analysis*

From Stanford NLP the Stanford emotional analysis package is used to do sentiment analysis on the processed tweets (Socher et al. 2013). The majority of systems performing sentiment analysis work by watching words one at a time, assigning negative, positive score, and then adding these scores together. While using previous approach it disregards word order, and which causes loss of significant information. Meanwhile, the Stanford NLP technique uses a deep learning method to represent complete sentences based on sentence structure.

This technique makes use of the Sentiment Treebank, that has sentiment labels for 215,154 phrases in parse trees composed from 11,855 sentences that introduces novel sentiment compositional challenges. To tackle this, the recursive neural tensor network was proposed. On a variety of measures, the model beats all previous techniques after being trained on the new treebank. The sentiment label prediction accuracy for every phrase has achieved 80.7%, a 9.7% enhancement over the group of features baselines. Only this model is the one that correctly captures the influence of conjunctions and negations, as well as their span on different levels of tree, for the phrases with negative and positive sentiment.

As per stated in the Stanford NLP method, neutral tweets get sentiment value of three, more positive tweets get sentiment value of two, more positive tweets get a sentiment value of 4, and more negative tweets get sentiment value of 0. Above detail is represented descriptively as follows.

$$Sentiment\ score = \begin{cases} 0 & if\ tweet\ or\ news\ is\ more\ negative \\ 1 & if\ tweet\ or\ news\ is\ negative \\ 2 & if\ tweet\ or\ news\ is\ neutral \\ 3 & if\ tweet\ or\ news\ is\ positive \\ 4 & if\ tweet\ or\ news\ is\ more\ positive \end{cases} \quad (1)$$

The overall sentiment of individual tweets posted on a given day is the aggregated sentiment of those tweets posted

Date	Open	High	Low	Close	Volume	Trend	Twitter Sentiments	Future Trend
2018-07-02	138.279	140.220	38.199	139.860	3405400	Positive	574	Positive
2018-07-03	140.649	0.940	139.369	139.570	1963200	Negative	720	Negative
2018-07-05	440.479	141.429	139.929	141.429	3744700	Positive	906	Negative
2018-07-06	141.529	142.940	141.169	142.479	2849000	Positive	837	Negative
2018-07-09	142.589	44.720	12.470	.389	3904700	Positive	714	Negative
2018-07-10	44.509	45.589	44.259	44.710	3777000	Positive	805	Negative
2018-07-11	144.0	46.190	144.0	.940	3526600	Positive	726	Negative
2018-07-12	45.850	146. 830	45.740	.449	3119500	Positive	913	Positive
2018-07-13	46.449	146. .979	5.800	145.899	3062600	Negative	859	Positive
2018-07-16	145.669	145.789	144.210	145.460	3468800	Negative	605	Positive
2018-07-17	144.75	145.0	143.339	143.490	5096700	Negative	1599	Negative

Table 2: View of the final dataset after pre-processing and sentiment analysis

No.	Algorithm	Abbreviation	Optimal parameter values
1	Gaussian Naïve Bayes	GNB	NA
2	Multinomial Naïve Bayes	MNB	alpha:0.2
3	Support Vector Machine	SVM	kemel:rbf, C: 0.5
4	Logistic Regression	LR	NA
5	Multilayer Perceptron	MLP	alpha: 0.0001, activation: tanh, solver: adam, learning_ rate: constant, hidden_layer_ sizes:(5)
6	K NearestNeighbor	KNN	n_neighbors:3
7	Classification and Regression Tree	CART	max_features:log2,min_samples_split:13,random_state:123,min_samples_leaf:1
8	Linear Discriminant Analysis	LDA	shrinkage:None,solver:lsqr
9	AdaBoost	AB	n_estimators:100,learning_rate:0.1
10	Gradient Boosting Classifier	GBM	n_estimators:250
11	Random Forest Classifier	RF	n_jobs:1,min_samples_leaf:1,n_estimators:20,random_state:123, criterion:gini, min_samples_split 5
12	Extra Tree	ET	_jobs: 1,min_samples_leaf:1,n_estimators:20,random_state:123, criterion: entropy,min_samples_split:6

Table 3: Machine learning algorithms used with optimal parameter values

on that day. If the total sentiment count for a given day is larger, the sentiment positivity on that day is also larger. Sentiment characteristics are generated in processed data files of social media as a consequence of the sentiment analysis stage.

3.4 Feature extraction

3.4.1 Yahoo finance data

Stock trend prediction is based on two characteristics: Trend and Future Trend, which are determined from current features available from stock price data file. These characteristics have nominal values that might be neutral, positive, or negative. The Trend feature's value for a a specific date may be calculated through the subtraction of the opening stocks from the closing stocks. The following equation describes the criterion for picking these

$$Trend_d = \begin{cases} Positive & if\ P_c - P_o > 0 \\ Neutral & if\ P_c - P_o = 0 \\ Negative & if\ P_c - P_o < 0 \end{cases} \quad (2)$$

In equation 2 $Trend_d$ denotes the trend, and P_c denotes the close price of stock, and P_o denotes the open price of stock on a given day.

The property that will be forecasted is the Future Trend feature. It is the difference between a stock's current closing price and the stock's closing price after n days. If the difference is positive, the trend after n days will also be positive. In the meantime, if the difference is 0, then the future trend of the stock would be neutral, whereas in case the difference is negative, then the future trend of the stock will move downward after n days.

The following equation can predict the future trend after n days.

$$Future\ Trend_n = \begin{cases} Positive & if\ P_{tc} - P_{nc} > 0 \\ Neutral & if\ P_{tc} - P_{nc} = 0 \\ Negative & if\ P_{tc} - P_{nc} < 0 \end{cases} \quad (3)$$

Where P_{tc} is today's closing price of the stock while P_{nc} is the closing price of stocks after n days.

For this study we've set n to 15, which means we'll be able to predict the stock's future trend for the next 15 days. Thus, we will investigate the influence of oil or gold trends and social media for forecasts for 15 days in the future.

3.4.2 Social media data

Sentiments over social media are considered to be a feature derived from social media files. For one date, these sentiments can be calculated through summation of all sentiments of tweets for that particular date.

3.4.3 Oil/gold data

From oil and gold price data, two features have been derived, i.e., *oil future trend* and *gold future trend*. These can be calculated in the same way as stated previously in yahoo finance data.

3.5 Final datasets

We have discussed the procedure of developing final datasets in this subsection which would be used for predicting stock.

3.5.1 Predicting stock market based on social media

By including the sentiment feature in the stock price data file, the final dataset for this subsystem is generated. Table 2 depicts a tabular representation of the dataset.

3.5.2 Predicting stock market based on oil trends

By including an oil future trend feature into Yahoo financial data files, the final dataset for this subsystem is generated. The resultant dataset is present in Table 2.

3.5.3 Stock market prediction using gold trends

By including a gold future trend feature into Yahoo financial data files, the final dataset for this subsystem is generated. The resultant dataset is presented in Table 2.

3.5.4 Stock market prediction using social media and oil/gold trends

Likewise, the final dataset for this subsystem is generated by including additional characteristics such as oil future trends, gold future trends, and social emotions into Yahoo financial data files. This dataset is equivalent to as presented in Table 2.

3.5.5 Impact of Covid19 on Stock market

For identifying the influence of Covid19 on the selected firms, we are using daily closing price from Yahoo Finance dataset. Now for performing comparative analysis between selected stock markets and Covid19, we have collected Covid19 infection rate from World Health Organization (WHO) Covid19 explorer web application.

3.6 Application of machine learning classifiers

Twelve machine learning classifiers were chosen for this investigation, and their prediction performance was compared. These classifiers will be trained first and then evaluated on final datasets to identify future stock market patterns. The final datasets for prediction system are divided into 70% (528 samples) training data and 30% (226 samples) testing data before machine learning techniques are applied. We built prediction models with Scikit-learn (Pedregosa et al. 2011), which is a machine learning toolkit developed in Python, to train and evaluate the algorithms. Table 3 lists the stock prediction algorithms that were used in this study.

3.6.1 *Standardizing the final datasets*

Prior to the application of the classifiers, the final datasets are standardised using Scikit-StandardScalar learns class. Since a few algorithms, such as LDA, and GNB, these algorithms consider input data to be in the Gaussian distribution form, standardisation is beneficial for converting characteristics with Gaussian distribution. After transformation, the attributes get a mean value of zero and a standard deviation value of one. Training dataset might be leaked to the testing dataset during dataset standardisation. A solid test harness with some strict distinction of teaching and testing is essential to overcome this problem. This involves data preprocessing, which allows the algorithm to learn the entire training dataset. Scikit-Pipeline learn's function is used for avoiding this. Pipelines turn data into a linear series that may be linked together, which result in an evaluable modelling process. It guarantees that standardisation is limited to every fold in the process of cross-validation (CV), therefore eliminating data leaking in the test harness. The main goal is to ensure that the entire pipelining method remains limited towards the training dataset that is currently accessible to testing.

3.6.2 *Classifiers performance evalutaion*

Various assessment criteria are used to assess the performance accuracy for all the chosen classifiers. Because of multi-class classification problem having non-uniform distribution of prediction classes, we used accuracy as the major classification metric, as well as F-measure, recall, and precision as within-class classification metrics. Accuracy can be defined as the classification statistic used to evaluate classifiers that may be represented.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (4)$$

If the prediction class distribution isn't uniform, then the accuracy won't be a suitable criterion to evaluate classifier performance. Hence, confusion matrix is utilised for the classification performance evaluation, and its accuracy, recall, and F-measure are determined.

Precision is a measurement of a model's ability to correctly identify samples, and it may be measured as below.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

The algorithm's true positive rate is represented by TP, and its false positive rate is represented by FP.

The recall of a model is measured by its ability to categorise as much potential data as possible. Below is the equation to calculate recall.

