Steyn, Dimitri H. W.; Greyling, Talita; Rossouw, Stephanie; Mwamba, John M.

Working Paper

Sentiment, emotions and stock market predictability in developed and emerging markets

GLO Discussion Paper, No. 502

Provided in Cooperation with:
Global Labor Organization (GLO)

Suggested Citation: Steyn, Dimitri H. W.; Greyling, Talita; Rossouw, Stephanie; Mwamba, John M. (2020) : Sentiment, emotions and stock market predictability in developed and emerging markets, GLO Discussion Paper, No. 502, Global Labor Organization (GLO), Essen

This Version is available at:
http://hdl.handle.net/10419/215436

Sentiment, emotions and stock market predictability in developed and emerging markets

Dimitri H. W. Steyn[1], Talita Greyling[2], Stephanie Rossouw[3], John M. Mwamba[4]

Highlights

- Sentiment and emotion derived from Tweets predict stock market movements.
- Machine learning algorithms accurately predict movements in stock markets.
- High-frequency, not only daily data, are significant predictors of stock markets.
- We find significant predictions in developing and emerging markets.

Abstract This paper investigates the predictability of stock market movements using text data extracted from the social media platform, Twitter. We analyse text data to determine the sentiment and the emotion embedded in the Tweets and use them as explanatory variables to predict stock market movements. The study contributes to the literature by analysing high-frequency data and comparing the results obtained from analysing emerging and developed markets, respectively. To this end, the study uses three different Machine Learning Classification Algorithms, the Naïve Bayes, K-Nearest Neighbours and the Support Vector Machine algorithm. Furthermore, we use several evaluation metrics such as the Precision, Recall, Specificity and the F-1 score to test and compare the performance of these algorithms. Lastly, we use the K-Fold Cross-Validation technique to validate the results of our machine learning models and the Variable Importance Analysis to show which variables play an important role in the prediction of our models. The predictability of the market movements is estimated by first including sentiment only and then sentiment with emotions. Our results indicate that investor sentiment and emotions derived from stock market-related Tweets are significant predictors of stock market movements, not only in developed markets but also in emerging markets.

Keywords: Sentiment Analysis, Classification, Stock Prediction, Machine Learning

JEL classification codes: C6, C8, G0, G4

1. Introduction

In this paper, our main aim is to investigate whether text extracted from social media, such as Twitter, and analysed to determine the sentiment and emotions of the text, predict stock market movements. Contrary to previous studies using daily data, we focus on high-frequency intraday data to accommodate investment decisions that are made continuously throughout the day. Furthermore, we investigate whether the sentiment and emotions of investors result in similar predictions of stock market movements in developed as well as emerging stock markets.

[1] School of Economics, University of Johannesburg, P O Box 524, Auckland Park, 2006, South Africa, email: dimitris@uj.ac.za
[2] Corresponding author: School of Economics, University of Johannesburg, P O Box 524, Auckland Park, 2006, South Africa, email: talitag@uj.ac.za , telephone number: +27 11 5592586, fax number: +27 11 5593039.
[3] Faculty of Business, Economics and Law, Auckland University of Technology, Private Bag 92006, Auckland, 1142, New Zealand, email: stephanie.rossouw@aut.ac.nz , telephone number: +64 9 921 9999 ext. 5710, fax number:+64 9 921 9340.
[4] School of Economics, University of Johannesburg, P O Box 524, Auckland Park, 2006, South Africa, email: johnmu@uj.ac.za

In recent years, the popularity of social media platforms such as Twitter have experienced tremendous global growth. In 2019, there were 3.2 billion social media users worldwide, accounting for approximately 42 per cent of the world population (Mohsin 2019). This allowed researchers to analyse the impact of social media on various study fields. Over the same period, Twitter has also expanded significantly, with 7 million new users each year (Statista 2019) and a total number of 330 million active daily users in 2019.

Stock market movement is explained as the up and down shift of a stock market, i.e. the deviations from its previous value. The upward shift represents positive returns, while the downward shift represents negative returns. Investor sentiment refers to the general perception (mood) of an individual stock or financial market. Traditionally, we derive investor sentiment through financial market measures such as trading volumes, average bond yield returns and put/call ratios (Bonga-Bonga & Mwamba 2011, Bodie et al. 2014). However, these do not accurately capture the sentiment or emotion of investors, which is often a catalyst in stock market price fluctuations (Da et al. 2015).

A more direct measure is needed that captures the psychological element of an investor's decision-making, which can explain deviations from fundamental values (Smith 2019). A likely solution to this problem is to analyse investor sentiment and emotions[5] from text extracted from social media, such as Twitter, which captures the psychological element of an investor's decision-making. This gives us the capability to instantaneously determine how investors feel about a particular stock, which contributes to a better understanding of the psychological determinants that drive the dynamics of stock markets.

Previous studies have analysed the relationship between the sentiment *or* emotions embedded in stock market-related text (also Tweets) and the performance *of either* individual stocks or the stock market as a whole[6] (Mc Kay 2018, Renault 2017, Broadstock & Zhang 2019). However, these studies considered *daily data* and not intraday high-frequency data. The advantage of using high-frequency data lies with investment decisions made throughout market trading hours, thus a need for high frequency predictions.

Furthermore, studies mainly analysed the predictability of stock market movements, using investor sentiment, in *developed markets* (Bollen et al. 2011, Zhang et al. 2010), whereas studies on emerging markets is very limited (Maree & Johnston 2015). No other study, to the knowledge of the authors, analysed both types of markets simultaneously, to determine whether *investor sentiment and emotions,* extracted from global Tweets, affects emerging and developed markets differently. In the current study, we derive investor sentiment from a set of global Tweets (adjusted for time

[5] See Greyling, T., Rossouw, S. & Afstereo. (2019). Gross National Happiness Index. University of Johannesburg and Afstereo [producers]. http://gnh.today/www.gnh.today, for similar analyses to measure sentiment,

[6] This happens through predicting the ETF (Exchange Traded Funds) that acts as a proxy for the entire market.

differences), allowing us to analyse the responsiveness (reactions) of emerging and developed markets to the same set of Tweets.

Additionally, studies that used Twitter for investor sentiment analysis have traditionally used conventional lexicons such as Hu and Liu (2004), Wordnet and WordNet-Affect (Jagdale et al. 2016). As pointed out by Chung and Liu (2011), some of these lexicons contain a disproportionate quantity of positive to negative sentiment ratios and do not contain sector-specific terms that could help to correctly capture the sentiment. The current paper employs the '*syuzhet*' package, which has traditionally been applied to sources of data such as TripAdvisor (Valdivia et al. 2017), fiction (Zehe et al. 2016) and Tweets to detect trending sentiments in political elections (Kolagani et al. 2017). This package is ideal for the analysis of stock market related Tweets. In saying that, we are aware of two studies that used the "syuzhet lexicon" concerning financial markets; Ageitos (2018) who studied the London Stock Exchange and Moritz (2018) who investigated financial asset pricing. However, these papers only used sentiment analyses and did not venture into the analyses of emotions, allowed for by the 'syuzhet package' (Naldi 2019, Elodie 2019), which makes our study unique.

Moreover, previous papers used, amongst other, standard statistical analysis or econometric methods to test the relationship between sentiment (emotions) and stock markets. These include methods such as correlation analysis, Granger Causality tests, and Ordinary Least Square estimations (Bollen et al. 2011, Zhao 2019, You et al. 2017, Nisar & Yeung 2018, Shen et al. 2018,). Some papers used more advanced methods and ventured into machine learning including Fuzzy or Neural Network, Support Vector Machine and Random Forest models (Bollen et al. 2011, Cropper 2011, Maree & Johnston 2015, Jadhav & Wakode 2017, Tabari et al. 2018, Maqsood et al. 2020), but none have *applied different machine learning approaches to test the robustness of the results.*

Considering the gaps mentioned above, this study expands upon existing literature on investor sentiment analysis and stock market prediction using the Twitter Application, with the following contributions:

i) Focusing on intraday high-frequency data rather than daily data (for robustness purposes we also report on daily data). Additionally, we use data extracted over nearly a one-year period. Similar studies use data from much shorter time periods, often only a few months or less.

ii) Comparing results for emerging and developed markets.

iii) Using both investor sentiment and emotions, such as fear, joy, anticipation, anger and trust of investors to predict market movements.

iv) Using three different Machine Learning Algorithms (Naïve Bayes, K-Nearest Neighbours and Support Vector Machines), utilising the evaluation metrics namely the Precision, Recall, Specificity and the F-1 score to compare these algorithms.

Additionally, we use the K-Fold Cross-Validation technique to validate the performance of the results for our machine learning models and the Variable Importance Analysis (VIA) to show which variables play an important role in the prediction of our models.

We analyse eight stock markets, including six developed countries (France, Germany, the UK, the USA, Japan and Spain) and two emerging markets (Poland and India). Our selection of the markets was determined by the frequency of the use of Twitter in those markets and the availability of high-frequency market-related data.

The rest of the paper is structured as follows: section 2 reviews the applicable theories and literature, section 3 explains the methodology followed, section 4 discusses the data, section 5 reports and presents the results, while section 6 concludes.

