EKONOMIA I PRAWO. ECONOMICS AND LAW
Volume 24, Issue 1, March 2025
p-ISSN 1898-2255, e-ISSN 2392-1625
www.apcz.umk.pl/EiP.pl

Original article
received 17.08.2024; accepted 21.01.2025; published 14.04.2025
doi: 10.12775/EiP.2025.01

Predictive capacities of social media in the financial market: ARX-GARCH model

JOANNA MICHALAK
corresponding author
Department of Economics, Faculty of Economic Sciences and Management, Nicolaus Copernicus University, Gagarina 13A, 87-100 Torun
Joanna.michalak@umk.pl
ORCID: https://orcid.org/0000-0002-1061-401X

Abstract

Motivation: Social media platforms have emerged as a new data source for social sciences. The data extracted from them, known as big social data, are characterized by their complexity and are described within the 'V' big data model. Literature has demonstrated the influence of investor activity, as captured by BSD, on the stock market. Modeling the stock market using Twitter data allows for the capture of real-time market sentiments, potentially enhancing the accuracy of financial forecasts. The value of such models lies in their ability to identify subtle yet significant signals that traditional methods might overlook.
Aim: The aim of this study is to explore the predictive capabilities of Twitter sentiment on financial markets, specifically focusing on the application of ARX-GARCH models to analyze the impact of both, negative and positive class of emotions on market volatility.
Results: Incorporating sentiment variables into the ARX-GARCH models did not significantly enhance their predictive capabilities. While sentiment variables did not broadly improve model performance, certain variables demonstrated statistical significance at various lag levels. This indicates that some sentiments might have a delayed impact on market returns, though the overall effect size was small. Among the sentiment indicators analyzed, those based on n-gram analysis and the bullish index outperformed others, including the volume of individual emotions like anger, fear, and sadness.

Keywords: behavioral economics; ARX-GARCH; sentiment analysis; social media news
JEL: C32; C45; G17; G41

1. Introduction

Social media are widely known for their descriptive capabilities in social science, particularly in predicting outcomes in the stock market, presidential elections, consumer trends and marketing campaigns (Jensen, 2017; Antweiler & Frank, 2004; Asur & Huberman, 2010; Audrino et Al., 2005; Bollen, Mao & Zeng, 2011; Mao et al., 2011).

The relationship between social media and the stock market is complex and multidimensional. In behavioral finance, social media activity is often used to characterize moods and emotions of individual investors (Michalak, 2021). This information can be integrated into market forecasting models that describe assets prices, volumes and volatility. In some case, this integration can improve the model accuracy and reduce stock market risk (Attigeri, Manohara Pai, Pai & Nayak, 2015).

The connection between real-world stock market categories and the indicators derived from Twitter is complex. Consequently, the predictive properties of social media can be considered on multiple levels, with the most important being psychological and statistical. The psychological level involves mechanisms that link offline and online experiences, affective states, and changes in situation perception. It is generally assumed that social media content influences investors financial decisions by altering risk attitudes (Porshnev et all, 2016) and emotions (Tyszka, 2010; Michalak & Kruszewski, 2021; Xiong et al., 2018; Ferrara & Yang, 2015). The statistical level involves the analyzing time series with characteristics such as nonlinearities, long memory processes, volatility, structural changes, and complex cyclicality (Kufel, 2010; Michalak & Kruszewski, 2021). Also, sentiment analysis is a key aspect of statistical analysis, focuses on effectiveness of feature space and its role in modeling the interdependencies between social media and the stock market, particularly in terms of predictive capabilities.

This article aims to explore the predictive capabilities of social media in financial markets, with particular focus on ARX-GARCH models. The analysis evaluates how social media generated information can enhance the accuracy of predictive models and addresses the challenges of processing large datasets. The analysis specifically focuses on affective responses (in terms of emotions) with Apple case study during 2016–2017.

Two issues regarding the study need to be addressed. First, in 2023, a significant change occurred. Twitter underwent rebranding and changed its name to X. The research presented in this study was conducted before the platform's name change. The methodology and results remain relevant and valuable. The author intentionally uses the old name "Twitter" which is significant in the context of the platform's evolution. Data were collected and filtered using API available at that time, ensuring consistency and accuracy in the analyses. The second issue, directly related to the first, concerns the literature. References to older works are due to the limitation of data collection by researchers following the rebranding, hence the exclusive reference to older publications.