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

In equation 6 FN is the false-negative classification of the algorithm, and F-measure represents the precision, as well as recall. F-measure calculation is shown as below.

$$F - \text{MEASURE} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

3.6.3 *Model validation for the proposed solution*

Validation approaches for prediction models involve replacement, holdout, and CVs like leave-more-out, leave-one-out and k-fold of CVs (Chou and Lin 2012; Kuhn and Johnson 2013).

With the help of tenfold CV, several tuning parameters utilised by algorithms to optimise performance of classification are investigated in terms of variance. When comparing algorithm performance (Thu et al. 2011; Chou and Lin 2012), the number of folds used is 10, as recommended by Kohavi (1995). As a result, the tenfold CV technique is used to create every prediction model over all conceivable parameter combinations in this study. The data from the training set is randomly split into 10 subgroups in this technique. For finding the prediction performance for the fitted model, a hold out set is utilised, whereas the fresh training set of nine subsets is utilised to develop all prediction models. Previous method is performed for 10 iterations on different datasets of training till all instances of the training dataset is utilised to test only once. The average of the accuracies gathered from ten iteration is then used to evaluate CV's overall accuracy. This method is employed in order to choose the optimal parameters for prediction models and to minimise over-fitting. The testing dataset isn't utilised in model development, yet is used to test the prediction performance of the relevant model after it has been completed. Prediction models are created by combining the optimal parameters and training datasets and later, the final models are implemented on testing datasets.

3.6.4 *Parameter optimization*

For avoiding under or over-fitting, machine learning classifiers employ multiple tuning parameters as per need. The fit function in the GridSearchCV class from Scikit-learn library generates a grid of tuned classification algorithms and creates a stable environment for training and tweaking of each machine learning algorithm. Once we have the ideal parameter values discovered, the whole training dataset is utilised to create the final model. On the basis of the training dataset, tenfold CV is used to pick optimal values for the tuning parameters, while the testing dataset is completely wiped during the CV process. The tuned parameter values are deemed optimum during the CV process in order to get the highest accuracy of classification. The optimum parameter values used in the associated Scikit-learn classification class are the parameters and values in the last column of Table 3.

4. PROPOSED SOLUTION

For accurate predictions, the suggested system takes into account all aspects of the data as well as the system itself. As a result, the suggested stock prediction system is separated into nine subsystems. These subsystems are described in this section.

4.1 Stock prediction system

4.1.1 Using social media

On the final datasets, extensive testing is done using ML models to predict future stock market patterns in the selected stock markets for the next 15 days. The algorithms chosen are trained first and later tested with the help of a tenfold CV, which is used for ensuring that every instance is utilised equally to train and test while decreasing variation. Algorithm parameter tweaking is also conducted to guarantee the choosing of the best parameter values for maximum accuracy of the prediction.

Feature	Polarize	Max-min scale	Fluctuation percentage
Price change			
Price change percentage			
Volume		v	
Amount		v	
SMA 10		V	v
MACD	v		
MACD SIGNAL	v		
MACD HIST	v		
CCI 24	v		
MTM 10	v		v
ROC 10	v		v
RSI 5			v
WNR 9	v	V	
SLOWK			v
SLOWD		v	v
ADOSC	V	v	
AR 26		v	
BR 26		v	
VR 26		v	v
BIAS 20	v		

Table 4: Selection for feature extension method

4.2 Using oil/gold trends

The chosen classifiers are applied to the final datasets of the specified stock markets over the next 15 days in order to forecast future trends on the final datasets. First, these classifiers are trained, followed by their testing on the final datasets using tenfold CV. For achieving optimum forecast accuracy, the parameters of the oil/gold price trend based prediction system are tuned as well.

Fig. 3. Comparison of mean testing accuracy between SelectKBest and PCA techniques for various values of K

4.3 *Feature extension*

For improving the performance of this prediction system, a feature extension has been added on top of the standard machine learning algorithms. We employ the most common technical indices derived from comparable research for feature extension. Max–min scaling, polarising, and computing fluctuation percentage are the three feature extension methods used in current research.

This process is relevant only for the meaningful extension methods to technical indices since not every technical index is appropriate for all three feature extension methods. We have selected relevant extension techniques while considering the ways in which the indices are computed. Table 4 depicts the technical indices and the related feature extension techniques.

4.4 *Feature selection/Dimensionality reduction*

In this study, dimensionality reduction or feature selection is used for developing an efficient prediction model. Scikit-feature learn's selection and deconstruction modules are used for feature selection for improving the performance and accuracy of algorithms on big datasets.

For feature selection, the modules allow a wide range of classes. In this study, for feature selection at various K values, we use PCA and SelectKBest Chi2 method Here, K represents the count of components that should be kept in PCA and the count of features that have to be selected in SelectKBest. The mean testing accuracy for the two approaches is depicted in Figure 3 using various values of K.

4.4.1 *SelectKBest*

SelectKBest is a univariate feature selection strategy that determines the top K features while ignoring the others. The Chi-squared statistics between each positive attribute and the classes are calculated via Chi2. The stats quantify interdependence of the characteristics. Those characteristics that are believed to be class-independent and irrelevant for classification are eliminated.

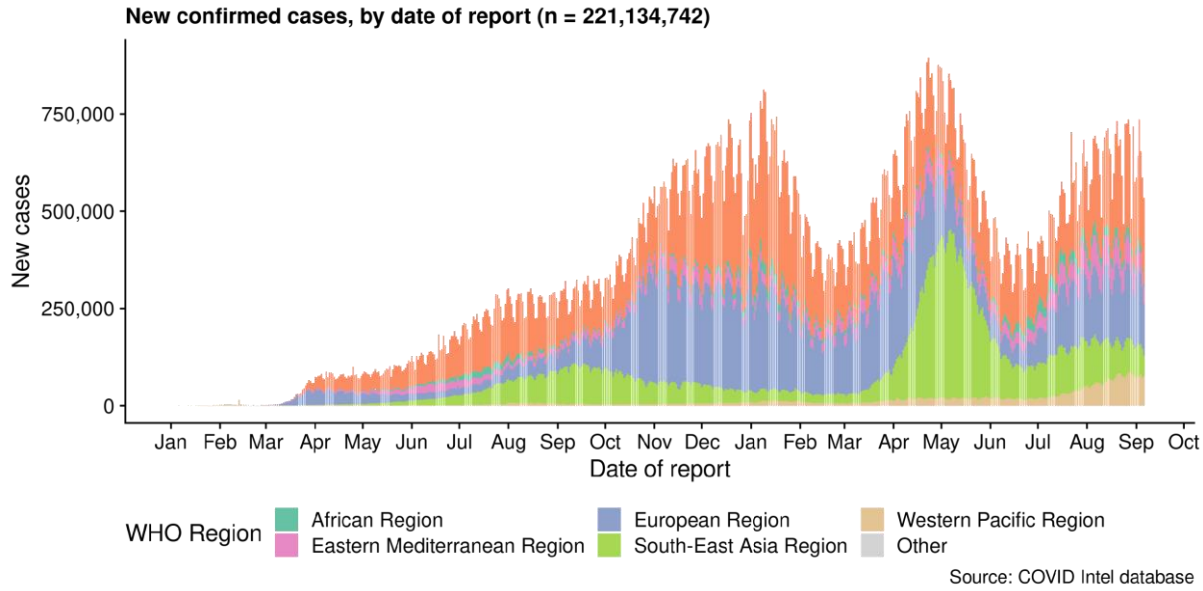

Fig. 4. Above chart shows new Covid19 cases World-wide on y-axis along with month on x-axis

Covid19 Waves	Duration
Before Covid19	1st July, 2018 to 20th March, 2020
1st wave	21st March, 2020 to 10th September, 2020
2nd wave	11th September, 2020 to 20th February, 2021
3rd wave	21st February, 2021 to 20th June 2021

Table 5: Covid19 timespan distribution

4.4.2 *PCA*

PCA uses singular value decomposition on the input data to project data into a smaller dimensional space, which is useful for removing redundant features and components from a dataset. We look for components that explain the most variation in PCA and retain the most information.

4.5 *Spam tweet reduction*

Spam reduction is performed in this study endeavour for achieving quality prediction outcomes. We performed training and testing on the MNB classifier for categorising raw tweets dataset into ham and spam tweets while noting the F-measure, precision, recall, and accuracy for reduction of such tweets. Due to its text classification accuracy, the MNB classifier is utilised for categorizing tweets as ham and spam (Afzal and Mehmood 2016). The trained algorithm, which has an 81.74% testing accuracy, is then utilised to categorise raw tweets into ham and spam messages. For each stock market, the percentage of ham and spam tweets is determined. After categorization, all tweets categorized as spam are eliminated, and to create final datasets, simple pre-processing as well as sentiment analysis of ham tweets are performed. The resulting datasets are then used for training and testing machine learning algorithms and then, their performance is recorded for comparison.

4.6 *Identifying classifier with consistent results*

It is critical to find a classifier that consistently produces good results under various factors, such as stock market prediction utilising news and social media data, spam tweets reduction, and dimensionality reduction. As a result, we examine all classifiers' prediction accuracies in various circumstances and find a classifier which consistently delivers excellent results in all scenarios in stock market prediction.

4.7 *Identifying stock markets that are highly impacted by social media*

In this study, the sentiment-based method is used to identify stock markets among the chosen stocks that are heavily influenced by social media. The Stanford NLP technique for sentiment analysis is used to discover sentiments in tweets. Stocks with higher levels of positive sentiment are thought to be impacted more by social media.