2. Literature review

Most studies cited in this section and elsewhere made use of daily data in the prediction of stock market returns or movement. As for studies focusing on *intraday data*, we direct the reader to the works of Bukovina (2016), in which he provides an overview of academic research related explicitly to the relationship between social media and capital markets. The overview is divided between social media platforms used to extract the data, namely Twitter, Facebook and Google. Bukovina (2016) analyses intraday data ─ though he does not consider stock market indices ─ rather individual-level stock returns. Furthermore, he analyses very high-frequency data at five-minute volatility, measured by absolute 5-minute returns, and Twitter sentiment and activity. He finds some statistically significant co-movements of intraday volatility and information from stock-related Tweets for all constituents of the Dow Jones Industrial Average. However, economically, the effects are of a negligible magnitude, and out-of-sample forecast performance is not improved when including Twitter sentiment and activity as exogenous variables. From a practical point of view, he finds that high-frequency Twitter information is not particularly useful for highly active investors with access to such data for intraday volatility assessment and forecasting, when considering individual-level stocks.

Very few studies have focused on *emerging stock market predictability*. Maree and Johnston (2015) extracted over 3.1 million Tweets within South Africa over a 55-day period to analyse the impact of emotion on stock market returns, using data from the JSE-ALSI (Johannesburg Securities Exchange All Share Index). The paper used a Granger causality method and a Spearman correlation test, and in contrast to the earlier studies, the paper found that the 'fatigue' emotion had a significant positive correlation with the JSE-ALSI market movements. A machine learning model, applied in the study, confirmed that the prediction of JSE-ALSI values does improve with the inclusion of the fatigue mood. The paper also found a significant negative correlation with the depressed mood and the JSE-

ALSI movements. Even though this paper extended the work done by Bollen et al. (2011), it did not address the effect of investor sentiment on other financial markets, neither individually nor holistically. The paper also has its limitations in that it used only 39 days of data collected on the JSE-ALSI and suggested that further studies be conducted using longer time-periods.

Bhardwaj et al. (2015) investigated the Indian stock market and focused on predicting the stock market status of the Sensex and NIFTY. These two market indexes represent the stocks for BSE (Bombay Stock Exchange) and NSE (National Stock Exchange), respectively. Specifically, under BSE there are 30 companies for Sensex, while under NSE there are 50 companies for Nifty. To do this, the authors extracted Sensex and Nifty live server data values at different intervals of time that could be used for predicting the stock market status. The drawback of this paper is its complete lack of sentiment classification techniques and, as a consequence, the fact that it does not make any real contribution to the prediction of stock market returns.

Maqsood et al. (2020), investigates a similar research question to ours but with a significant difference. In theirs, they consider four countries, namely the US, Hong Kong, Turkey and Pakistan, and use deep learning-based models along with event sentiment for stock exchange prediction. They explore the effect of some of the most significant events from 2012 to 2016. These events are categorised into local and global events for each country according to their impact. For example, for the US they use the 2012 Mexican and US elections as local events, and Gaza under attack in 2014, Brexit 2016 and Refugee Welcome in 2015 as global events. Maqsood et al. (2020) use a Twitter dataset to calculate the sentiment analysis for each of their eight events. Their results show that stock market performance improves by using the sentiment for significant events. However, they calculate investor sentiment by using an intensive dataset of Tweets regarding international events and do not include the usage of emotions.

Das and Chen (2007) and Antweiler and Frank (2004) are some of the most prominent studies that analysed *investor sentiment* using a social media platform. Das and Chen (2007) used stock message boards, whereas Antweiler and Frank (2004) analysed over 1.5 million messages posted on Yahoo! Finance and attempted to investigate whether stock message boards can cause stock price changes for 45 companies. Accordingly, Das and Chen (2007) found evidence supporting a relationship between stock returns and investor sentiment. The study also found that investor sentiment through social media applications contains an idiosyncratic component.

Similarly, Antweiler and Frank (2004) found ─ using the Dow Jones index with a Naïve Bayes model and Time-series panel regressions ─ that stock messages assist in the prediction of market volatility. Although the results obtained did not yield a big economic impact, they generated a statistically significant result. The authors also conclude that message posting assists in the prediction of volatility, and suggest that message postings can provide helpful insights into studies using high-

frequency data analysis. However, neither of these studies employed the effect of emotion in stock prediction and relied on computational linguistics methods.

Oliveira et al. (2017) summarise the literature by distinguishing between specific dimensions relevant to the papers. For example, the source from which data was extracted such as blogs, financial data, and Google searches; the methods followed to analyse the sentiment; the methods used to merge the distinct sources; the period (daily or monthly) used to analyse the data; the type of stocks analysed, such as individual or portfolio; the methods used to predict the relationships, for example, multiple regression; and the statistical tests used to verify the significance of the sentiment. None of the works cited by Oliveira et al. (2017) attempted to predict survey sentiment indices, which is addressed in this study. Earlier studies, from 1988 to 2010, adopted surveys, financial data, message boards (e.g., ragingbull.com) and news (e.g., Wall Street Journal) to create the sentiment and attention indicators. After 2011, Web 2.0 services, such as microblogs (e.g., Twitter, StockTwits) and Google searches, have also been adopted. Some financial measures (e.g., closed-end fund discount) and survey values, such as the American Association of Individual Investors (AAII) have also been used (Oliveira et al. 2017).

Yang et al. (2015) considered market influencers and found that there is evidence of a financial community on Twitter and that the weighted sentiment of its most influential contributors has significant predictive power for market movement. Similarly, in our study, we are cognizant of the *effect of market influencers, but also recognise the importance of viral news*, when extracting Tweets.

In terms of *emotions* and stock market predictability, there are several studies which investigate these relationships, such as Tabari et al. (2018), Zhang et al. (2010), Bollen et al. (2011), Maree and Johnston (2015) and Rao and Srivastava (2012). Of specific noteworthiness is the study done by Zhang et al. (2010) who extracted Twitter feeds over a six-month period, ranging from 8100 to over 43000 daily Tweets and then extracted public moods on Twitter and used three different baseline measures, including the volume of daily Tweets, the total number of followers and the retweet volume. The study extracted Tweets by filtering for emotional keywords such as ‘fear’, ‘hope’ and ‘worry’. The Twitter mood feeds were analysed with the stock returns from the S&P500, Dow Jones and NASDAQ indices, in an attempt to predict stock market movements using sentiment analysis. The paper, using correlation analysis, found that emotions, and in particular, ‘hope’, ‘worry’ and ‘fear’, seem to display a significant negative correlation with the three stock indices. The paper also found, however, these same emotions to correlate positively with stock market volatility. However, Zhang et al. (2010), only applied a correlation analysis on the stock indices and mood sentiment and failed to employ more advanced methods such as predictive machine learning algorithms. Moreover, the paper employed a basic approach to capturing emotion by merely counting the daily Tweets containing the keywords and using this as a metric for emotions.

Bollen et al. (2011) investigate whether public emotions can predict the Dow Jones index. The authors use Twitter data, which they collected over a 10-month period, which amounted to almost 10 million Tweets. With this data, they generated a multidimensional time series of public emotions, which included six dimensions, namely being calm, alert, sure, vital, kind and happy. To determine if these emotions did predict market movements, they made use of Granger causality and a machine learning approach known as a 'Fuzzy Neural Network'. They found that positive and negative emotions do not improve prediction accuracy any better than a baseline model based purely on its past predictions of stock market performance. However, when only using the emotions 'calm' and 'happy', the accuracy of the model to predict the stock market movements improved significantly. They found a reduction of 6 per cent in prediction error and an accuracy of 87.6 per cent in determining the up or down movements in the market.

The results of Bollen et al. (2011) is in direct contrast to that of Zhang et al. (2010) who found significance in using positive and negative sentiment in the prediction of stock market price changes. However, the results of Bollen et al. (2011) indicate the significance of including public emotions in the prediction of stock market price changes. A drawback of the Bollen et al. (2011) study lies with their failure to analyse the effect of global stock prediction and only focusing on the Dow Jones index. Furthermore, this study was implemented at a time when Twitter had 68 million users. Since then, Twitter has grown to 330 million users, representing an increase of over 385 per cent and, as highlighted by studies such as those of Bollen et al. (2011) and Abbes (2015), to employ a study with Twitter having grown so much in user base and popularity, could provide better insights into stock predictability.

Considering the above review of the literature (also see Table A1 in Appendix A for an overview of studies), we find that none of these studies specifically analysed high-frequency intraday data and secondly, the majority of the works consider single stock market indices in developed countries such as the Dow Jones or the S&P 500. Very few analyse more than one stock market, and there are a minimal number of papers that analyse emerging markets (Maree & Johnston 2015, Bhardwaj et al. 2015).

3. Methodology

In this section we explain (i) the machine learning classification algorithms; (ii) the evaluation metrics used to determine how good the predictions of the models are; (iii) the K-Fold Cross-Validation as a robustness test and (iv) the Variable Importance Analysis (VIA) to determine which variables have the most significant impact on the prediction of stock returns. As mentioned earlier, we make use of high-frequency data in our analyses; however, we do repeat all classification algorithms, evaluation metrics, the K-Fold Cross-Validation and the VIA, using daily data, as a robustness test.