The rest of this article is organized as follows. Section 2 describes literature review about properties of BSD that affect research process. Section 3 describes the methods employed in the research. Section 4 contains the results and section 5 concludes.

2. Literature review

Big Social Data (BSD) is a subset of the broader Big Data trend, characterized by the "5V": volume, velocity, variety, veracity and value (Laney, 2001; Michalak & Kruszewski, 2021; Manovich, 2011; Reinsel, Gantz & Rydning, 2017). These characteristics pose significant challenges for data processing, particularly within the social context that BSD encompasses. BSD is subject to a paradox: reliable patterns can only emerge when data is aggregated (Jansen et al., 2009, p. 2169). While aggregation facilitates an accurate representation of the information stream, data sets reflecting individual users often fall short of capturing the full scope of the phenomenon (Michalak & Kruszewski, 2021, pp. 133–143; Bello-Orgaz et al., 2016). The first assumption of this study is that the analysis should focus on companies that generate significant discussion on Twitter (Manovich, 2011). This approach expands the semantic space, introducing challenges related to its complexity. These challenges fall under the semantic analysis factor, which includes data quantity, including the number of tweets, retweets, likes, replies and mentions. This high level of interaction creates a complex, high-dimensional semantic space (Liu, 2012). However, this complexity also complicates sentiment analysis.

Additionally, the semantic analysis factor is impacted by the platform's functionality. BSD is characterized by metadata provided by the social networks from which it originates. Twitter's functionality defines specific features of its data. For instance, if we consider space K with all messages, this space can be segmented by keywords, though retrieval is limited by Twitter's API limitations. The analysis of dependencies is closely tied to the quality of the collected data, i.e. the set D. Filtering and managing Twitter's message space is challenging due to irrelevant information (spam and irony), Twitter-specific characteristics (#, $, @) and multimedia content (photos, videos and links). Furthermore, the quality of semantic space depends on data trustworthiness, a critical factor that involves verifying data credibility, especially in the context of misinformation (Iskiawa, 2015). This implicate the second assumption for this study. Preparing multiple versions of preprocessing sentiment analysis will help reduce the complexity of feature space. Other issues related to semantic factor include stemming, preprocessing scenarios, vectorization methods, and feature selection methods. The latter is connected to training set issues, domain mismatch and the use of pseudo-training sets. A different domain for the training set can lead to biased classifiers, as demonstrated by studies such as those by Go (et al., 2009) Classifiers trained on such pseudo-training sets may approximate the rules established for their creation, resulting in biased outcomes.

Traditional economic analyses often tend to overlook the significance of market sentiment, treating affective factors as secondary. Even when behavioral or experimental economics address sentiment, they typically rely on conventional data sources. However, the advent Big Data has shifted the analytical environment from controlled laboratory settings to the more complex and uncontrolled, real-world scenarios. In this context, the complexity of Big Social Data requires careful consideration of how emotions are identified and aggregated. Different psychological theories, such as those proposed by Ekman (1972), Plutchik (1980), and Feldman-Barrett (2018), offer differing frameworks for understanding emotions. These theories stand in contrast to those commonly accepted in behavioral finance, such as those proposed by Nofsinger. New paradigm influencing interpretation process. The selection of emotions to be considered and the underlying psychological theory play are crucial in predictive modeling. The categorization and aggregation of emotions directly affects the efficacy of analysis. This introduces the third assumption: the need to consider a broader range of emotions beyond the simplistic positive/negative classification.

The application of Lisa Feldman Barrett's (2018) theory of constructed emotions involves interpreting emotions as social constructs. Emotions are conceptualized as responses arising from the interplay of the physical properties of the body and the brain, which is embedded within culture and society. An emotion emerges because society (along with its established set of cultural factors) teaches individuals how to interpret a specific combination of events associated with that emotion. Based on this knowledge, the brain constructs the emotional experience, in which concepts surrounding the event and the identification of affective niches play a crucial role. In the context of this definition, it can be assumed that an investor continuously constructs affective realism, which arises from (1) interoceptive predictions and (2) external events (in this case, messages on social media). These events, perceived as stimuli, are processed through visual and auditory perception, stimulating the investor. During this process, the brain draws on past experiences and concepts, including its knowledge of finance, to contextualize the stimuli. Investors operate based on concepts regarding the financial market that they have learned through societal and academic exposure. They accept the principles governing its operation and the indicators used to describe it. If any of these conditions change, the brain may generate predictions associated with unpleasant emotions (Feldman Barret, 2018). Thus, it should be assumed that the behaviors of investors, along with their emotions, are complex and dynamically evolving categories. Through stimulation by social media messages, an emotion can trigger investor actions aligned with widely accepted behavioral patterns, consistent with the direction of the emotion.