4.8 *Identifying stock markets that are more affected by Covid19*

To identify the most affected stock market by Covid19, we first have to identify the impact for individual stock market. Therefore this analysis is divided into two segments, first is to identify the impact of Covid19 on individual stock markets and then comparing the impact to identify the most impacted stock market. The first segment is further divided into partitions that identify impact of different waves of Covid19. From Figure 4 we can divide the Covid19 waves as per the Table 5. Starting time of the first wave is considered when number of cases world-wide increased from 10K and the end of every wave is considered when the number of cases decreases to the point after which they start increasing again. Even though the Covid19 still haven't ended we set the ending timespan of 3rd wave to 20th of June, 2021 when the number of infection cases are at lowest before increasing again. Before Covid19 timespan is considered when there were less than 10K infection cases world-wide. From three waves categorised the max number of infection cases are observed in 3rd wave, 2nd wave and 1st wave respectively.
In first segment of this analysis, we used descriptive analysis. To perform descriptive analysis we chose Welch test for identification of deviation of selected stock market returns using their mean for different waves of Covid19. For the second segment of this analysis, we used variance of mean stock closing prices between two timespans.

4.9 *Application of deep learning in stock prediction*

For stock market data prediction, we use neural networks, particularly MLP, because they perform better. We also incorporate deep learning into MLP by introducing hidden layers. For determining the ideal count for hidden layers in this study, we raised hidden layers individually and recorded the prediction accuracies. Authors (Wassan et. al, 2021 and Dogra et.al,) presented sentiment analysis for amazon products, and banking. Further (Chohan, 2021) presented related to the sentiment for social networks as well.

4.10 *Hybrid algorithm*

To achieve greater performance, several ensemble techniques are utilised to combine predictions from separate classifiers. Individual predictions from the top classifiers are merged with the help of the voting ensemble approach for getting the maximum prediction accuracy. The voting ensemble technique was chosen since there are small variations within it as compared to other ensemble methods, and while it maintains good performance. (Kim et al. 2003).

Using the majority voting ensemble technique, we integrate the predictions of ET, RF, and GBM classifiers. These are the classifiers that outperformed the rest of the pack in our prediction models. For implementing majority voting, the Python Voting Classifier class is utilised. Majority class values forecasted by every different classifier are indicated by projected class value for a particular sample in this approach is the class value that indicates. For instance, in our issue, if the forecast for a particular sample is GBM →Positive, ET →Negative, and RF →Negative then on the basis of the majority class value, the Voting Classifier will classify the sample as "Negative".

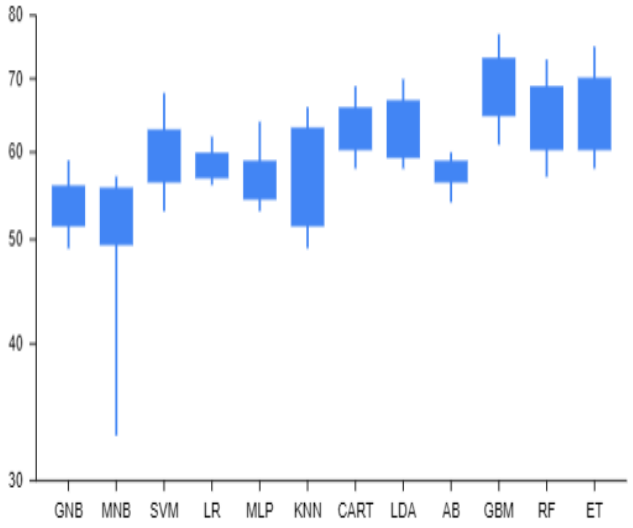

Fig. 5. Accuracy comparison between selected algorithms using box plot distribution (IBM dataset)

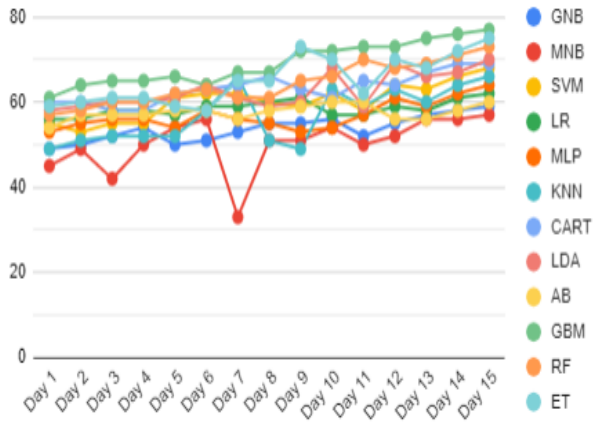

Fig. 6. Prediction accuracies of algorithms using social media sentiments for 15 days (IBM dataset)

Classes	**Metrics**						**Algorithms**					
		GNB	MNB	SVM	LR	MLP	KNN	CART	LDA	AB	GBM	RF	ET
Negative	Precision (%)	50	100	78	62	53	58	62	66	84	74	81	56
	Recall (%)	52	4	28	44	28	64	55	48	15	75	65	60
	F-measure (%)	48	3	31	48	40	63	53	60	28	76	74	57
Neutral	Precision (%)	0	Na	Na	Na	Na	0	Na	Na	Na	Na	Na	Na
	Recall (%)	0	Na	Na	Na	Na	0	Na	Na	Na	Na	Na	Na
	F-measure (%)	0	Na	Na	Na	Na	0	Na	Na	Na	Na	Na	Na
Positive	Precision (%)	72	68	76	65	62	78	74	73	67	84	79	76
	Recall (%)	52	100	92	83	86	71	75	87	97	83	88	78
	F-measure (%)	64	72	81	77	75	73	76	81	77	85	84	73

Table 6: F-measure, recall, and precision accuracies of classification using all the algorithms (IBM dataset)

5. RESULTS OF SOLUTION AND DISCUSSION

5.1 Stock prediction system

The following sub-section is used for discussing the outcomes of the suggested algorithms on the IBM stock market dataset.

5.1.1 Using social media

a) Results of tenfold CV

Figure 5 depicts a comparison of performance for the various classification methods using the tenfold CV algorithm over the IBM dataset.

The box plot depicts every algorithm's performance on the tenfold CV regarding the average accuracy on the training dataset. The borders of the box identify the 1st and 3rd quartiles. The lines are stretched to the most extreme data points both on the higher value and the lower value as well. Figure 5 depicts the accuracy of twelve algorithms as a whole prior to standardisation and parameter tweaking. Across the 12 algorithms, average accuracy is in the range of 48% to 65%.

GBM classifier shows the best accuracy which is 65.17%, trailed by LDA and RF, which have average accuracies of 62.73% and 61.27%, respectively. CART has a lower average accuracy of 60.27%, trailed by ET, SVM, LR, and MLP, which have average accuracies of 60.26%, 58.2%, 56.03% and 55.37%, respectively. AB, MNB, and GNB perform the worst, with MNB having the lowest average accuracy of 48.4%.

b) Results of the independent testing dataset

The accuracy of testing dataset ranges from 49.00% to 77.57%. Figure 6 depicts the performance of several algorithms over the next 15 days. On day 15, the GBM classifier clearly has the greatest accuracy of 77.57%, trailed by the ET classifier's accuracy of 75.86%. MNB has the lowest accuracy of 33.00% on the seventh day after the deal was made.

The results of using social media emotions to predict stock market future trends show that GBM has the highest prediction accuracy on day 15, followed by ET on the same day. On the 15th day, GBM achieves an accuracy of 74.20% without employing the social media sentiment component,

resulting in a 3.96% drop in prediction accuracy. The graph shows that while the impact of emotions on prediction accuracy fluctuated from day 1 to day 9, it became more consistent from day 10 to day 15.

The findings of the training and testing datasets show that GBM gives high performance on both, whereas RF, ET gives high performance on the testing dataset. The improved performance of RF on the testing dataset might be caused due to dataset standardisation and parameter tweaking.

For each class in two prediction models, three accuracy metrics, namely F-measure, recall, and precision, are collected in order to assess the accuracy of the classifiers in discriminating among the three future trend classes (negative, positive, and neutral). We demonstrate F-measure, recall, and precision for prediction models based only on social media. The values for all indicators for the sentiment-based prediction are shown in Table 6.

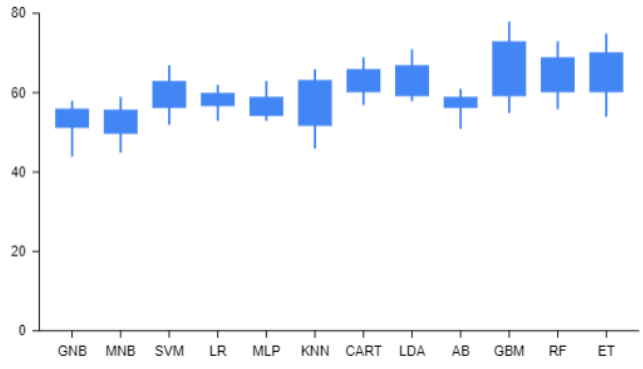

Fig. 7. Accuracy comparison between selected algorithms using box plot distribution (IBM dataset)

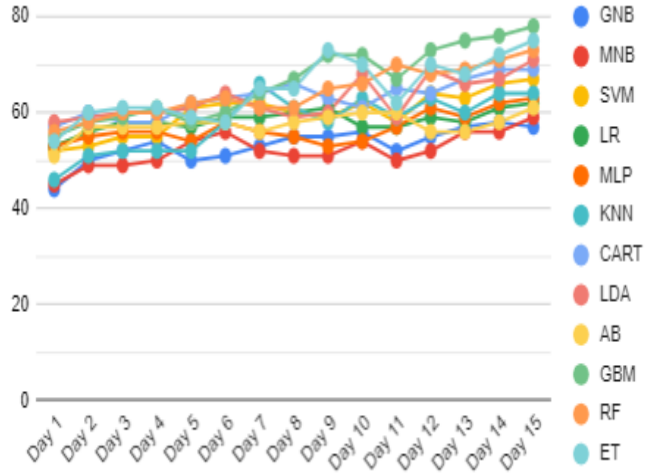

Fig. 8. Prediction accuracies of algorithms using social media sentiments for 15 days (IBM dataset)