3.1 Machine Learning Classification Algorithms

This study makes use of three Machine Learning Classification Algorithms, namely Naïve Bayes, K-Nearest Neighbours and the Support Vector Machine.

3.1.1 Naïve Bayes (NB)

The Naïve Bayes classifier is a supervised machine learning algorithm that is based on the Bayes theorem. The NB classifier is a well-known and extensively used benchmark and evaluation model for classification problems relating specifically to text-based categorisation (Rennie et al. 2003). This makes the NB model the ideal choice as a benchmark and evaluation machine learning model to compare against other more sophisticated machine learning classification models. In essence, the NB is a simple technique to develop a way of classifying data, relying on the assumption that the features within the data are independent of one another. Given a classification problem, the NB can be represented by:

$$\boldsymbol{y} = (y_1, y_2, y_3, \dots, y_n) \quad (1)$$

with $\boldsymbol{y}$ being a vector with *n*-independent features, assigning the probabilities:

$$P(K_m \mid y_1, y_2, y_3, \dots, y_n) \quad (2)$$

for each possible class K_m. This method becomes cumbersome when the number of features in the model becomes large. Thus, the model can be reconstructed as:

$$P(K_m \mid \boldsymbol{y}) = \frac{P(K_m)P(\boldsymbol{y} \mid K_m)}{P(\boldsymbol{y})} \quad (3)$$

in which $P(K_m)$ is the prior probability, $P(\boldsymbol{y} \mid K_m)$ is the likelihood, $P(\boldsymbol{y})$ is the evidence and $P(K_m \mid \boldsymbol{y})$ is the posterior probability. However, since the NB assumes independence, only the numerator is of interest. Thus, it can be expressed as:

$$P(K_m, y_1, y_2, y_3, \dots, y_n) \quad (4)$$

The joint model, under the NB assumption, can equivalently be expressed as:

$$P(K_m) \prod_{i=1}^{m} P(y_i \mid K_m) \quad (5)$$

Despite the simplicity in the assumptions of the NB, studies such as that of Rennie et al. (2003), find that the NB model can perform competitively with complex models such as the SVM.

3.1.2 K-Nearest Neighbours (KNN)

In machine learning and pattern recognition, the K-NN model is a method employed in supervised machine learning for classification and is a non-parametric method. The K-NN approach is a simple

method that stores and classifies unseen data based on a similarity measure. Principally, the data is classified according to the proximity the data has to its neighbours; the data is assigned a class most common among a selected *k* nearest neighbours.

The K-NN is an instance-based learning algorithm, in that, the K-NN hypothesis about the training data can evolve with the data. That is, the K-NN can compare new instances with prior seen instances instead of explicit generalisations about the data. The advantage of the K-NN is that it can adapt itself to new unseen data points. Common distance methods used to classify data with the K-NN method are shown in equations 6 to 8:

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad Euclidean\ Distance \tag{6}$$

$$\sum_{i=1}^{k}|x_i - y_i| \quad Manhattan\ Distance \tag{7}$$

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{\frac{1}{q}} \quad Minkowski\ Distance \tag{8}$$

As a general rule-of-thumb, the larger the *k*-value, the more accurate the classification, as it dramatically reduces noise. However, the rule-of-thumb suggests the following:

$$k = \sqrt{n} \tag{9}$$

However, when dealing with a 2-class problem and $\sqrt{n}$ being even, equation 9 becomes:

$$k = \sqrt{n} \pm 1 \tag{10}$$

where k is the number of neighbours and n is the total number of data points. For binary classification models, the *k* parameter should be an odd number to avoid tied votes (Hall et al. 2008). Following Qian and Rasheed (2007), the Euclidean distance was selected.

3.1.3 Support Vector Machine (SVM)

A Support Vector Machine (SVM) is a supervised machine learning approach that analyses data through a classification algorithm. Provided with a set of training data, the instances are labelled according to two classes, thus resulting in the SVM being a binary machine learning classifier. The objective of the SVM model is to obtain a hyperplane from an *n*-dimensional space (with *n* being the number of features in the model) that classifies the data into either one of two classes.

For classification to occur, there has to be a hyperplane that separates the two classes. Several hyperplanes could fit the SVM model; however, the objective is to find a plane that has the maximum

margin. In other words, suppose data points exist that belong to either one of two classes, and the objective is to determine the class for a new data point. In the case of the SVM, a data point is an *n*-dimensional vector space, and the hyperplane is chosen as an (*n-1*)-dimensional hyperplane to separate the data.

In an attempt to obtain a hyperplane that separates the set of data into two classes: "up" and "down" with two optimal margin lines called support vectors can be given by:

$$w_0^T \chi + b_0 = 1 \tag{11}$$

$$w_0^T \chi + b_0 = -1 \tag{12}$$

in which w is a weight vector, x is a vector made up of the inputs and b is the bias. The SVM, therefore, solves the following quadratic optimisation problem:

$$\min_w \frac{1}{2} w^T \cdot w \tag{13}$$

$$\text{Subject to: } y_i(w^T \cdot x_i + b) \geq 1 \tag{14}$$

Equation 14 represents a linear separation line. However, if the separation line is nonlinear, one is required to use the Kernel trick by replacing the linear equation 14 with a kernel function of the data.

3.2 Evaluation Metrics

This section discusses the evaluation metrics used to evaluate how good the predictions are, made by our machine learning classification algorithms.

To describe the performance of a classification model on a set of test data for which the true values are known, we make use of four parameters. These parameters are namely: true positives, true negatives, false positives and false negatives. True positive and true negatives are the observations that are correctly predicted. In contrast, the false positives and false negatives are those not correctly predicted, and the aim is to minimise these observations.

3.2.1 Accuracy

Accuracy is the most intuitive evaluation metric and is known as the ratio of the predictions that were classified correctly to the total number of observations. Accuracy is defined as follows:

$$Classification\ Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{15}$$

3.2.2 Precision

Precision is defined as the total number of true positives compared to the total number of predicted positives. Precision is a good measure to determine when the costs of false positives are high and are expressed as follows:

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

3.2.3 Recall

Recall is defined as the total proportion of true positives to the total number of actual positives. Recall is expressed as follows:

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

For prediction purposes, ideally, the precision and recall should both be maximised. In this case, combining these two metrics in a single, simple metric, provides what is known as the F1-Score.

3.2.4 The F1-Score

The F1-Score is the harmonic mean between the precision and the recall. The harmonic mean is chosen over the simple average, because it penalises extreme values. To achieve a balanced model between recall and precision, the F1-Score has to be close to 100 per cent. Hence, the F1-Score will tell us how balanced the data is. The F1-score is defined as follows:

$$F1 = 2 * \frac{(Precision * Recall)}{(Recall + Precision)} \tag{18}$$

3.3 K-Fold Cross-Validation

K-Fold Cross-Validation is a model validation technique that validates the performance results of a machine learning model. Although the splitting of data into a train-test split is an approach to model validation (Sanjay 2018), as performed in this study, it can lead to high bias. This is because one could miss out on important information within the data that was not used for training or testing. This tends to be the case when dealing with limited data, such as daily data. However, using intraday data, reflecting hourly stock returns, which is vast in numbers, one could expect consistent results.

More specifically, the K-Fold Cross-Validation technique gives us a comprehensive measure of our model's performance throughout the *whole dataset.* The method is trained and validated multiple times and is more thorough than the previous evaluation methods, as it can potentially find a very accurate data split. It addresses selection bias and assists with the overfitting of data. By dividing the dataset into a training and validation set, we can check that our model performs well on data seen or

not seen during training. Without cross-validation, we would not be able to establish whether our model behaves well using any data.

In our study, the K-Fold Cross-Validation is tested on the more complex SVM-K model to assess the robustness of the findings. Since the chosen training-test split was 80:20, 5-fold cross-validation was chosen (Qian & Rasheed 2007). This means that the data is split into an 80:20 division through five iterations and on each iteration, a new set of data is chosen as the test set. The five results are then averaged to provide a single performance metric[7].

3.4 Variable Importance Analysis (VIA)

Measuring the importance of a variable (feature) in a machine learning model is essential (Wei et al. 2015). The analysis of which variables have the most significant impact on the prediction of stock movements can help with the understanding of the significance of certain features and could point out patterns to be used by individuals or institutional investors.

The method we use is called Variable Importance Analysis (VIA) and is a technique that illustrates *which features (variables) play an important role in the prediction of a machine learning model*. In this case, VIA will demonstrate the importance of sentiment and the different emotions, in the prediction of stock market movements.

4. Data

4.1 Twitter Data

Twitter data in the form of extracted Tweets, can be obtained in several ways. However, due to recent changes in the 'terms of use policy' of the Twitter platform, existing Twitter datasets are no longer publicly available; this policy also extends to the distribution of datasets for academic purposes (Watters 2011).