In analyzing BSD, one must address several critical statistical factors to ensure accurate and reliable results. Real-time nature of data introduces challenges such as mood variability, where fluctuations in mood over the data collection period can impact the analysis (Chen et al., 2009). Additionally, the presence of a significant outliers, which may arise from the intensity of discussions or API errors, can skew analysis results leading to non-cyclical patterns in the data. In fact, it is essential to recognize time patterns, such as the time of day or day of the week when activity levels increases. Cyclic patterns, particularly those that align with stock market hours, must be considered to avoid bias. Incorporating binary variables to model seasonality or applying seasonal differencing techniques, as demonstrated by Kufel (2010), may be necessary. Furthermore, addressing statistical issues such as heteroskedasticity, cyclicity, autocorrelation, and the potential for long-memory processes is vital to ensure the accuracy and reliability of BSD analysis. The fourth assumption involves selecting an appropriate econometric method that can effectively handle the variability and complexity inherent in BSD. Given the dynamic and high-dimensional nature of social media data, is essential to choose a robust econometric model that can effectively manage these challenges. Such a model should account for the high volatility of the data and intricate relationships between various factors, ensuring accurate and reliable analysis.

The final factor concerns the topic discussed. Researchers have identified several groups of sentiment-associated factors relevant to changes and communications on social media (Bomfim, 2003; Brenner et al., 2009; Moebert, 2009; Lucca & Moench, 2015; Ross, 1989). Including macroeconomics news (Jabeen et al., 2022), fundamental data and technical analysis (Michalak & Kruszewski, 2021), behavioral and psychological factors (Nosfinger, 2005) and political events (Liang et al., 2023). The complex nature of the social media landscape and its multifaceted influence mean that each of these groups is significant for big social data analysis (Abdollahi, 2023) and can serve as a lead topic in dataset. This brings us to the final assumption: thematic analysis must be conducted and verified to determine whether the indicators it represents could possess better predictive properties than those representing emotions. This step is essential to identify the most robust indicators for modeling the relationship between social media data and market behavior.

3. Methods and data

The study utilizes GARCH-class models to analyze the impact of Twitter on stock market. However, the research process also necessitates the construction of indicators representing emotions/sentiment based on sentiment analysis and text mining methods.

Let D represent the set of tweets, d_n $(n=1...i)$. Let X denote the set of features, where, in sentiment analysis, features are defined as words or selected word sequences (n-grams). Each set D can be characterized by $j = 1...m$ unique features used in D. Each tweet d_i is represented as a sparse vector of these m features. Let Y denote the finite set of labels representing classes, $Y = \{y_1, y_2, ... y_l\}$, where l is the number of labels, such as Y = {positive, negative} (Hastie, Tibshirani, and Friedman, 2001). In the case, the number of classes is fixed at 2. The task of the supervised machine learning module is to classify tweets into one of the two classes, y_1 or y_2.

The sentiment analysis was conducted using b0th supervised and unsupervised approaches. In supervised machine learning Senti140 dataset was selected as the training set to feed the MNB and BNB classifiers (with workflow presented in scheme 1). The labels within this dataset were generated automatically based on a emoticon keyword criterion. Go (et al., 2009) assumed that in short tweets, an emoticon is sufficient to represent the overall sentiment of tweet.

Both, training set and Apple set of tweets were prepared by using the same pre-processing scenario, based on (Effrosynidis, Symeonidis, and Arampatzis, 2017) work, which is in line with recommendations of (Krouska et al., 2016; Chen, Huang, Tian, and Qu, 2009; Zobal, 2017). Pre-processing involved following steps: removing numbers, replacing repeated punctuation with a single "multiMarks" token, normalizing text to lowercase, identifying slang and abbreviations, converting elongated words to their correct form, stopwords removal, stemming by using the Porter stemmer, removing punctuation, emoticon, and Twitter-specific characteristic. Text vectorization was carried out using the bag of words model and TF-IDF. Both, with feature modifications via n-gram (n=1,2,3). Multinomial Naïve Bayes (MNB), and Bernoulli Naïve Bayes (BNB) from Naïve Bayes classifiers family with cross-validation procedure were use to classification. Beside its simplicity MNB and BNB are one of the most effective classifier in text mining domain.