For positive future trend class, GBM has the best accuracy (84.00%), whereas MNB has the highest recall (100.0%). GBM also has the best overall performance (F-measure, recall, and precision are respectively 85.00%, 83.00, and 84.00%). Except for the MLP, all algorithms have accuracy values of more than 62.00% and recall values greater than 71.00 % except for GNB (52.00%). MNB achieves maximum precision (100.0%) for the negative future trend class, however, it has very weak recall (4.00%) and an F-measure of (3.00%). As a result, the algorithms have a poor recall (4.00–72.00 %) and F-measure (3.00–73.00 %) for the negative future trend class. Moreover, GBM performs rather well in the classification of the negative future trend class Precision, recall, and F-measure were all 74%, 75%, and 76%, respectively. According to Table 6, the major issue is the algorithms' unfavourable classification rate for the neutral future trend class. GNB and KNN show 0% precision, recall, and F-measure, whilst other remaining algorithms show no precision, recall, or F-measure. Here, the confusion matrix clearly shows that the most significant mistake cause is the reclassification of the neutral future trend class into positive or negative future trend classes. The explanation for this might be due to the minimal amount of samples in this class. Despite the fact that the dataset is unbalanced, certain classifiers outperformed the others significantly. It is obvious that no matter what the imbalance dataset, GBM and RF are both effectual in providing the best accuracy for the negative and positive future trend classes.

5.1.2 Using oil/gold trends

a) Results of tenfold CV

Figure 7 depicts a comparison of performance for the various classification methods using the tenfold CV algorithm over the IBM dataset. Across the 12 algorithms, average accuracy ranges from 49% to 63%. The LDA classifier has the greatest average accuracy of 62.80%, followed by the GBM and RF classifiers with 62.35% and 61.23%, respectively. CART doesn't perform as expected, and gives an average accuracy of 60.07%, followed by ET, SVM, and LR classifiers, which have average accuracies of 59.87%, 57.93%, and 55.83%, respectively. GNB, KNN, AB, and MLP classifiers perform poorly, with GNB having the lowest average accuracy of 49.31%.

b) Results of the independent testing dataset

The accuracy of stock market prediction using oil/gold price trends ranges from 40.70% to 74.20% across the tested dataset. Figure 8 depicts the performance of several algorithms over the course of 15 days. On day 15, the GBM classifier clearly obtains the greatest accuracy of 74.20%, trailed by RF with an accuracy of 70.40%. On the 1st day, GNB has the lowest accuracy of 40.70%. The data also reveals that the highest level of accuracy is achieved on the 15th day, trailed by the 14th day. On the 13th day, the RF achieves an accuracy of 66.95% without utilizing the oil/gold trend feature, resulting in a 2.97% drop in forecast accuracy without using the oil/gold price trends.

The accuracy, F-measure, and recall for the oil or gold trend-based prediction model are shown in Table 7. The highest accuracy of 76.00% was achieved by the RF classifier for the positive future trend class, whereas MNB and AB, (both 100%) attain the highest recall.

Still, The RF classifier's overall performance is at top, with accuracy, recall, and F-measure of 76.00%, 85.00%, and 81.00%, respectively. Except for AB (58.00%) and MLP(61.00%) , the general accuracy is greater than 62.00% and the recall is greater than 70.00% except for GNB with (66.00%) and KNN with (69%).

Maximum accuracy is attained with AB (100.0%) for the negative future trend class, but it has extremely low recall (9.00%) and F-measure (21.00%). MNB has the lowest precision, recall, and F-measure of (0.0%) for negative future trend class, whereas CART and KNN has the highest recall (65.00%). As a result, the algorithms have a poor recall in range of (0.00–62.00 %) and F-measure in range of (0.00–64.00 %) for the negative future trend class. The RF classifier performs reasonably well in the negative future trend class, with precision, recall, and F-measure of 77.00%, 63.00%, and 70.00%, respectively.

Classes	Metrics	Algorithms											
		GNB	MNB	SVM	LR	MLP	KNN	CART	LDA	AB	GBM	RF	ET
Negative	Precision (%)	60	0	92	89	79	65	60	79	100	74	77	65
	Recall (%)	57	0	22	25	24	65	65	41	9	62	63	55
	F-measure (%)	58	0	37	36	33	65	66	58	21	69	70	58
Neutral	Precision (%)	0	0	0	0	0	0	0	0	0	0	0	0
	Recall (%)	0	0	0	0	0	0	0	0	0	0	0	0
	F-measure (%)	0	0	0	0	0	0	0	0	0	0	0	0
Positive	Precision (%)	64	57	63	62	61	71	70	66	58	74	76	69
	Recall (%)	66	100	99	98	96	69	68	91	100	82	85	76
	F-measure (%)	65	73	77	76	76	70	67	78	74	78	81	71

Table 7: F-measure, recall, and precision accuracies of classification using all the algorithms (IBM dataset) (oil/gold price prediction)

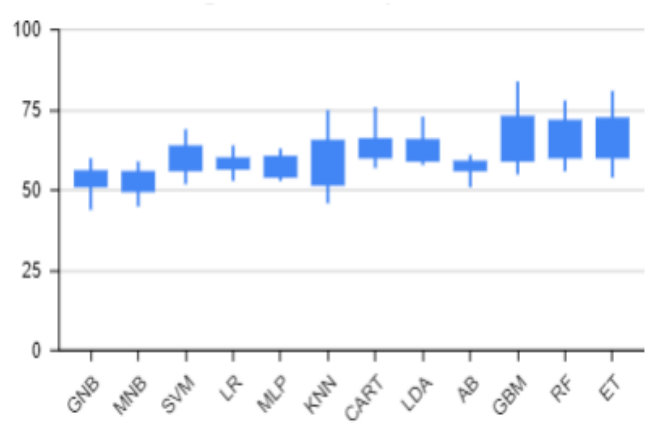

Fig. 9. Accuracy comparison between selected algorithms using box plot distribution (IBM dataset)

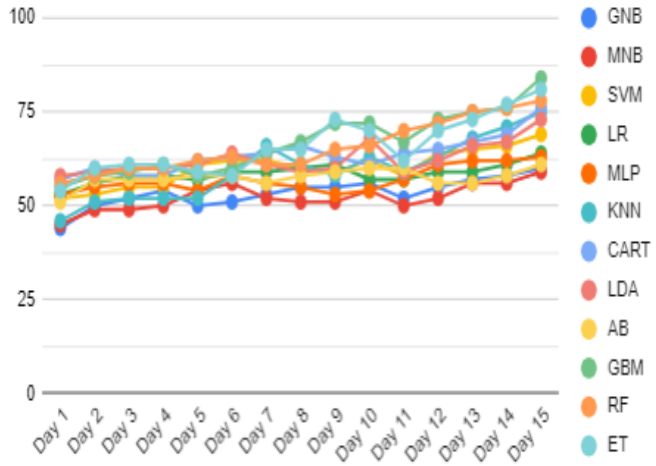

Fig. 10. Prediction accuracies of algorithms using social media sentiments for 15 days (IBM dataset)

Table 7 shows that the algorithms' poor classification performance for the neutral future trend class is also a major challenge in the oil/gold based prediction system. All classifiers in this category have no F-measure, recall, or accuracy. The most common mistake, according to the confusion matrix, is the reclassification of neutral future trend classes into negative or positive future trend classes. The explanation for this might be due to the minimal amount of samples in this class. To a large degree, certain classifiers outperform the other classifiers in this classification task. No matter the imbalance dataset, both GBM and RF are effective in displaying the maximum accuracies for the negative and positive future trend classes.

According to Tables 6 and 7, the RF and GBM classifier performed well in terms of F-measure, recall, and precision on the testing datasets, therefore we can conclude that they might be suggested for stock market prediction.

5.1.3 Using social media sentiments and oil/gold trends

a) Results of tenfold CV

Figure 9 depicts comparison of performance with the various classification methods on the tenfold CV over the IBM dataset. Across the twelve algorithms, average accuracy ranges from 49.47% to 62.73%.

The GBM classifier has the highest average accuracy of 62.73%, trailed by the RF, LDA, and ET classifiers, which have average accuracies of 62.53%, 62.47%, and 60.93%, respectively. CART performs poorly, with an average accuracy of 60.53%, followed by SVM, KNN, and LR classifiers, which have average accuracies of 58.2%, 56.17%, and 56.03%, respectively. The GNB, AB, and MNB classifiers perform the worst, with GNB having the lowest average accuracy of 49.47%.

b) Results of the independent testing dataset

The accuracy for stock prediction utilising social media emotions and oil/gold price trends ranges from 40.37% to 80.20% across the testing dataset. Figure 10 depicts the performance of several classifiers over a 15-day period on the IBM testing dataset. On the 15th day, the GBM classifier obtains the greatest accuracy of 80.20%, followed by ET at 76.11%. On the 1st day, GNB has the lowest accuracy of 40.37%. The results also reveal that the highest level of accuracy is achieved on the 15th day.

After the 9th day, the unstable performance of some algorithm becomes stable, and the average prediction accuracies of most algorithms rise. The GNB's overall accuracy is low, with a substantial decline in accuracy (40.37%) recorded on the 1st day.