The Twitter API does offer several subscription packages for Twitter data extraction, ranging from a free standard-user API to an enterprise API. Given the financial constraints in selecting individual Twitter API subscriptions, the standard-user API was selected for this study. The standard-user API allows 180 search requests per 15-minute window using the user authentication and 450 requests per 15-minute window through the application authenticator. To adhere to the constraints of the standard-user API, the algorithm was adjusted to extract as many Tweets as possible, while remaining within the request limit.

The standard-user API only allows Twitter data extraction for a period of up to 7 days before the date of implementing the request. Such an iterative procedure was followed to extract the Twitter data. The algorithm was set to extract a minimum of 12 000 daily Tweets, which was the optimal amount of

[7] This allows for 80 per cent training and 20 per cent testing of the data. Ultimately, a fifth of the data is used for validation and the remaining four-fifths of the data are used for training and testing the data, swapping one fifth out with every iteration.

Tweet requests, on average, that could be requested without exceeding the rate limits, it being rate-limited. This is important, as exceeding the rate limit could result in the Twitter developer account being blacklisted from using the standard-user API. The data extracted did not always achieve the daily target of 12 000 Tweets, as the Twitter API is dependent on factors such as connectivity and network traffic volume. On average, 9 191 Tweets were extracted daily, over almost a year (340 days). The total number of Tweets extracted are close to 3 million[8] over the period.[9]

To extract a Tweet, a general keyword, based on relevance and popularity, namely '*stock market*' was selected. A preliminary examination was performed to assess whether this keyword resulted in Tweets with excessive noise (in terms of having irrelevant Tweets in the data, not specifically related to stock markets). Fortunately, the results showed that the Tweets contained little noise. The Twitter data extracted include a time stamp based on the Coordinated Universal Time (UTC) of each Tweet. We used an algorithm to adjust the UTC time to the time-zone of each individual market in the study. To derive hourly data, we aggregated the sentiment or emotion scores per hour.

Using R-programming, the *'twitteR'* R-package was used to extract the Twitter data. The '*twitteR*' code was set to extract and include Tweets based on a mixture of popularity (user influence and trending Tweets) and ordinary Tweets, not categorised as popular (Gentry 2016). This package option assists with our contribution in capturing the effects of viral news feeds and social media user-influence (see Gholampour (2019), which similarly analysis high followers Twitter accounts) analysing Tweets . Daily expectations of returns index. *Journal of Empirical Finance*, *54*, pp.236-252.. Once we extracted the relevant Tweets and cleaned the data, we used the 'syuzhet' package and 'syuzhet' lexicon to determine the sentiment of each Tweet, and the 'NRC lexicon' to determine the emotions of each Tweet.

Table 1 shows Tweets extracted from Twitter for a particular day. The results from a standard lexicon are contrasted against the 'syuzhet' lexicon. The standard lexicon, in this case, the Hu and Liu lexicon, classifies the Tweets in a simplistic discrete manner, whereas the 'syuzhet' lexicon continuously weighs the text.

The advantage of the 'syuzhet' lexicon for sentiment analysis, included in the 'syuzhet' package, can be seen in Table 1, the frequency and choice of words are weighed separately, unlike the standard lexicons. In Table 1, the first Tweet states: "Technical Damage After Trump Threatens Higher Tariffs - $S&P500", which intuitively can be interpreted as a negative sentiment. However, the standard lexicon incorrectly declares the Tweet as a neutral statement, giving it a score of zero, whereas the 'syuzhet' lexicon gives it a score of -1.35. This illustrates the dominance the 'syuzhet' lexicon has over other lexicons by utilising independent weighting and frequency techniques.

[8] Exact number of Tweets extracted are 2 996 295

[9] The analyses were done in two phases: in January 2020, all models were run, to get preliminary results. However, we continued to extract Tweets and ran all models a second time on Tweets extracted up to 6 March 2020. These are the results reported in the paper.

Table 1 Lexicon Comparison

Tweet	Standard Lexicon Score	Syuzhet Lexicon Score
"Technical Damage After Trump Threatens Higher Tariffs-$S&P500."	0	-1.35
"Double STORM cell signals for SP500 persist while US bond yields and inflation gauges continue to decline."	-1	-0.95
"Stocks appear to be on a decent footing for the next few months. The S&P500 may decline in May."	0	-0.50
"Mini Speculators trim their bullish bets this week."	1	-0.80
"Positive market outlook. #SP500."	1	0.75

Source: Authors' results

Furthermore, we use the *NRC Lexicon* in the 'syuzhet' package developed by Saif Mohammad (Mohammed et al. 2013) which returns the emotion scores for each Tweet extracted. The *NRC Lexicon* is a list of English words and their associations with eight basic emotions; anger, fear, anticipation, trust, surprise, sadness, joy and disgust as well as two sentiments; negative and positive. The annotations were done manually by crowd-sourcing. The NRC function returns a data frame in which each row represents a Tweet from the original file. The columns include each emotion type, as well as the positive or negative sentiment value as per the example below.

If we take the Tweet "I love dogs; they are such good companions"; then the NRC function will return the following data frame (table 2):

Table 2 An example of an NRC data frame

anger	fear	anticipation	trust	surprise	sadness	joy	disgust	positive	negative
0	0	0	0	0	0	2	0	2	0

Source: Authors' results

For each of the almost 3 million Tweets extracted, we used the 'syuzhet' package to derive sentiment scores as well as emotion scores (as explained above). Therefore, we created a dataset of sentiment and emotions scores to be used in the models to predict market movement in the eight markets under investigation.

4.2 Financial Data

The financial data collected in this study consist of eight financial indices for eight countries and were selected based on high-frequency intraday data availability and the Twitter userbase of each country. The countries, as well as the respective stock market indices included in the study, are shown in Table 3.

Table 3 Countries and relevant stock indices

Country	Stock Exchange	Market Index
France	Euronext Paris	CAC 40
Germany	Frankfurt Stock Exchange	DAX
India	National Stock Exchange of India	NIFTY 50
Japan	Tokyo Stock Exchange	Nikkei 225
Poland	Warsaw Stock Exchange	WIG20
Spain	Madrid Stock Exchange General Index	IBEX 35
UK	London Stock Exchange	FTSE 100
USA	NYSE/NASDAQ/CBOE	S&P 500
Total		***8 Indices***

Source: Authors' compilation

Based on the 2019 Morgan Stanley Capital International market classification review (MSCI 2019), India and Poland are classified as emerging markets, whereas the other markets are developed markets. The combination, using developed and emerging markets, allows for a holistic view on the impact that investor sentiment analysis derived from social media at a global level has on stock market performance.

Furthermore, the use of both emerging and developed markets could illustrate how different markets react to investor sentiment derived from social media. The assumption is that different markets might be more or less sensitive to investment sentiment and thus react differently to the same information derived from global Tweets. The time period of response can also differ between markets with some markets reacting immediately, while others might have a delayed response or no response at all.

The financial data was obtained through the Swiss Banking Group Dukascopy, for the period 01-04-2019 to 06-03-2020. As mentioned previously, we focused on high-frequency financial and sentiment data, rather than daily data, due to the continuous trading in markets and the increased use of high-frequency algorithmic trading. Previously high-frequency data was not easily accessible, but with technical developments, accessibility has improved. We also obtained daily data and repeated all models using the daily financial, sentiment and emotion data as a robustness check of our results for high-frequency data[10]. We need to highlight that high-frequency data likely include a higher percentage of noise than daily data. Tweets with excessive noise, for example the number of Tweets that are not relevant to the analyses of stock markets, as a percentage of total extracted Tweets per hour, might be higher during certain hours analysed compared to daily data. Nonetheless, it is essential that analyses of intraday data should be performed, as high frequency information is a necessity when trading.

[10] All daily data results are available on request

Finally, a simple algorithm is implemented to process the stock returns into a positive or a negative market movement. The positive stock returns are classified as 'up' and coded as one, whereas the negative returns are classified as 'down' and coded as zero. Since the market filtering process only considers data during market trading-hours, the effect of a 'no price change' is not included (the inclusion of this would nonetheless have been negligible).

After the above-mentioned manipulation of data, the data were now suitable for analyses using machine learning models.

5. Results

In this section, we report on the results from i) the evaluation metrics for each of the algorithms used, ii) the K-Fold Cross-Validation technique and iii) the Variable Importance Analysis (VIA) using high frequency intraday data. All analyses were also repeated using daily data, as a robustness test. These results are not reported, though all results are available from the authors.

5.1 Evaluation Metrics

Tables 4-6 show the evaluation metrics for each of the classification algorithms used, namely NB, K-NN and SVM-K for each of the eight stock markets under analysis. Subsequently, each evaluation metric, namely accuracy, recall, precision and F1-Score, are reported. Each metric highlights an important finding of the ability to predict stock market movements. We distinguish between two models, including i) only sentiment scores, and ii) sentiment and emotions. We firstly discuss the results relating to developed markets and secondly the findings associated with emerging markets.

5.1.1 Developed markets

As mentioned earlier, we analyse six developed markets. Here, we only report and discuss the USA results, since all the developed markets show very similar results. The reader can compare the USA results to that of the other developed markets (see Appendix B Tables B1-B5 for the results for the UK, France, Germany, Japan and Spain).