Unsupervised analysis was performed with NRC Emotion Lexicon and VADER. NRC contains 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). VADER is a NLTK dictionary with positive/negative/neutral classes. Both VADER and

NRC are assigned to the general domain, which means that some words specific to the financial domain may be misclassified in terms of sentiment. The issue of domain mismatch is not the only limitation of the dictionary-based approach. This method fails to account for semantic context, lacks flexibility in handling ambiguity, and has a limited vocabulary range. Furthermore, it does not accommodate multilingual texts effectively.

Twitter and stock market data was scaled using the normalization formula:

$$x' = \frac{x - min(x)}{max(x) - min(x)} \quad , \tag{1}$$

Where:
x – the original value,
min(x) – the minimum value in the dataset,
max(x) – the maximum value in the dataset.
Logarithmic returns in calculated with formula:

$$r_t = ln\, P_t - ln\, P_{t-1}, \tag{2}$$

Where:

P – the closing price.

Initially, the ARX-GARCH model was applied following the procedure outlined by Polshenev (et al., 2016). The GARCH model, with an added sentiment variable in the mean equation describes the return process using the equations (3). Considering the discussion presented in Polshenev (et. al., 2016) and Fiszeder (2009), the inclusion of the GARCH component in the model enables the modeling of conditional heteroskedasticity in time series representing financial markets. A notable characteristic of these series is the presence of periods with high or low volatility. The ARX component, on the other hand, facilitates the incorporation of an exogenous variable into the modeling process. Including the variable X helps account for the direct impact of sentiment and allows for its quantitative assessment. Furthermore, the model's structure permits the inclusion of lags for both exogenous and endogenous variables, enabling the capture of delayed effects of sentiment on investor behavior. All calculations were performed in Python using ARCH package, with documentation and code. Consequently, all formulas align with the notation provided on the site. The degree of differentiation of the variables was determined using the ADF test.

$$r_t = \mu + \Sigma_{i=1}^{p} \phi_i r_{t-i} + \Sigma_{j=1}^{q} \beta_j X_{t-j} + \varepsilon_t \ , \tag{3}$$

Where:
r_t – is the return rate at the time t,
μ_t – is the intercept term,
ϕ – are the autoregressive coefficients for lagged returns ,
ε_t – is the random error in period t,
X_{t-j} – represents the exogenous variable at lag j,
β_j – are the coefficients for the lagged exogenous variable.

Random errors are described by the following formula:

$$\varepsilon_t = \sigma_t z_t \ , \tag{4}$$

where:
z_t – a white noise process representing the random shocks.

The conditional Variance Equation for GARCH (1,1) process is given by:

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \ , \tag{5}$$

where:
σ_t^2 – conditional variance,
ω – the constant,
α – the parameter associated with the squared residuals from the previous period
β – the parameter related to the previous conditional variance.

As part of the procedure for testing the impact of sentiment, it was decided to analyze the effect of each sentiment separately, rather than estimating the combined effect of emotion groups. This decision stems from the multicollinearity of variables resulting from the procedure used to construct the indicators and the classification of the original dataset into emotion categories.

For each model in Tables 5 and 6, a lag of order 3 was used in the ARX component. Introducing lags allows capturing the delayed effect of investor reactions over time. Three lags are sufficient to account for the delayed impact of sentiment on market volatility without introducing excessive complexity into the model. This choice represents a balance between model complexity and its accuracy.

The data for Apple Inc. was collected from January 1, 2016, to December 31, 2017, resulting in a total of 808,218 messages, averaging of 2,204 messages per day, with no zero activity. The topics discussed in relation to the company included: (1) financial results, (2) management and CEO decisions, (3) products or services, (4) opinions on various events and their implications, and (5) spam. Tweets with multiple cashtags were generally low in information content and were categorized as a spam. Messages containing a single cashtag made up 66% of the Apple dataset.