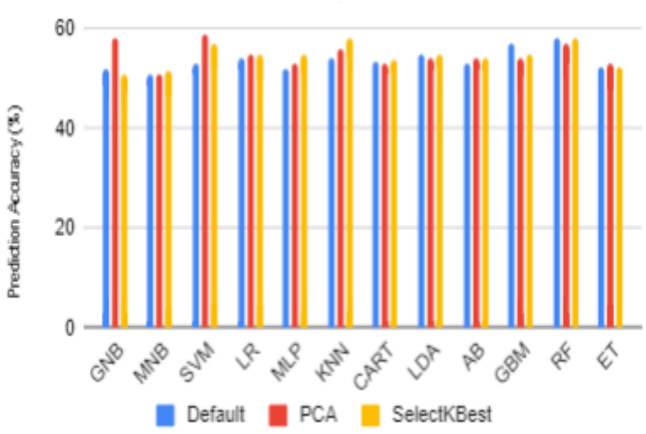

Fig. 11. Comparison of accuracies before and after the applicarion of SelectKBest and PCA (IBM dataset)

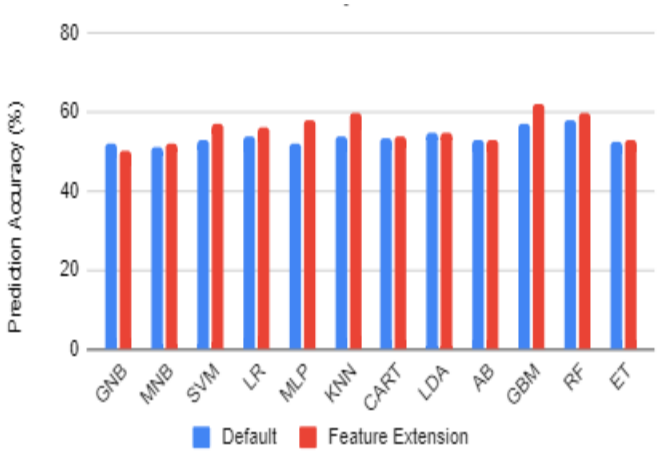

Fig. 12. Accuracies comparison before and after the feature extension on IBM stock dataset

The findings of the three prediction models mentioned in previous subsections reveal that MNB and GNB perform the worst on the training and testing datasets. GNB's poor accuracy might be related to its assumption of that the input data is in Gaussian distribution form since it works best with data having Gaussian distribution. Similarly, MNB outperforms with unique features, however our dataset contains both nominal and discrete characteristics, reducing MNB performance.

Comparison of results with oil/gold and social media datasets

To compare the effect of both of our attributes we can compare the highest accuracies achieved by both results. With social media sentiments only we achieved a maximum accuracy of 77.57% and with oil/gold trends dataset we achieved a maximum accuracy of 74.2%. It can be clearly seen that both give a good accuracy score even if used individually. When we combined both the datasets of oil/gold trends with social media sentiments datasets we get a very good accuracy of 80.20%. From these stats it can clearly be seen that if we combine both the datasets they increase the accuracy of stock market prediction.

5.2 Feature extension

We can deduce from the accuracy findings shown in Figure 12 that the feature extension technique enhances the accuracies of the majority of the classifiers (SVM, MLP, LR, GBM, KNN, RF, CART, MNB, and ET), whereas the GNB shows a decline in accuracy. There are no differences in accuracy between AB and LDA.

5.3 Feature selection/Dimensionality reduction

According to the results, the accuracies of the majority of the classifiers (SVM, GNB, LR, KNN, MLP, MNB, RF, AB, and ET) enhance by a single or the two feature selection approaches, whereas GBM shows an accuracy loss, LDA and CART show little to no change in accuracy after dimensionality reduction or feature selection. As shown in Fig. 3, the components or number of features on which both methods attain maximum accuracy is nine out of eleven in

this case. SelectKBest, as shown in Fig. 10, is the best approach for enhancing prediction accuracy when compared to PCA. The findings revealed that there could be an increase in the performance of the classifiers or it is possible to gain the same performance with the help of a feature subset.

5.4 *Spam reduction*

The percentage split of spam tweets in Fig. 12 reveals that spammers have a greater impact on the HPQ stock market than on the ORCL stock market. Approximately 14% of HPQ tweets and 6% of ORCL tweets are judged to be spam. RHT is the stock that suffers the least from spammers (1% spam tweets only). Similarly, the LSE is proven to be more affected by spammers than the rest of the stock exchanges (12.4% of spam tweets), while others aren't affected at all. The rationale for other markets might be that they are stock markets that are rarely addressed, as seen by the number of tweets shown in Table 1.

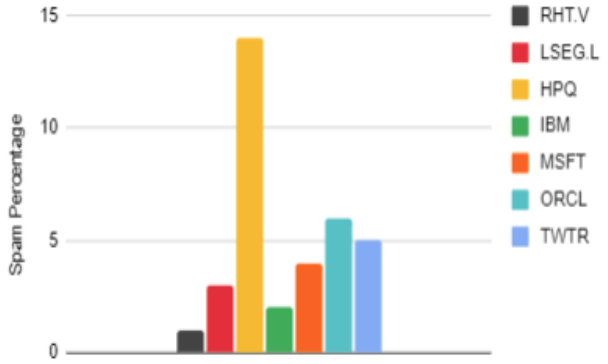

Fig. 13. Spam tweet percentage for all the chosen stock markets

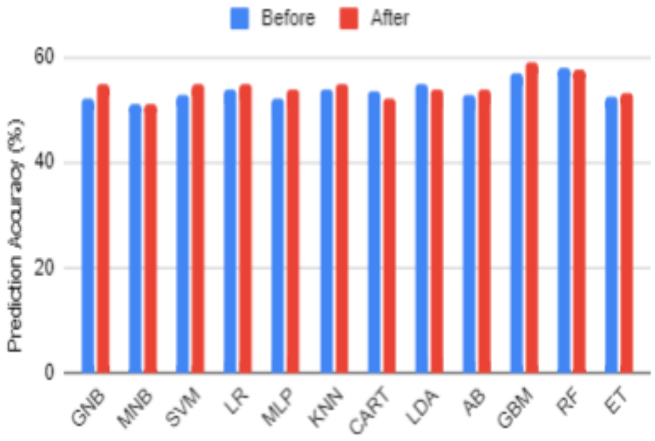

Fig. 14. Accuracy comparison of classifiers before and after applying spam reduction

Following the removal of spam tweets the prediction accuracies for majority of the classifiers (SVM, MLP, GNB, LR, GBM, ET, KNN and AB) increase, indicating their resilience. GNB shows the greatest gain of 3% in accuracy after spam removal. Likewise, several of the classifiers' accuracies (CART, LDA, and RF) are reduced, whereas MNB don't show any change at all after removal of spam. Classifier accuracy comparisons before and after spam minimization are shown in Fig. 13. From above findings it is clear that spam reduction have a positive impact on most of the classifier's accuracy.

		HPQ	**IBM**			**LSEG.L**		**MSFT**		**ORCL**		**RHT.V**	**TWTR**	
	Mean	Std. Dev	Mean	Std. Dev	Mean	Std. Dev	Mean	Std. Dev	Mean	Std. Dev	Mean	Std. Dev	Mean	Std. Dev
Before Covid19	21.11	2.36	136.46	10.26	5,655.25	1,355.42	128.35	21.55	425.00	3.55	0.54	0.40	34.64	4.69
1st wave	16.83	1.45	121.56	5.87	7,528.07	2,100.71	192.15	20.27	53.95	2.45	0.30	0.06	32.91	4.93
2nd wave	22.12	3.02	121.79	4.93	8,786.76	411.87	218.48	11.63	60.36	2.55	0.41	0.21	49.24	7.41
3rd wave	31.48	2.03	137.32	9.12	7,841.66	910.13	245.97	10.32	74.72	5.72	0.53	0.08	63.01	6.81

B. Table 8: Mean and Std. deviation for waves in Covid19

5.5 *Identifying classifier with consistent results*

Based on the findings of several subsystems described in earlier subsections, it was concluded that GBM is the most ideal classifier for producing consistent results for multiple reasons, which are as follows.

- It has the greatest prediction accuracy of (77.53%) when it comes to stock prediction with sentiment data.
- It has the best prediction accuracy of (78.2%) when utilising oil or gold price trends to predict stocks.
- Its prediction accuracy improves by 5% when feature extension methods are used.
- Its prediction accuracy increases by 2% after spam removal.
- It has the best F-measure, prediction accuracy, recall, and precision performance.
- It gives a high performance on training and testing datasets.

The highest performance of GBM might be due to the fact that our issue is a multiclass one, and GBM has been designed

for such issues. The second explanation might be because our dataset includes both numerical and categorical characteristics, and GBM can effectively work with this kind of dataset.
Because of the GBM classifier's reliable findings, the method may be suggested for the prediction of stock market trends.

5.6 Identifying stock markets that are highly impacted by social media

Findings of our research clearly demonstrate that social media has the most effect on the IBM stock market, followed by TWTR, whereas LSEG.L and HPQ stocks have the least influence, as evidenced by the spikes in Figure 15. Likewise, the NYSE is proven to be highly impacted by social media than other stock markets.