If we consider the movement of the stock markets in the USA using the S&P500 against the sentiment scores (the first model, see Table 4), the model shows an accuracy of 55.73 per cent (NB), 59.21 per cent (K-NN) and 53.55 per cent (SVM-K). This implies that considering all three the machine learning models at least more than 53 per cent of the unseen test-set data is predicted correctly. An accuracy measure of more than 53 per cent is significant since it conforms to the efficient market hypothesis that stock prices are predictable with an accuracy greater than 50 per cent, as highlighted by Qian and Rasheed (2007).

Furthermore, the recall measure, which indicates the proportion of correctly classified positives, as a proportion of all positives (true positives), yields 89.43 per cent (NB), 70.19 per cent (K-NN) and 92.77 (SVM-K) per cent, which are notable results (see Table 4). In terms of the current study, the

recall measure indicates the proportion of correctly predicted positive returns to the total number of actual positive returns on the stock market. The higher the recall percentage, the lower the number of false negatives that occur. A false negative is when the prediction indicates a negative return on the stock, while the return was actually positive. In terms of investing, a false negative is a far more concerning metric than a false positive. It is of the utmost importance that false negative predications should be minimised since, theoretically, there is no upper limit to the number of losses suffered by incorrectly taking a short position, which is a position to sell stock (Bank 2019). Thus, if a prediction model gives a high false negative rate (FNR), and the model is used to inform investment decisions, it can lead to huge losses.

The precision metric, which is defined as the total number of true positives (positive returns) to the total number of predicted positives (predicted positive returns), is a good measure to determine whether the costs of false positives (incorrectly predicted positive returns) are high. The precision metric is 55.14 per cent (NB), 59.03 per cent (K-NN) and 53.68 per cent (SVM-K). The model predicts on average 55 per cent or more true positives (correct positive returns). This is a reasonable result as "up movements" in the stock market can be predicted correctly 55 out of 100 times, implying that investment returns should be positive on average.

Table 4 USA Evaluation Metrics

Measure	**Sentiment (model 1)**	**Sentiment & Emotion (model 2)**
Naïve Bayes (NB)		
Accuracy	55.73	50.76
Recall	89.42	55.08
Precision	55.14	53.06
F1-Score	68.67	53.79
K-Nearest Neighbours (K-NN)		
Accuracy	59.21	57.22
Recall	70.19	69.23
Precision	59.03	57.38
F1-Score	64.93	63.54
Support Vector Machine - Kernel (SVM-K)		
Accuracy	53.55	53.55
Recall	92.77	62.65
Precision	53.68	54.53
F1-Score	68.38	59.14

Source: Authors' calculations

The next evaluation metric, namely the F1-Score, has the benefit that it considers both the recall and precision values, which separately, are at times difficult to interpret. The F1-Score is the harmonic mean of the two metrics. If the F1-Score is 100 per cent, it indicates perfect precision accuracy, whereas if it is 0 per cent, it shows the worst possible prediction accuracy. The F1-Score for the USA prediction model is 68.67 per cent (NB), 64.93 per cent (K-NN) and 68.38 per cent (SVM-K). These results are significant, as it shows more than two-thirds accuracy between the precision and recall metrics. This indicates that intraday frequency stock returns on the US market can present good recall and precision metrics.

Considering the second model in Table 4, that includes the movement of the stock markets in the USA (S&P500) against the sentiment scores and the eight emotion scores, we find the evaluation metrics not as notable as when only the sentiment scores are considered. The inclusion of emotion seems to weaken the predictive power of the results. This is true considering all the evaluation metrics related to all three the different machine learning models. Specifically the recall metric's performance is much worse than in model 1. This indicates that a higher rate of false negatives is predicted and, as previously explained, a false negative prediction can lead to significant losses on stock markets. The F1-Score of the second model is also considerably lower than in the first model, looking at the SVM-K model, the F1-Score is 59.14 per cent for model 2, compared to 68.38 per cent for model 1, thus an almost 9 per cent decrease in precision accuracy.

For the results on the rest of the developed markets, see Appendix B, Tables 10 to 14. The reader will note that for the UK (FTSE 100), France (CAC 40), Germany (DAX), Spain (IBEX 35) and Japan (Nikkei 225), the results are similar to that of the USA. The results indicate that high-frequency intraday sentiment data can successfully be used to predict stock market movements. However, we find that sentiment with emotions can weaken the predictive power of the models. However, at least two out of the three machine learning models render F1-Scores of above 50 per cent.

Therefore, we conclude that sentiment scores derived from Twitter can successfully predict high frequency stock market movements in developed countries. These findings are robust, as the prediction ability of sentiment is tested using three different machine learning models (NB, K-NN and SVM-K) and using an array of evaluation metrics.

Using the sentiment and emotion model is also an option in the prediction of stock market movements, though its performance is weaker than when only sentiment is considered. In saying that, the sentiment and emotion model still produces predictability of above 50 per cent for at least two out of the three machine learning models.

5.1.2 Emerging markets

In our analysis we include two emerging markets, *India* (NIFTY) and *Poland*[11] (WIG 20) to determine if the predictability of market movement, using sentiment and sentiment and emotion, renders similar results as in the developed countries. India is discussed first, as this is a compelling case, with India having the seventh-highest active Twitter user base, numbering 7.75 million users in the world (Statista 2019) (see Table 5).

Table 5 India Evaluation metrics

Measure	**Sentiment (model 1)**	**Sentiment & Emotion (model 2)**
Naïve Bayes (NB)		
Accuracy	54.91	55.92
Recall	47.82	52.02
Precision	52.00	46.79
F1-Score	51.37	59.94
K-Nearest Neighbours (K-NN)		
Accuracy	55.97	56.50
Recall	45.82	32.34
Precision	53.32	54.94
F1-Score	50.27	40.57
Support Vector Machine - Kernel (SVM-K)		
Accuracy	54.91	58.08
Recall	28.97	30.97
Precision	58.15	59.14
F1-Score	36.53	38.64

Source: Authors' calculations

If we consider the movement of the stock markets in *India* using the NIFTY against the sentiment scores (model 1 in Table 5), the model shows an accuracy of 54.91 per cent (NB), 55.97 (K-NN) per cent and 54.91 per cent (SVM-K), which is above the 50 per cent threshold. This implies, considering the results of all three the machine learning algorithms, that at least 54 per cent of the predictions out of the total number of possible predictions, were classified correctly. These results are similar to that of developed markets, as discussed in section (5.1.1). However, if we consider the other evaluation metrics we notice that the recall and the F1-score in some instances do not perform as well as in predicting market movement in the developed markets (section 5.1.1). The lower recall percentage, indicates a higher number of false negatives. As previously explained, in terms of investment, false negatives are a concerning metric, as it can increase losses suffered on incorrectly taking a short position on a stock. Therefore, if investors consider these models to predict market movement, they

[11] Note that in certain classifications Poland has recently been reclassified as an emerging market, although MSCI (2019) still classifies it as an emerging market.

need to stay alert to this finding, to minimise losses. The poorer performance of the recall measure also negatively influences the performance of the F1-score. The lower recall and F1-Score, might reflect noise in the model; thus, if a better noise reduction model could be provided, over perhaps an extended period of more than a year, the recall and F1-Score should improve markedly.

However, what is interesting is that if we consider the sentiment and emotion model (model 2 in Table 5), we find the accuracy of the machine learning models more significant than if we only consider sentiment by itself, namely accuracy levels of 55.92 per cent (NB), 56.5 per cent (K-NN) and 58.08 per cent (SVM-K). It seems that in India, an emerging market, a model that includes emotions and sentiment achieves higher predictive power than a model that only relies on sentiment, this is in contrast to developed countries where results show that including emotion weakens the models

These findings are significant, suggesting that the use of Twitter data can predict stock market movements using investor sentiment, or sentiment and emotion. It also seems as if emerging markets have better inherent performance accuracy if sentiment and emotion are included in the models than developed markets. To substantiate this idea, another emerging market, Poland, is discussed.

Similarly, if we consider the movement of the stock markets in *Poland* against the sentiment scores (model 1 in Table 6), the model shows a prediction accuracy of 57.29 per cent (NB), 52.48 (K-NN) per cent and 54.40 per cent (SVM-K), which is significant as it is above the threshold of 50 per cent. This finding is in line with the developed markets and the findings for India.

If we consider the market movement, against the sentiment and emotion model, we obtain accuracy measures of 43.35 per cent (NB), 52.96 per cent (K-NN) and 53.92 per cent (SVM-K); thus, two out of the three machine learning algorithms are above the threshold of 50 per cent. These findings are similar to those of developed markets, showing that if emotion is included in the model, it might weaken the accuracy of the model, though the gap between the results of the two models is not as severe for Poland, as it is in developed markets (see section 5.1.1).

Once again we notice the weaker performance of the recall and the F1-score, similar to the results analysing the Indian market, which indicates that special attention should be given to these measures when analysing emerging markets.