Table 1 presents basic statistics on sentiment indicators. High variability across sentiments can be observed, as indicated by the coefficients of variation (CV). For instance, the CV for "sadness" is particularly high at 167%, suggest a wide fluctuation over time. The positive and negative sentiment values differ significantly between the NRC and VADER methods. Negative NRC has a mean of 54 tweets while negative VADER has much higher mean of 170 tweets. This may suggest that VADER might detect more negative sentiment rather than NRC. Machine Learning sentiments indicators show the highest average values for both positive and negative sentiments. The lower CVs (57% and 65%) suggest that these values are more consistent compared to NRC and VADER. The neutral sentiment measured with VADER approach has the highest mean value among all sentiments, which is expected as neutral sentiments often dominate social media discussion. The minimum and maximum values shows that the occurrence of certain sentiments can spike, lead to outliers in time series.

The n-gram analysis (n = 1 to 2) allowed for identification of dominant tokens in the discussion. However, these tokens are rare, indicating a dispersion of the conversation. The most frequently occurring bigram was 'iphone 7,' with a frequency of 27.193, the new version of the iOS 10 system (6.372). Other common terms paired with 'iphone' included 'pro,' '6,' 'new,' and 'sale.'

Key bigrams related to decision-makers include 'Tim Cook' (9.848 occurences) and investors like 'Warren Buffett' (2.118 occurences). The stock market was referenced through bigrams such as 'stock market,' 'share stock,' and 'stock price.'

4. Results

The results for sentiment analysis using supervised machine learning approaches are presented in Tables 2 and 3. The effectiveness of various vectorization methods was compared, specifically the Bag-of-Words and TF-IDF methods, combined with different n-gram values. Combining 1-gram features with higher-order n-grams (2-gram and 3-gram) yielded the best results. For further analysis, the CountVectorizer n-gram (1,2) and TF-IDF n-gram (1,2) representations were selected. The use of 3-grams was omitted due to the increased dimensionality of the feature space. For both the MNB and BNB classifiers, classification accuracy results were identical. Additionally, both classifiers produced similar accuracy values, leading to the decision to proceed with the Multinomial Naive Bayes (MNB) classifier.

In the subsequent analysis, a 10-fold cross-validation approach was employed. This method was chosen to verify whether the model's performance across 10 different folds aligns with the results presented in table 3. Cross-validation provides a more objective assessment, as relying on a single, isolated test set can introduce biases due to specific language nuances that might occur randomly in that set. A potential limitation of cross-validation is the number of folds used in the analysis, which can depend on the complexity of the classifier. However, given that the Multinomial Naive Bayes (MNB) classifier is not computationally intensive, 10-fold cross-validation was deemed appropriate. According to the data presented in table 2, the CountVectorizer n-gram (1,2) representation yields consistent results across the folds, as does the TF-IDF n-gram (1,2) representation.

Sentiment analysis across three different approaches reveals different patterns in the emotional content of tweets. According to the NRC method, positive sentiment dominates, comprising 10.60% of the total messages volume, followed by negative emotions such as fear (1.94%), anger (1.37%) and sadness (1.19%). The VADER approach similarly indicates a positive sentiment at 32.67% with 15.74% of tweets being negative and 51,59% neutral. In contrast the machine learning approach suggest an even higher positive sentiment at 63.85% with 36.15% classified as negative. Overall, the results demonstrate that positive sentiment consistently outweighs negative emotions in discussions about Apple, although the proportion of positive to negative sentiment varies depending on the methos use.

All analyzed variables (normalized by (1) formula), except for 'anger', exhibits stationarity with ADF test. The variable "anger" was differenced to the second degree.

Table 4 and 6 contains results for ARX-GARCH two years analysis. The results indicate that while the X in ARX component is statistically significant for some lags, their predictive performance remains limited. MSE and RMSE values across the different models, indicating that predictions deviate actual outcome. Models with X = 'anger', 'fear', 'sadness', 'negative NRC' and "negative VADER' struggle to accurately predict return based on emotional sentiment data. The "Bullish Index" had one of the higher MSE and RMSE values (0.0058024 and 0.0761739, respectively), suggesting that even variables traditionally considered as strong indicators of market trends do not significantly enhance the model's predictive capability over this period.