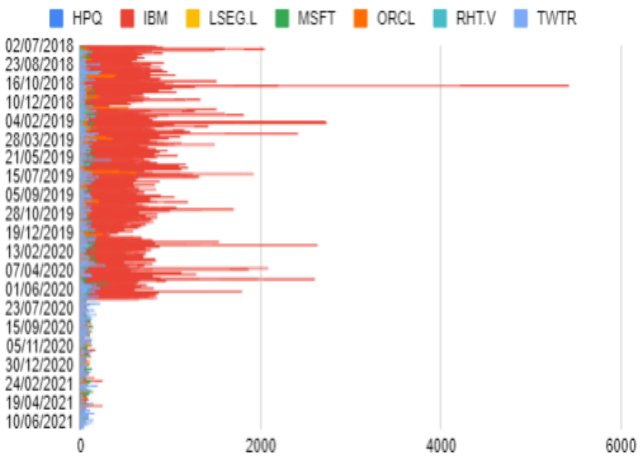

Fig. 15. Impact of social media on chosen stock markets

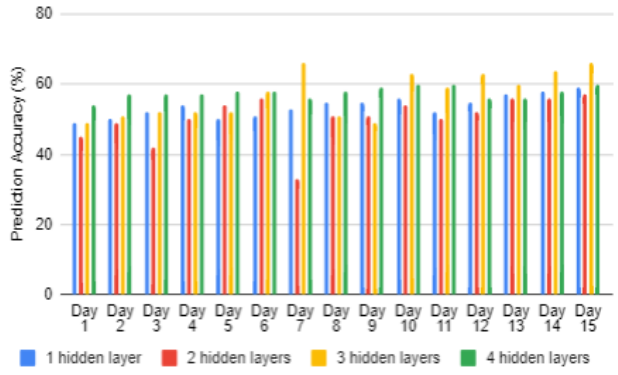

Fig. 16. Comparing accuracy of the neural network with different number of hidden layers

5.7 Identifying stock markets that are more affected by Covid19

The Covid19 epidemic has had a substantial influence on the socioeconomic status of the majority of the world's capital markets. It is undeniably a watershed moment in the operations of several industries, and in the growth paths of entire economies. It definitely had a substantial impact on stock prices in several areas.

The impact of one wave timespan is calculated by subtracting mean value of previous timespan. For 1st wave we subtract the mean value of a stock market of before Covid19 timespan. From the results of descriptive analysis in Table 8 it is clear that closing price of different stock markets falls throughout all waves of Covid19. Closing price fall is very steep on 1st wave of Covid19. For ORCL, the mean closing price was 425.00 but after 1st wave it went down to 53.95 which is an astonishing decrease of 87.31%. The second stock market that was negatively affected in 1st wave was RHT.V with a decrease of 45.05%. Other than these two stock markets, HPQ, IBM and TWTR were also negatively affected and their stock prices decreased 20.27%, 10.92% and 4.98% respectively. Remaining stock markets weren't affected negatively in 1st wave of Covid19. MSFT was the stock market with maximum positively impacted having a 49.71% increase in mean closing stock price followed by LSEG.L with a 33.12% increase.
Furthermore, the impact of 2nd wave is positive for all the stock markets and only LSEG.L stock market got a decrease in closing price in 3rd wave. Overall impact of Covid19 on all the selected stock markets can be easily observed by comparing mean values of before Covid19 and 3rd wave. ORCL and RHT.V were negatively impacted during complete Covid19 timespan with 82.42% and 2.77% decrease respectively. Whereas MSFT, TWTR were the stock markets with the most positive impact on mean value of 91.64% and 81.90% respectively.
From our stats as above we can conclude that most of the stock markets have a negative association during the first wave; however for the 2nd wave this relationship becomes positive for most of the stock markets. The reason for this can be that the second wave has a lower amount of uncertainty than the first and people and companies have evolved to the conditions. Moreover it is also observed that the companies that could do most of their work actions remotely were very positively impacted by Covid19. Opposite happened for the stock markets that were mostly relying on physical customers dealing.

5.8 Improving accuracy using deep learning

The prediction accuracy of our neural network was improved by up to 8.5% using deep learning. As demonstrated in Figure 16, raising the number of hidden layers up to 3 gives best prediction accuracy. Following the increase in the number of hidden layers, the neural network's accuracy falls. Its performance doesn't increase while employing four hidden layers, indicating that three hidden layers are the best amount for this issue.

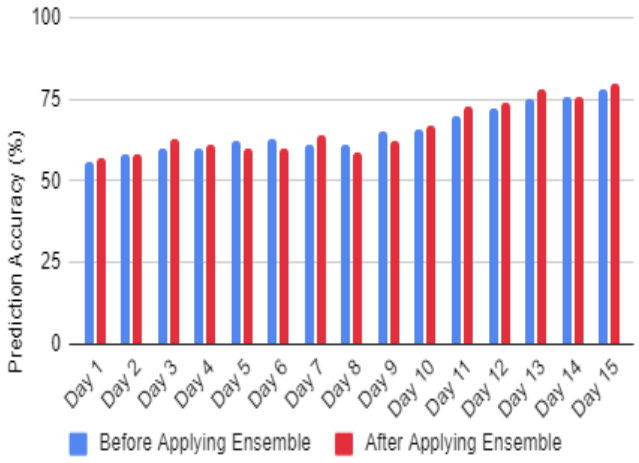

Fig. 17. Results of Voting Classifier on RF algorithm accuracy

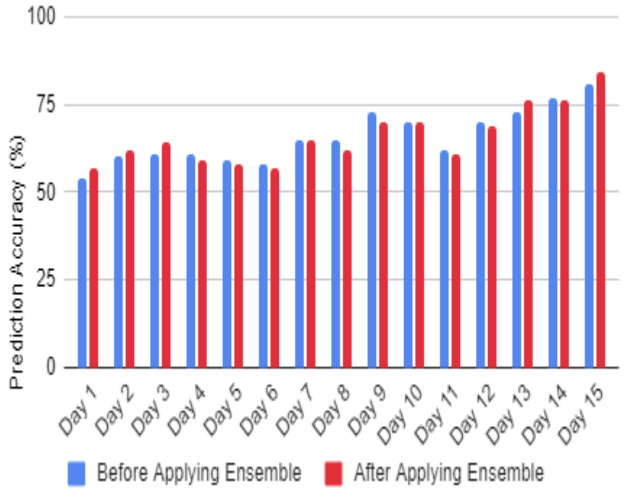

Fig. 18. Results of Voting Classifier on ET algorithm accuracy

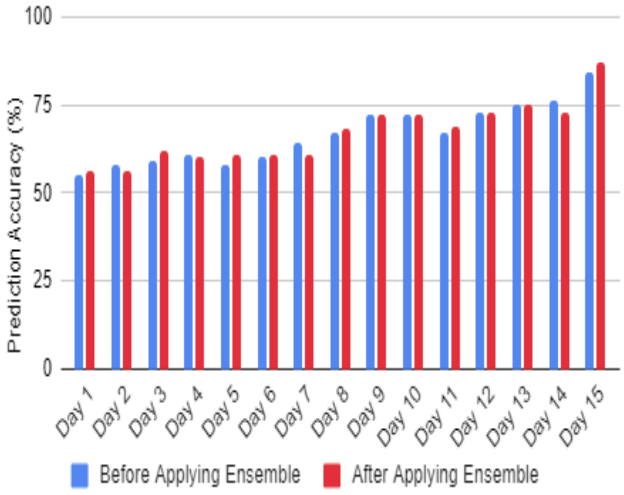

Fig. 19. Results of Voting Classifier on RF algorithm accuracy

5.9 Hybrid algorithm

The prediction accuracies of the ET, RF, and GBM classifiers increased when their separate predictions were merged. As illustrated in Figure 17, RF prediction accuracies are displayed before and after ensembles are applied to the IBM social media final dataset. After using the voting ensemble technique, the highest prediction accuracy of the RF classifier rises from 78.23% to 80.42%% on the 15th day.

Likewise, the ET classifier's maximum prediction accuracy rises from 81.16% to 84.4% on the 15th day (Figure 18), indicating a considerable improvement in accuracy. It's just 4th day on which does the accuracy of ET prediction drop.

Finally, as presented in Figure 19, the highest prediction accuracy of GBM enhances from 84.86% to 87.2% on the 15th day.

Based on the results, we can infer that ensemble techniques improve classifier prediction accuracy and may be applied in any field, which includes stock prediction, to improve the accuracies of individual classifiers.

6. Conclusion

The findings of this research suggest that social media and oil/gold price data can be effectively utilized to predict stock market trends. Our study showed that social media has a greater impact on stock prediction on the eighth day when sentiment features are considered, while oil or gold price fluctuations have a greater impact on the 15th and 14th days. The integration of social media emotions with oil price movements resulted in increased accuracy, and most of the classifiers demonstrated overall improved accuracy. Our research also explored various elements of data and algorithms used for prediction. Feature selection, feature extension, spam tweets reduction, and oil price trends had a positive impact on the performance of most classifiers. The impact of COVID-19 on the stock market was also analyzed, and it was found that the pandemic had both positive and negative impacts on different stock markets. In addition, we investigated the social media impact on different stock markets and found that IBM, NYSE, and TWTR equities were more impacted by social media. Moreover, we discovered that the VotingEnsemble technique using an ensemble of individual classifiers improved the accuracy of prediction for most of the classifiers. In conclusion, the use of social media and oil/gold price data as external inputs along with traditional historical data can improve the accuracy of stock market trend prediction. Our research highlights the importance of feature selection, feature extension, spam tweets reduction, and oil price trends for improved prediction performance. Moreover, the integration of COVID-19 impact analysis and social media sentiment analysis can provide a more comprehensive understanding of stock market trends.

5.10 Future Work

In order to further improve the accuracy of stock market prediction models, there are several areas that can be explored in future research. One such area is the analysis of the impact of the commodity market on stocks. It is well-known that fluctuations in commodity prices have a ripple effect on stock prices, and incorporating this data into prediction models could improve their performance. Additionally, the impact of global factors on stock markets is another important area of study. While it can be

challenging to quantify the influence of outside events on a country's financial markets, understanding this global factor effect could provide valuable insights for investors. However, due to the complexity and multitude of factors involved, it is a challenging task to accurately link the impact of different events from other countries on the stock market of any given country. Further research in these areas could lead to more accurate and reliable stock market prediction models, and provide valuable information for investors to make informed financial decisions.