More research is needed on the predictability of high frequency sentiment data on the movement of stock markets in emerging markets, before these models can be fully accepted. However, the models were significant regarding accuracy and precision evaluation metrics, which is an indication that these models have the potential to be significant predictors of stock market movements if additional research is undertaken.

Table 6 Poland Evaluation metrics

Measure	Sentiment (model 1)	Sentiment & Emotion (model 2)
Naïve Bayes (NB)		
Accuracy	57.29	43.35
Recall	67.26	69.02
Precision	59.14	42.91
F1-Score	50.81	52.81
K-Nearest Neighbours (K-NN)		
Accuracy	52.48	52.96
Recall	38.17	39.23
Precision	46.16	46.87
F1-Score	41.77	42.70
Support Vector Machine - Kernel (SVM-K)		
Accuracy	54.40	53.92
Recall	28.38	30.51
Precision	37.29	38.36
F1-Score	32.81	35.79

Source: Authors' calculations.

Emerging markets are becoming the catalyst of global economic growth, providing over 40 per cent of global GDP, and for the period 2009 to 2014, the top companies from emerging market economies experienced more than double the growth rates of the top companies from developed market economies (Renoult 2019). Factors likely responsible for the aforementioned are higher growth potential and higher risk premium requirements from emerging markets, due to adverse socio-economic conditions typically existing in emerging economies (Renoult 2019).

The results for India and Poland highlight a significant finding. Stock market movements for India and Poland can be successfully predicted using Twitter data. This is the case for both the sentiment only and the sentiment with emotions model. The models for both emerging countries yield good predictive accuracy.

5.2 Robustness Test: K-Fold Cross-Validation

As discussed in section 3.3, K-fold cross-validation is a model validation technique that validates the performance results of a machine learning model. In particular, it gives us a comprehensive measure of our model's performance throughout the *whole dataset.* The method is trained and validated multiple times and is more thorough than the previous evaluation methods, as it can potentially find a data split that is very accurate. It addresses selection bias and assists with the overfitting of data. By dividing the dataset into a training and validation set, we can concretely check that our model performs well on data seen or not seen during training. Without cross-validation, we would not be able to establish if our model behaves well using any data. We use the K-fold cross-validation test on the

more complex SVM-K model to assess the robustness of our findings. Since the chosen training-test split was 80:20, 5-fold cross-validation was chosen. We present the results in Table 7.

Using the USA as an example, Table 7 shows that, initially, the USA results were biased. The cross-validation technique yields an accuracy score of 52.60 per cent. This is marginally lower than the initial results of 53. 55 (see Table 4 for the evaluation metric "accuracy" under the SVM-K model). However, this finding is unsurprising, as the USA has the world's most active Twitter userbase and is the world's leading economy (World Bank 2019). Therefore the potential choice of the split of the data might lead to somewhat biased results. This result, of prediction accuracy, is similar to previous studies conducted on the USA stock markets in which it was also found that the choice of the training-test split of the data might lead to biased results (Ruan et al. 2018).

With regard to the sentiment with emotion model, the results yield a 52.34 per cent predictive accuracy compared to 53.55 per cent (see Table 4 for the accuracy evaluation metric under the SVM-K model), which once again highlights the marginal bias in the model. However, with the models being validated on different test sets, we can now report with confidence that the models are robust and give accurate predictions of the market movements using sentiment or sentiment and emotion. This finding suggests that sentiment and sentiment with emotions derived from Tweets are valid predictors of stock market movements in the USA.

This finding holds for the other stock markets (see Table 7), using sentiment and sentiment with emotions models, for the UK (the first percentage shown is for sentiment only and the second for the sentiment with emotion) (52.99 52.01), Germany (55.60; 55.25), France (55.05; 53.67), Spain (53.94; 55.22), Japan (50.57; 50.95) (compared to the accuracy evaluation metrics under the SVM-K models in Tables 10-14 in Appendix B), Poland (55.38; 55.38) and India (54.48; 53.32), (compared to the accuracy evaluation metrics under the SVM-K models in Table 5 for India and Table 6 for Poland), all of the metrics using the K-Fold Cross-Validation method highlight minor biases in the estimated models, though these biases were corrected with the validation.

Table 7 Robustness Test: K-Fold Cross-Validation of SVM-K

Performance Accuracy		
Country	**Sentiment (model 1)**	**Emotion & Sentiment (model 2)**
USA	54.80	54.34
UK	52.99	52.01
Germany	55.60	55.25
Poland	55.38	55.38
France	55.05	53.67
Spain	53.94	55.22
India	54.48	53.32
Japan	50.57	50.95

Source: Authors' calculations.

Thus, if either sentiment or sentiment and emotions are included in the predication of market movement we find that, with the correction of the previous minor biased evaluation metrics, all models in all the markets under investigation, using high-frequency data, are significant in predicting stock market movements (see Table 7).

What is most notable, is the reconfirmation of the significance of Poland and India – the two emerging markets in the set of countries. Thus not only can sentiment and sentiment with emotion be used to predict stock market movements in developed countries, but these models are also applicable to emerging markets.

5.3 Variable Importance Analysis (VIA)

As discussed in section 3.4, VIA illustrates the importance of sentiment and the eight different emotions included in the prediction of stock market movements. Firstly, when analysing sentiment only in terms of stock movement prediction it provides good prediction accuracy, showing that sentiment is an important variable. Thus we can accept that based on VIA sentiment is a significant factor.

However, when analysing the significance of sentiment with the eight emotions, the results highlight that not only sentiment is an accurate predictor of stock market movements, but certain emotions also play important roles. Therefore, in the next section it is important to highlight which emotions in each specific market are most significant in predicting stock market movements.

In the USA the emotions that are most significant in predicting stock market movements are 'joy', 'trust' and 'anticipation'. Thus, if these variables are used in the prediction of the model, it is most likely that 'joy', 'trust' and 'positive anticipation' will predict the upward movement of the market. The same holds for the other markets, for example in the UK the emotions 'anger', 'disgust' and 'fear' are the most significant indicators of market movements, with 'anger', 'disgust' and 'fear' predicting downward movements in the market. Similarly, different emotions are found to be most significant in different markets under evaluation. This implies that if these emotions are detected within a Tweet, significant prediction accuracy can be obtained.

Table 8 Most significant VIA Emotions in different markets

Most significant emotions							
USA	**UK**	**Germany**	**Poland**	**France**	**Spain**	**India**	**Japan**
Joy	Anger	Trust	Anger	Joy	Sadness	Trust	Disgust
Trust	Disgust	Fear	Trust	Sadness	Trust	Joy	Fear
Anticipation	Surprise	Anger	Fear	Trust	Disgust	Fear	Anticipation

Source: Authors' calculations.

Table 8 reveals that the emotions found most often to be significant in predicting market movements, in the current study are, 'fear' and 'trust'. These emotions also frequently contribute to market volatility.

Often investors can succumb to the emotion 'fear', as was recently seen with the announcement of the Corona virus (COVID-19), and previously also the Asian crisis, the 'tech bubble' and the financial crisis of 2008, which had severe negative effects on financial markets. The negative effects are due to investors' fears of losses, which result in the selling of stocks. The selling of stocks further contributes to price decreases, and thus the 'fear' of losses is realised. Therefore, if the emotion 'fear' is found to be a significant predictor of market movements, it is an indicator that markets will likely be moving downwards.

For investors to buy stocks, the emotion 'trust' (and positive anticipation, which could also have a positive effect), among others, is important. Firstly, they need to trust the performance of financial markets in general, and secondly to purchase stock they need to 'trust' the specific stock and believe that it will offer positive returns. Therefore, if the emotion 'trust' is revealed as an emotion of a Tweet, it is likely a predictor of an upward market movement.

These results are similar to the findings on sentiment and stock markets, showing a positive relationship in general (Brown & Cliff 2004), as well as a positive relationship for specifically sentiment derived from Tweets (Zhang et al. 2010 and Li et al. 2014).

5.4 Robustness check using daily data rather than intraday data

To test the robustness of our results, we repeated all analyses using daily data. We found the results to be very similar to those discussed in section 5.1 to 5.3, with the difference that, in most instances, the evaluation metrics and the K-Fold Cross-Validation test revealed somewhat better levels of performance across all models ('sentiment' and 'sentiment and emotion') and all machine learning algorithms. For example, for the USA, the sentiment model showed an accuracy of 60 per cent using daily data, compared to 55.73 per cent using intraday data (NB), and using the t K-Fold Cross-Validation of the SVM-K model it showed an accuracy of 59.58 per cent using daily data, compared to 54.80 using intraday data. The main reason for this is the higher frequency of Tweets per day, compared to Tweets per hour and a reduced level of 'noise' in daily data compared to high frequency intraday data. As previously mentioned, these results are not reported in the paper, as the main focus and contribution of the current research is analyses of high frequency data, though all results on daily data are available on request.

6. Conclusion and Recommendations

This study investigated whether text data, extracted from the social media platform Twitter, and analysed to determine the *sentiment and emotions*, predict stock market movements.