Bullish Index and fear NRC has shown some statistical significance in certain lags, with p-value of 0.049 and 0.014 for the second and third lags, respectively (fear). This suggest that fear sentiment might have a delayed but significant impact on returns. Public discussion about "Tim Cook" could influence market returns, especially with one and three day delay.

However, the challenge of explaining variation over two-year period is evident, particularly when add variables from BSD. To better understand the dynamics and refine the models, a shorter time frame analysis was conducted – 2017-01-01:2017-7-29'. This analysis reveal that certain variables, such as Bullish Index and Negative VADER, exhibits more evident over shorter period, aligning with the understanding that financial markets are often influenced by more immediate sentiment and news.

5. Conclusion

The aim of this paper was capture predictive capabilities of social media sentiment on example of Twitter and Apple. The chosen method for describing the processes was ARX-GARCH model. Predictive capabilities refers to model's accuracy. The second section presents the factors influencing modeling of Twitter variables' impact on the stock market. This is author's contribution to the existing literature, considering both text mining analysis and econometric analysis in comprehensive manner. Most publications exclusively on just one of these aspects.

The findings indicate that the inclusion of emotional sentiment and other independent variables in the ARX-GARCH models over a two-year period does not lead to improvements in predictive performance. The generally low R-squared values, combined with the MSE and RMSE metrics, suggest that these models do not adequately capture the complexity of the relationships between the studied variables and financial return volatility. On the other hand, the statistical significance of certain lags has shown that Twitter cannot be ignored in the stock market.

In the literature, results are often summarized in a way that highlights Twitter's ability to predict volatility, prices and movements in the stock market. However, this capability is demonstrated for selected indicators. For instance, the pioneering study by Bollen (et all., 2011) showed that among all four dimensions of sentiment, only 'calm' improved the model's properties and DJIA forecasts. They used Self-Organizing Feature Maps (SOFM) neural networks, which are believed to capture process complexities better than the GARCH model. Pagalu (et al., 2016) analyzed yield increases/decreases and Twitter sentiment's ability to reflect these movements. Analyzing the direction of changes is inherently simpler than analyzing residual volatility. Roa & Strivastava (2012) et al., who developed the Bullish Index, compared its properties to logarithmic returns but ultimately selected the OLS model for forecasting. The performance of forecasts depends on the chosen time window, which aligns with the author's conclusion about the need to identify an appropriate time window for forecasting. The author noted that as the time window lengthens, the R^2 value increases only up to a certain point and then begins to decrease, with a critical value being one month.

Zhang conducted sentiment analysis differently from the approach presented in this article. His method used a straightforward assumption—tweets containing keywords like "fear" were used to construct a fear index. This approach was based on the brief nature of messages and the assumption that the words used represent the sentiment of the message (similar to Go's use of emoticons in constructing Senti140). After aggregating daily messages, the percentage of messages representing sentiment relative to the total number of messages sent each day was calculated, and these indicators were then correlated with the stock market. Zhang (et al., 2011) demonstrated correlations between basic indicators and the Dow Jones, NASDAQ, and S&P 500. He also showed correlations with the VIX index. In conclusion, Zhang found that increased market uncertainty is associated with a rise in the use of words indicating uncertainty. Mittel demonstrated a causal relationship with a 3-4 day lag for only two sentiment dimensions: calmness and happiness.

Thus, the results obtained in this study do not differ significantly from those presented in the literature. They confirm the issue of adapting the time window, indicator, and method to detect correlations between BSD and the stock market.

References

Abdollahi, H. (2023). *Oil price volatility and new evidence from news and Twitter*. Energy Economics, 122, 106711. https://doi.org/10.1016/j.eneco.2023.106711.

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, 1259-1294. doi:10.1111/j.1540-6261.2004.00662.x.

Asur, S. & Huberman, B. (2010). *Predicting the future with social media*. In Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (Vol. 1, pp. 492-499). IEEE Computer Society.

Attigeri, G.V., Manohara Pai, M.M., Pai, R.M. & Nayak, A. (2015). *Stock market prediction: A big data approach*. TENCON 2015 – 2015 IEEE Region 10 Conference, 1–5. https://doi.org/10.1109/TENCON.2015.7373006

Audrino, F., Sigrist, F. & Ballinari, D. (2020). *The impact of sentiment and attention measures on stock market volatility*. International Journal of Forecasting, 36(2), 334–357. ISSN 0169-2070. https://doi.org/10.1016/j.ijforecast.2019.05.010.