References

[1] - Hegazy O, Soliman OS, Salam MA (2014) A machine learning model for stock market prediction. Int J Comput Sci Telecommun 4(12):16–23

[2] - Shen S, Jiang H, Zhang T (2012) Stock market forecasting using machine learning algorithms. Department of Electrical Engineering, Stanford University, Stanford, pp 1–5

[3] - Chen L, Qiao Z, Wang M, Wang C, Du R, Stanley HE (2018) Which artifcial intelligence algorithm better predicts the Chinese stock market? IEEE Access 6:48625–48633

[4] - Yetis Y, Kaplan H, Jamshidi M (2014) Stock market prediction by using artifcial neural network. In: IEEE WAC, pp 718–722

[5] - Ou P, Wang H (2009) Prediction of stock market index movement by ten data mining techniques. Mod Appl Sci 3(12):28

[6] - Urolagin S (2017) Text mining of tweet for sentiment classifcation and association with stock prices. In: IEEE ICCA, pp 384–388

[7] - Khatri SK, Srivastava A (2016) Using sentimental analysis in prediction of stock market investment. In: IEEE 5th international conference ICRITO, pp 566–569

[8] - 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

[9] - Dang LM, Sadeghi-Niaraki A, Huynh HD, Min K, Moon H (2018) Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. IEEE Access 6:55392–55404

[10] - Jeon S, Hong B, Chang V (2018) Pattern graph tracking-based stock price prediction using big data. J Future Gener Comput Syst. https://doi.org/10.1016/j.future.2017.02.010

[11] - Li Q, Wang T, Li P, Liu L, Gong Q, Chen Y (2014a) The efect of news and public mood on stock movements. J Inf Sci 278:826–840. https://doi.org/10.1016/j.ins.2014.03.096

[12] - Li X, Huang X, Deng X, Zhu S (2014b) Enhancing quantitative intra-day stock return prediction by integrating both market news and stock prices information. J Neuro Comput 142:228–238

[13] - Li X, Xie H, Chen L, Wang J, Deng X (2014c) News impact on stock price return via sentiment analysis. J Knowl-Based Syst 69:14–23. https://doi.org/10.1016/j.knosys.2014.04.022

[14] - Brown GW, Clif MT (2004) Investor sentiment and the near-term stock market. J Empir Financ 11(1):1–27

[15] - Sedhai S, Sun A (2018) Semi-supervised spam detection in Twitter stream. IEEE Trans Comput Soc Syst 5(1):169–175

[16] - He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of IEEE conference on CVPR 2016, pp 770–778

[17] - Noda K, Yamaguchi Y, Nakadai K, Okuno HG, Ogata T (2015) Audiovisual speech recognition using deep learning. J Appl Intell 42(4):722–737

[18] - Tsai CF, Lin YC, Yen DC, Chen YM (2011) Predicting stock returns by classifier ensembles. J Appl Soft Comput. https://doi.org/10.1016/j.asoc.2010.10.001

[19] - Tayal D, Komaragiri S (2009) Comparative analysis of the impact of blogging and micro-blogging on market performance. Int J Comput Sci Eng 1(3):176–182

[20] - Yuan B (2016) Sentiment analysis of Twitter data. M.S. thesis, Department of Computer Science, Rensselaer Polytechnic Institute, New York

[21] - Joshi R, Tekchandani R (2016) Comparative analysis of Twitter data using supervised classifers. In: IEEE international conference ICICT, 3 pp 1–6

[22] - Qasem M, Thulasiram R, Thulasiram P (2015) Twitter sentiment classifcation using machine learning techniques for stock markets. In: IEEE international conference on ICACCI, Kochi, India, pp 834–840

[23] - Chakraborty P, Pria US, Rony M, Majumdar MA (2017) Predicting stock movement using sentiment analysis of Twitter feed. In: IEEE 6th international conference ICIEV-ISCMHT, pp 1–6

[24] - Yan D, Zhou G, Zhao X, Tian Y, Yang F (2016) Predicting stock using microblog moods. J China Commun 13(8):244–257

[25] - Makrehchi M, Shah S, Liao W (2013) Stock prediction using eventbased sentiment analysis. In: IEEE/WIC/ACM international joint conference on WI and IAT,1, pp 337–342

[26] - Cao J, Cui H, Shi H, Jiao L (2016) Big data: a parallel particle swarm optimization-back-propagation neural network algorithm based on MapReduce. PLoS ONE 11(6):e0157551

[27] - Cheng S, Shi Y, Qin Q, Bai R (2013) Swarm intelligence in big data analytics. In: International conference on intelligent data engineering and automated learning. Springer, Berlin, pp 417–42

[28] - Blum C, Li X (2008) Swarm intelligence in optimization. In: Dorigo M(ed) Swarm intelligence. Springer, Berlin, pp 43–85

[29] - Hassanien AE, Emary E (2016) Swarm intelligence: principles, advances, and applications. CRC Press, Boca Raton

[30] - Dorigo M (1992) Learning and natural algorithms. Ph.D. Thesis, Politecnico di Milano, Milano, Italy

[31] - Eberhart R, Kennedy J (1995) Particle swarm optimization. In: Proceedings of the IEEE international conference on neural networks, pp 1942–1948

[32] - Brezočnik L, Fister I, Podgorelec V (2018) Swarm intelligence algorithms for feature selection: a review. Appl Sci 8(9):1521

[33] - Jayaraman V, Sultana HP (2019) Artifcial gravitational cuckoo search algorithm along with particle bee optimized

associative memory neural network for feature selection in heart disease classifcation. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-019-01193-6
[34] - Yang XS, Deb S (2009) Cuckoo search via Lévy fights. In: 2009 world congress on nature & biologically inspired computing (NaBIC). IEEE, pp 210–214
[35] - Wang G, Dai D (2013) Network intrusion detection based on the improved artificial fish swarm algorithm. J Comput 8(11):2990–2996
[36] - Li X (2003) A new intelligent optimization-artifcial fsh swarm algorithm. Ph.D. Thesis, Zhejiang University, Hangzhou, China
[37] - Enache AC, Sgarciu V, Petrescu-Niţă A (2015) Intelligent feature selection method rooted in Binary Bat Algorithm for intrusion detection. In: 2015 IEEE 10th Jubilee international symposium on applied computational intelligence and informatics. IEEE, pp 517–521
[38] - Yang XS (2010) A new metaheuristic bat-inspired algorithm. In Nature inspired cooperative strategies for optimization (NICSO 2010) Springer, Berlin, pp 65–74
[39] - Seth JK, Chandra S (2016) Intrusion detection based on key feature selection using binary GWO. In: 2016 3rd international conference on computing for sustainable global development (INDIACom). IEEE, pp 3735–3740
[40] - Mirjalili S, Mirjalili SM, Lewis A (2014) Grey wolf optimizer. J Adv Eng Softw 69:46–61
[41] - Mohammadi FG, Abadeh MS (2014) Image steganalysis using a bee colony based feature selection algorithm. J Eng Appl Artif Intell 31:35–43
[42] - Karaboga D (2005) An idea based on honey bee swarm for numerical optimization. Technical report-tr06, Erciyes University, Engineering Faculty, Computer Engineering Department, vol 200, pp 1
[43] - Chhikara RR, Sharma P, Singh L (2018) An improved dynamic discrete frefy algorithm for blind image steganalysis. Int J Mach Learn Cybern 9(5):821–835
[44] - Yang X-S (2008) Firefy algorithm. In: Nature-inspired metaheuristic algorithms. Luniver Press, Beckington, pp 128
[45] - Passino KM (2002) Biomimicry of bacterial foraging for distributed optimization and control. IEEE Control Syst Mag 22:52–67
[46] - Wang H, Jing X, Niu B (2016) Bacterial-inspired feature selection algorithm and its application in fault diagnosis of complex structures. In: 2016 IEEE congress on evolutionary computation (CEC). IEEE, pp 3809–3816
[47] - Sattiraju M, Manikantan K, Ramachandran S (2013) Adaptive BPSO based feature selection and skin detection based background removal for enhanced face recognition. In: 2013 4th national conference on computer vision, pattern recognition, image processing and graphics (NCVPRIPG). IEEE, pp 1–4
[48] - Hu Z, Chiong R, Pranata I, Susilo W, Bao Y (2016) Identifying malicious web domains using machine learning techniques with
[49] - Saraç E, Özel SA (2014) An ant colony optimization based feature selection for web page classifcation. Sci World J 2014:649260. https://doi.org/10.1155/2014/649260
[50] - Ibrahim RA, Ewees AA, Oliva D, Elaziz MA, Lu S (2019) Improved salp swarm algorithm based on particle swarm optimization for feature selection. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-018-1031-9
[51] - Mirjalili S, Gandomi AH, Mirjalili SZ, Saremi S, Faris H, Mirjalili SM (2017) Salp Swarm Algorithm: a bio-inspired optimizer for engineering design problems. Adv Eng Softw 114:163–191
[52] - Moslehi F, Haeri A (2019) A novel hybrid wrapper–flter approach based on genetic algorithm, particle swarm optimization for feature subset selection. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-019-01364-5
[53] - Zhong X, Enke D (2016) Forecasting daily stock market return using dimensionality reduction. Exp Syst Appl 67:126–139. https://doi.org/10.1016/j.eswa.2016.09.027
[54] - Sedhai S, Sun A (2015) HSpam14: a collection of 14 million tweets for hashtag-oriented spam research. In: 38th ACM conference on SIGIR, pp 223–232
[55] - Mork, K.A., Olsen, O. and Mysen, H.T., 1994. Macroeconomic responses to oil price increases and decreases in seven OECD countries. The Energy Journal, 15(4).
[56] - Chen W, Yeo CK, Lau CT, Lee BS (2017a) A study on real-time lowquality content detection on Twitter from the users' perspective. PLoS ONE 12(8):e0182487
[57] - Chen W, Zhang Y, Yeo CK, Lau CT, Lee BS (2017b) Stock market prediction using neural network through news on online social networks. In: IEEE international ISC2, pp 1–6
[58] - Al-Zoubi A, Faris H (2017) Spam profle detection in social networks based on public features. In: IEEE 8th international conference ICICS, pp 130–135
[59] - Hamilton, J.D., 1983. Oil and the macroeconomy since World War II. Journal of political economy, 91(2), pp.228-248.
[60] - Jones, D.W., Leiby, P.N. and Paik, I.K., 2004. Oil price shocks and the macroeconomy: what has been learned since 1996. The Energy Journal, 25(2).
[61] - Cunado, J. and De Gracia, F.P., 2005. Oil prices, economic activity and inflation: evidence for some Asian countries. The Quarterly Review of Economics and Finance, 45(1), pp.65-83.
[62] - Baláž, P. and Londarev, A., 2006. Ropa a jej postavenie v globalizácii svetového hospodárstva [Oil and its position in the process of globalization of the world economy]. Politická ekonomie, 2006(4), pp.508-528.
[63] - Metcalfe, K.A., Birenbaum-Carmeli, D., Lubinski, J., Gronwald, J., Lynch, H., Moller, P., Ghadirian, P., Foulkes, W.D., Klijn, J., Friedman, E. and Kim-Sing, C., 2008. International variation in rates of uptake of preventive options in BRCA1 and BRCA2 mutation carriers. International journal of cancer, 122(9), pp.2017-2022.
[64] - Cologni, A. and Manera, M., 2008. Oil prices, inflation and interest rates in a structural cointegrated VAR model for the G-7 countries. Energy economics, 30(3), pp.856-888.
[65] - Barsky RB, Kilian L. Oil and the macroeconomy since the 1970s. Journal of Economic Perspectives. 2004 Dec;18(4):115-34.
[66] - Hamilton, J.D., 1988. Rational-expectations econometric analysis of changes in regime: An investigation of the term structure of interest rates. Journal of Economic Dynamics and Control, 12(2-3), pp.385-423.