Previous studies have investigated the likelihood of sentiment derived from Twitter to predict stock market movements, though those papers mostly used daily data and not high frequency intraday data,

which is a necessity, considering that trading occurs throughout the day and not only once a day. Very few of these papers investigated more than one market and the majority of papers analysed stock markets in developed countries only (Bollen et al. 2011, Zhang et al. 2010). Not one of the previous studies compared the results of models in developed and emerging markets, to establish if the models are significant in both types of markets. Previous papers used text data or Tweets extracted for relatively short time spans, while the current study includes Tweets over a time span of almost a year. In previous investigations either emotions, or sentiment of investors, were analysed to predict stock market movements (see Table A1). Not one of them, to our knowledge, considered both. Previous studies only used correlation analysis, basic econometric models or a single machine learning technique in their analyses (Ruan et al. 2018 & Zhang et al. 2010). In addition, studies did not consider a range of evaluation metrics or complete any robustness tests on their machine learning algorithms (see Table A1).

In the current study we addressed these shortcomings and contributed to the literature as follows: To derive the sentiment and emotions of Tweets, we employed an artificial intelligence supervised machine learning approach, with the use of 'syuzhet', developed by Jockers (2017), which includes the 'syuzhet' lexicon to analyse sentiment - and the 'NRC' lexicon to analyse emotions. We analysed high frequency intraday data, both for sentiment and emotion and for financial data, obtained through the Swiss Banking Group Dukascopy. In out analyses we investigated eight markets, six developed markets and two emerging markets. We applied three machine learning models, namely Naïve Bayes, K-Nearest Neighbours and the Support Vector Machine algorithm. Additionally, we used evaluation metrics; the Precision, Recall, Specificity and F-1 Score to evaluate the results of these algorithms. Lastly, we used the K-Fold Cross-Validation technique as a robustness check of the performance of our machine learning models and the Variable Importance Analysis (VIA) to show which emotions played an important role in the prediction of stock market movements.

Our findings suggest that a keyword like 'stock market' can be used to accurately predict and explain movements of stock markets in developed and emerging markets, with similar prediction accuracy shown in these markets. The exception was that the recall, and F1-score evaluation metrics performed slightly weaker in the emerging markets than in the developed markets, which might indicate more 'noise' in these datasets. Thus, our findings suggest that sentiment and emotions derived from Tweets are significant predictors of stock market movements, not only for a single market but for multiple markets in developed and emerging markets. We find an intraday accuracy measure above 50 per cent for all markets, using any of the three machine learning algorithms (except for Japan in which only two out of the three algorithms are above 50%), which is an acceptable level to predict stock market movements.

These results are important for portfolio planning. Also, it emphasises the need for careful consideration of the social media as a vehicle to derive investors' sentiment towards a specific stock

or stock market as a whole. Investors can exploit this knowledge to achieve financial gains through better-informed decision-making ability.

The fact that these models show that sentiment and emotion can accurately predict market movements in emerging markets and not only developed markets, is of interest, as the risks and returns in emerging markets are higher than in developed markets. Therefore, more information can increase trust in these markets.

A recommendation for further studies is the need to focus purely on emerging markets and high frequency data, since our research reveals the potential financial gains in these markets. However, more studies on different emerging markets are needed.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Abbes, H. (2015). *Tweets Sentiment and their Impact*. S.l.: Liege University.

Ageitos, E. C. (2018). *Experiment on sentiment analysis over LSE (London Stock Exchange) Twitter data.* Trabajo de Fin de Grado Escuela de Ingeniería de Telecomunicación Grado en Ingeniería de Tecnologías de Telecomunicación.

Antweiler, W. & Frank, M. Z. (2004). Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance*, 59(3): 1259-1294.

Bank, E. (2019). *Long vs. Short Stocks*. Available at: https://finance.zacks.com/long-vs-short-stocks-5351.html

Bhardwaj, A., Narayan, Y., Pawana, V. & Dutta, M. (2015). Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty. *Procedia Computer Science*, 70: 85-91.

Bodie, Z., Kane, A. & Marcus, A. J. (2014). *Investments*. 10th ed. New York: Mc Graw Hill Education.

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

Bonga-Bonga, L. & Mwamba, J. M. W. (2011). The predictability of stock market returns in South Africa: Parametric vs non-parametric methods. *South African Journal of Economics*, 79(3): 301-311.

Broadstock, D. & Zhang, D. (2019). Social-media and intraday stock returns: The pricing power of sentiment. *Finance Research Letters*, 30: 116-123.

Brown, G. W. & Cliff, M. T. (2004). Investor sentiment and the near-term stock market. *Journal of empirical finance*, *11*(1), 1-27.

Bukovina, J. (2016). Social media big data and capital markets — An overview. *Journal of Behavioral and Experimental Finance*, 11(C): 18-26.

Chung, S. & Liu, S. (2011). *Predicting Stock Market Fluctuations from Twitter*. s.l.: s.n.

Cropper, A. (2011). *Modelling Stock Volume Using Twitter*. London: Oxford University.

Das, S. R. & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9): 1375-1388.

Da, Z., Engelberg, J. & Gao, P. (2015). The Sum of All FEARS: Investor Sentiment and Asset Prices. *Review of Financial Studies*, 28(1): 1-32.

Elodie, M. (2019). *Twitter and its relationship with returns and trading volume of European stocks*. Louvain School of Management, Université Catholique de Louvain, 2019. Prom.: D'Hondt, Catherine; Desagre, Christophe. http://hdl.handle.net/2078.1/thesis:20734

Gentry, J. (2016). Package 'twitteR'. Available at: https://cran.r-project.org/web/packages/twitteR/twitteR.pdf

Gholampour, V. (2019). Daily expectations of returns index. *Journal of Empirical Finance*, *54*:236-252

Hall, P., Park, B. U. & Samworth, R. J. (2008). Choice of Neighbor Order in Nearest-Neighbor Classification. *The Annals of Statistics*, 36(5): 2135-2152.

Hu, M. & Liu, B. (2004). *Mining and Summarizing Customer Reviews*. Seattle, Washington, USA, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

Jadhav, R. & Wakode, M. S. (2017). Survey: Sentiment Analysis of Twitter Data for Stock Market Prediction. *International Journal of Advanced Research in Computer and Communication Engineering*, 6(3): 507-509.

Jagdale, R. S., Shirsat, V. S. & Deshmukh, S. N. (2016). Sentiment Analysis of Events from Twitter Using Open Source Tool. *International Journal of Computer Science and Mobile Computing*, 5(4): 475-485.

Jockers, M. (2017). Introduction to the Syuzhet Package. Available at: https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html

Kolagani, S. H. D., Negahban, A. & Witt, C. (2017). Identifying trending sentiments in the 2016 us presidential election: A case study of twitter analytics. *Issues in Information Systems*, 18(2): 80-86.

Li, Q., Wang, T., Li, P., Liu, L., Gong, Q. & Chen, Y. (2014). The effect of news and public mood on stock movements. *Information Sciences*, 278: 826-840.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1): 1-167.

Maqsood, H., Mehmood, I., Maqsood, M., Yasir, M., Afzal, S., Aadil, F., Selim, M. M. & Muhammad, K. (2020). A local and global event sentiment-based efficient stock exchange forecasting using deep learning. *International Journal of Information Management*, 50: 432-451.

Maree, S. & Johnston, K. (2015). Critical Insights into the Design of Big Data Analytics Research: How Twitter 'Moods' Predict Stock Exchange Index Movement. *The African Journal of Information and Communication*, 15: 53-67.

Mc Kay, D. (2018). *Investigating the Effect of Sentiment in High-Frequency Financial Markets*. Dublin: The University of Dublin.

Mohsin, M. (2019). *10 Social Media Statistics You Need to Know in 202*0. Available at: https://www.oberlo.com/blog/social-media-marketing-statistics

MSCI. (2019). *Results of the MSCI 2019 global market accessibility review*. Available at: https://www.msci.com/market-classification

Moritz, B. (2018). *Applications of Textual Analysis and Machine Learning in Asset Pricing*. Dissertation an der Fakult¨at fu¨r Mathematik, Informatik und Statistik der Ludwig–Maximilians–Universita¨t Mu¨nchen.

Mohammad, S., Kiritchenko, S. & Zhu, X. (2013). *In Proceedings of the seventh international workshop on Semantic Evaluation Exercises*, June 2013, Atlanta, USA.

Naldi, M. (2019). *A review of sentiment computation methods with R packages*. Available at: https://arxiv.org/abs/1901.08319

Nisar, T. M. & Yeung, M. (2018). Twitter as a tool for forecasting stock market movements: A short-window event study. *The Journal of Finance and Data Science*, 4: 101-119.

Oliviera, N., Cortez, P & Areal, N. (2017). The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices. *Expert Systems with Applications*, 73: 125-144.

Qian, B. & Rasheed, K. (2007). Stock market prediction with multiple classifiers. *Applied Intelligence*, 26(1): 25-33.

Rao, T. & Srivastava, S. (2012). *Analyzing Stock Market Movements Using Twitter Sentiment Analysis.* Delhi, International Conference on Advances in Social Networks Analysis and Mining.

Renault, T. (2017). Intraday online investor sentiment and return patterns in the U.S. stock market. *Journal of Banking & Finance*, 84: 25-40.

Rennie, J. D. M., Shih, L., Teevan, J. & Karger, D. R. (2003). *Tackling the Poor Assumptions of Naive Bayes Text Classifiers*. Washington DC, Massachusetts Institute of Technology.

Renoult, S. (2019). *Doom and Gloom in the Developed Economies: Time to Invest in Emerging Markets*. Available at: https://riskmagazine.nl/article/2019-11-26-doom-and-gloom-in-the-developed-economies-time-to-invest-in-emerging-markets

Ruan, Y., Durresi, A. & Alfantoukh, L. (2018). Using Twitter trust network for stock market analysis. *Knowledge-Based Systems*, pp. 1-12.

Sanjay, M. (2018). *Why and how to Cross-Validate a Model?* Available at: https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f

Shen, D., Liu, L., & Zhang, Y. (2018). Quantifying the cross-sectional relationship between online sentiment and the skewness of stock returns. *Physica A*, 490: 928-934.

Smith, T. (2019). *Market Sentiment.* Available at: https://www.investopedia.com/terms/m/marketsentiment.asp

Statista. (2019). *Leading countries based on number of Twitter users as of July 2019 (in millions).* Available at: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

Tabari, N., Biswas, P., Praneeth, B., Seyeditabari, A., Hadzikadic, M. & Zadrozny. W. (2018). *Causality Analysis of Twitter Sentiments and Stock Market Returns*. Melbourne, Association for Computational Linguistics, pp. 11-19.

Valdivia, A., Luz´on, M. V. & Herrera, F. (2017). Sentiment analysis in TripAdvisor. *IEEE Intelligent Systems*, 32(4):72–77.

Watters, A. (2011). *How Recent Changes to Twitter's Terms of Service Might Hurt Academic Research.* Available at: https://readwrite.com/2011/03/03/how_recent_changes_to_twitters_terms_of_service_mi/

Wei, P., Lu, Z. & Song, J. (2015). Variable importance analysis: A comprehensive review. *Reliability Engineering and System Safety*, 142(1): 399-432.

World Bank. (2019). GDP (current US$). Available at: https://data.worldbank.org/indicator/ny.gdp.mktp.cd?view=map

Yang, S. Y., Mo, S. Y. K. & Liu, A. (2015). Twitter financial community sentiment and its predictive relationship to stock market movement. *Quantitative Finance*, 15(10): 1637-1656.

You, W., Guo, Y. & Peng, C. (2017). Twitter's daily happiness sentiment and the predictability of stock returns. *Finance Research Letters*, 23 (C): 58-64.

Zehe, A., Becker, M., Hettinger, L., Hotho, A., Reger, I. & Jannidis, F. (2016). *Prediction of happy endings in German novels based on sentiment information*. In 3rd Workshop on Interactions between Data Mining and Natural Language Processing, Riva del Garda, Italy, 2016.

Zhang, X., Fuehres, H. & Gloor, P. A. (2010). Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear". *Procedia - Social and Behavioral Sciences*, 26: 55-62.

Zhao, R. (2019). Quantifying the correlation and prediction of daily happiness sentiment and stock return: The Case of Singapore. *Physica A*, 533: 1-9.

Appendix A

Table A1 Summary of most significant studies.

Paper name	Emerging market (give names)	Developed markets (names)	Sentiment	Emotion (give words)	Influencers	Viral	Daily	High frequency	Type of model	Type of machine	Control variables
Bollen et al. (2011)		Dow Jones Industrial Average (DJIA could be used as a US proxy)		Calm, Alert, Sure, Vital, Kind and Happy			✓		Granger Causality, Multiple Linear Regression	Fuzzy Neural Network	None but they include cross-validate by checking affect on thanksgiving and presidential campaign day
Maree and Johnston (2015)	JSE ALSI (South Africa)			Depression, Tension, Anger, Vigor, Fatigue and Confusion			✓		Spearman Correlation, Granger Causality	Neural Network	None
Tabari et al. (2018)		A Tweet was considered stock related if it contains at least one of the stock symbols of the first 100 most frequent stock symbols that were included in SemEval dataset form	✓		✓	✓	✓		Granger Causality	SVM, Random Forest	None
Rao and Srivastava (2012)		NASDAQ, DJIA (Both are USA) and then they included companies: Amazon, Apple, Dell, eBay, etc	✓ - Using Tweets got Bullishness, Message Volume and Agreement				✓		Correlation, Granger Causality, OLS and then used Expert Model Mining system to see R square and Error-values		None
Zhang et al. (2010)		NASDAQ, Dow Jones and S&P 500 (ALL USA)		Hope, Happy, Fear, Worry, Nervous, Anxious, Upset, Positive, Negative	✓	✓	✓		Correlation analysis		Yes- Chicago Board Options Volatility Index (VIX) as an external benchmark of investor fear
Abbes (2015)		FTSE100 (UK)	✓		✓	✓	✓		Causality, linear regression, Breusch-pagan, Shapiro-Wilk		None

						and Kolmogorov-Smirnov, logistic		
You et al. (2017)	Ten international stock markets	✓			✓	Granger non-causality in quintiles, Quantile regressions		None
Jadhav and Wakode (2017)	S&P 500 (USA)	✓	✓	✓	✓	Logistic, correlation	SVM, Random Forest	None
Zhao (2019)	Singapore stock market	✓			✓	Linear quantile regression, nonlinear contemporaneous correlation tests, VAR model, Granger causality		None
Maqsood et al. (2020)	Four Countries	Event sentiment			✓	Linear regression	SVR Neural Network	None
Ruan et al. (2018)	Eight firms in SP500	✓-valence	✓	✓	✓	Correlation, MAE, Linear Regression,		Yes-compared treating authors equally with those that are not 'equal.'

Source: Authors' compilation.

Appendix B

Table B1 UK evaluation metrics

Measure	Sentiment (model 1)	Sentiment & Emotion (model 2)
	Naïve Bayes (NB)	
Accuracy	52.19	49.96
Recall	23.48	14.59
Precision	52.88	45.59
F1-Score	32.21	21.54
	K-Nearest Neighbours (K-NN)	
Accuracy	55.90	52.56
Recall	56.81	57.56
Precision	56.01	52.68
F1-Score	56.41	55.00
	Support Vector Machine - Kernel (SVM-K)	
Accuracy	55.23	55.02
Recall	54.75	62.19
Precision	94.31	54.85
F1-Score	69.13	58.28

Source: Authors' calculations.

Table B.2 Germany evaluation metrics

Measure	Sentiment (model 1)	Sentiment & Emotion (model 2)
	Naïve Bayes (NB)	
Accuracy	54.81	46.66
Recall	100.00	16.89
Precision	54.81	45.08
F1-Score	71.12	24.13
	K-Nearest Neighbours (K-NN)	
Accuracy	54.25	53.40
Recall	77.00	69.55
Precision	55.41	55.14
F1-Score	64.39	61.48
	Support Vector Machine - Kernel (SVM-K)	
Accuracy	55.37	56.21
Recall	96.15	89.23
Precision	55.31	56.13
F1-Score	70.08	68.80

Source: Authors' calculations.

Table B3 Japan evaluation metrics

Measure	Sentiment (model 1)	Sentiment & Emotion (model 2)
Naïve Bayes (NB)		
Accuracy	45.64	49.27
Recall	57.95	47.24
Precision	47.63	50.10
F1-Score	52.27	48.63
K-Nearest Neighbours (K-NN)		
Accuracy	54.12	51.09
Recall	69.86	61.52
Precision	54.29	52.00
F1-Score	61.07	56.35
Support Vector Machine - Kernel (SVM-K)		
Accuracy	53.52	51.70
Recall	57.95	44.86
Precision	54.22	52.70
F1-Score	56.02	50.45

Source: Authors' calculations.

Table B4 France evaluation metrics

Measure	Sentiment (model 1)	Sentiment & Emotion (model 2)
Naïve Bayes (NB)		
Accuracy	55.51	52.55
Recall	100.00	86.72
Precision	55.33	54.14
F1-Score	71.57	66.55
K-Nearest Neighbours (K-NN)		
Accuracy	55.87	52.18
Recall	98.61	84.64
Precision	55.58	53.97
F1-Score	71.44	65.81
Support Vector Machine - Kernel (SVM-K)		
Accuracy	54.40	50.34
Recall	71.44	61.03
Precision	56.05	53.20
F1-Score	62.79	56.84

Source: Authors' calculations.

Table B5 Spain evaluation metrics

Measure	Sentiment (model 1)	Sentiment & Emotion (model 2)
	Naïve Bayes (NB)	
Accuracy	55.18	52.94
Recall	89.86	84.86
Precision	55.25	54.02
F1-Score	68.31	65.91
	K-Nearest Neighbours (K-NN)	
Accuracy	55.87	46.07
Recall	71.44	64.14
Precision	57.25	49.03
F1-Score	63.54	55.54
	Support Vector Machine - Kernel (SVM-K)	
Accuracy	54.43	51.81
Recall	94.86	87.00
Precision	54.63	53.29
F1-Score	69.18	65.98

Source: Authors' calculations.