[67] - Mork, K.A., 1989. Oil and the macroeconomy when prices go up and down: an extension of Hamilton's results. Journal of political Economy, 97(3), pp.740-744.
[68] - Ferderer, J.P., 1996. Oil price volatility and the macroeconomy. Journal of macroeconomics, 18(1), pp.1-26.
[69] - Davis, S.J. and Haltiwanger, J., 2001. Sectoral job creation and destruction responses to oil price changes. Journal of monetary economics, 48(3), pp.465-512.
[70] - Lee, K. and Ni, S., 2002. On the dynamic effects of oil price shocks: a study using industry level data. Journal of Monetary economics, 49(4), pp.823-852.
[71] - Hamilton, J.D., 2003. What is an oil shock?. Journal of econometrics, 113(2), pp.363-398.
[72] - Zheng, P., Allen, W.B., Roesler, K., Williams, M.E., Zhang, S., Li, J., Glassman, K., Ranch, J., Nubel, D., Solawetz, W. and Bhattramakki, D., 2008. A phenylalanine in DGAT is a key determinant of oil content and composition in maize. Nature genetics, 40(3), pp.367-372.
[73] - Lardic, S. and Mignon, V., 2006. The impact of oil prices on GDP in European countries: An empirical investigation based on asymmetric cointegration. Energy policy, 34(18), pp.3910-3915.
[74] - Lardic, S. and Mignon, V., 2008. Oil prices and economic activity: An asymmetric cointegration approach. Energy Economics, 30(3), pp.847-855.
[75] - Cologni, A. and Manera, M., 2009. The asymmetric effects of oil shocks on output growth: A Markov–Switching analysis for the G-7 countries. Economic Modelling, 26(1), pp.1-29.
[76] - Kilian, L., 2008. The economic effects of energy price shocks. Journal of economic literature, 46(4), pp.871-909.
[77] - Kilian, L., 2009. Not all oil price shocks are alike: Disentangling demand and supply shocks in the crude oil market. American Economic Review, 99(3), pp.1053-69.
[78] - Jones, C.M. and Kaul, G., 1996. Oil and the stock markets. The journal of Finance, 51(2), pp.463-491.
[79] - Diatchenko, L., Lau, Y.F., Campbell, A.P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E.D. and Siebert, P.D., 1996. Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. Proceedings of the National Academy of Sciences, 93(12), pp.6025-6030.
[80] - Henriques, I. and Sadorsky, P., 1999. The relationship between environmental commitment and managerial perceptions of stakeholder importance. Academy of management Journal, 42(1), pp.87-99.
[81] - Ciner, C., 2001. Energy shocks and financial markets: nonlinear linkages. Studies in Nonlinear Dynamics & Econometrics, 5(3).
[82] - Park, J. and Ratti, R.A., 2008. Oil price shocks and stock markets in the US and 13 European countries. Energy economics, 30(5), pp.2587-2608.
[83] - Apergis, N. and Miller, S.M., 2009. Do structural oil-market shocks affect stock prices?. Energy economics, 31(4), pp.569-575.
[84]Wassan, S., Chen, X., Shen, T., Waqar, M., & Jhanjhi, N. Z. (2021). Amazon product sentiment analysis using machine learning techniques. Revista Argentina de Clínica Psicológica, 30(1), 695.
[85]Dogra, V., Singh, A., Verma, S., Kavita, Jhanjhi, N.Z., Talib, M.N. (2021). Analyzing DistilBERT for Sentiment Classification of Banking Financial News. In: Peng, SL., Hsieh, SY., Gopalakrishnan, S., Duraisamy, B. (eds) Intelligent Computing and Innovation on Data Science. Lecture Notes in Networks and Systems, vol 248. Springer, Singapore. https://doi.org/10.1007/978-981-16-3153-5_53
[86]Chouhan, K., Yadav, M., Rout, R. K., Sahoo, K. S., Jhanjhi, N. Z., Masud, M., & Aljahdali, S. Sentiment Analysis with Tweets Behaviour in Twitter Streaming API.

Dr. Mustansar Ali Ghazanfar is a Senior Lecturer and Course Leader for AI at the University of East London. He earned his M.S. and Ph.D. degrees in software engineering and electrical engineering, respectively, from the University of Southampton, UK. With research interests in data mining, machine learning, and recommender systems, he leads a team of M.S. and Ph.D. students in the field of machine learning and recommender systems. He has published more than 60 articles in his field of expertise.

Madiha Anwar hold MSC in Artificial Intelligence from University of East London. Her area of interest is Machine Learning, Fintech, and stock market predictions.

Amin Karami received the BS degree in computer engineering from Qazvin

Islamic Azad University, Iran in 2008, and MS in informatics from University of Skövde in 2011. He received the Ph.D. degree in computer science from Universitat Politècnica de Catalunya, Spain in 2015. He is a Senior Lecturer in the Department of Computer Science and Digital Technologies (CS&DT). He is the programme leader for MSc Big Data Technologies and the PG academic leader at CS&DT. His research interests include big data technologies, computational intelligence, blockchain and optimization techniques.

Nadeem Qazi earned his PhD in Machine Learning (ANN) from Cranfield

University UK.

He is a Senior Lecturer in Computer Science and Informatics at the School of Architecture, Computing and Engineering. He is the Co-Head of Intelligent Systems Research Group. His research interests include AI, machine learning, computer vision, text mining, data visualisation and analytics, mobile application and social media mining.

Prof. Dr. Noor Zaman Jhanjhi (N.Z Jhanjhi) is currently working as **Professor** in Computer Science, **Program Director** for the Postgraduate Research Programmes in computer science, at the School of Computer Science at Taylor's University, Malaysia. He has been nominated as the **world's top 2%** research scientist globally for the year **2022**. He has been nominated Malaysia's number **1st** researcher by Scopus in terms of research publications for the year 2022. He has been nominated as **outstanding faculty** by the **MDEC Malaysia** for the year 2022. He has highly indexed publications in WoS/ISI/SCI/Scopus, and his collective research Impact factor is more than **700** plus points. His H index is **42**, while more **than 500** publications are on his credit. He has several international Patents on his account, including **Australian, German, and Japanese**. He edited/authored more than **40** research books published by world-class publishers, including **Springer, Taylors and Frances, Willeys, Intech Open, IGI Global USA**, etc. He has excellent experience supervising and co-supervising postgraduate students, and more than **30 Postgraduate scholars** graduated under his supervision. Dr. Jhanjhi serves as Associate Editor and Editorial Assistant Board for several reputable journals, such as PeerJ Computer Science, CMC Computers, Materials & Continua, Computer Systems Science and Engineering CSSE and Frontier in Communication and Networks. He received **Outstanding Associate Editor** for IEEE ACCESS. Active reviewer for a series of top-tier journals and has been awarded globally as a **top 1%** reviewer by Publons (Web of Science). He is an external Ph.D./Master **thesis examiner/evaluator** for **several universities** globally. He has completed more than **40** internationally funded research grants successfully. He has served as a **Keynote/Invited speaker** for more than **60** international conferences globally, **chaired** international conference sessions internationally. He has vast experience in academic qualifications, including **ABET, NCAAA, and NCEAC** for **10** years. His research areas include Cybersecurity, IoT security, Wireless security, Data Science, Software Engineering, and UAVs.
https://expert.taylors.edu.my/cv/noorzaman.jhanjhi

Engr. Dr. Ali Javed is an Associate Professor in the Software Engineering

department at UET Taxila. With a Ph.D. in Computer Engineering from UET Taxila and the University of Michigan, USA, and a Post Doctorate from Oakland University, MI, USA, his research interests include video summarization, image processing, computer vision, software quality, multimedia forensics, machine learning, and medical image processing. As an accomplished academic, he has authored numerous publications and has a vast knowledge of his field.

Dr Sin Wee Lee is an accomplished academic with a PhD in

Neurocomputing and currently serves as the Deputy Head of the School of Computing at Arden University. His research interests include artificial neural networks, network architecture, semi-supervised learning, and computer networking. With his expertise in these areas, he has contributed significantly to the field of artificial intelligence and its application in network management and computer networks.