Bello-Orgaz, G., Jung, J.J. & Camacho, D. (2016). *Social big data: Recent achievements and new challenges*. Information Fusion, 28, 45-59.

Bollen, J., Mao, H. & Zeng, X. (2011). *Twitter mood predicts the stock market.*

Journal of Computational Science, 2(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007.
Bomfim, A.N. (2003). *Pre-announcement effects, news effects, and volatility: Monetary policy and the stock market*. Journal of Banking & Finance, 27(1), 133–151.
Brenner, M., Pasquariello, P. & Subrahmanyam, M. (2009). *On the volatility and comovement of US financial markets around macroeconomic news announcements*. Journal of Financial and Quantitative Analysis, 44(6), 1265–1289.
Chen, J., Huang, H., Tian, S. & Qu, Y. (2009). *Feature selection for text classification with Naïve Bayes*. Expert Systems with Applications, 36, 5432––5435. doi: 10.1016/j.eswa.2008.06.054.
Effrosynidis, D., Symeonidis, S. & Arampatzis, A. (2017). *A Comparison of Pre-processing Techniques for Twitter Sentiment Analysis*. In" 21st International Conference on Theory and Practice of Digital Libraries (TPDL 2017), LNCS 10450 (pp. 421–426). Thessaloniki, Greece: Springer. doi: 10.1007/978-3-319-67008-9_31
Ekman, P. (1972). *Universals and cultural differences in facial expressions of emotion*. In: J. Cole (Ed.), Nebraska symposium on motivation, 1971 (pp. 207–283). Lincoln: University of Nebraska Press.
Feldman Barrett, L. (2018). *Jak powstają emocje? Sekretne życie mózgu*. Warszawa: CeDeWu Sp. z o.o.
Ferrara, E. & Yang, Z. (2015). *Measuring emotional contagion in social media*. PLoS ONE, 10(11), e0142390. https://doi.org/10.1371/journal.pone.0142390.
Fiszeder, P. (2009). *Modele klasy GARCH w empirycznych badaniach finansowy*ch [eBook]. Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika.
Go, A., Bhayani, R. & Huang, L. (2009). *Twitter sentiment classification using distant supervision*. CS224N Project Report, Stanford, 1(2009).
Hastie, T., Tibshirani, R. & Friedman, J. (2001). *Elements of statistical learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA. Online Access (Accessed on March 15, 2019): Retrieved from https://web.stanford.edu/~hastie/Papers/ESLII.pdf.
Ishikawa, H. (2015). *Social Big Data Mining*. CRC Press. https://doi.org/10.1201/b18223.
Jabeen, A., Yasir, M., Ansari, Y., Yasmin, S., Moon, J. & Rho, S. (2022). *An empirical study of macroeconomic factors and stock returns in the context of economic uncertainty news sentiment using machine learning*. Complexity.
Jansen, B., Zhang, M., Sobel, K. & Chowdury, A. (2009). *Twitter Power: Tweets as electronic word of mouth*. Journal of the American Society for Information Science and Technology, 60(11), 2169–2188.
Jensen, E. (2017). *Putting the methodological brakes on claims to measure national happiness through Twitter: Methodological limitations in social*

media analytics. PloS one, 12, e0180080. https://doi.org/10.1371/journal.pone.0180080.

Krouska, A., Troussas, C. & Virvou, M. (2016). *The effect of preprocessing techniques on Twitter sentiment analysis*. In: 2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA) (pp. 1–6). IEEE. doi: 10.1109/IISA.2016.7785373.

Kufel, T. (2010). *Cykle gospodarcze i modele ekonometryczne*. Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika.

Laney, D. (2001). *3D data management: controlling data volume, velocity, and variety*. Meta Group Research Note, 6, 70–73.

Liang, F. & Lu, S. (2023). *The dynamics of event-based political influencers on Twitter*. Social Media + Society, 9(3), 1–12. https://doi.org/10.1177/20563051231177.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining* (1st ed.). Springer Cham. https://doi.org/10.1007/978-3-031-02145-9.

Lucca, D.O. & Moench, E. (2015). *The pre-FOMC announcement drift*. Journal of Finance, 70(1), 329-371.

Manovich, L. (2011). *Trending: The promises and the challenges of big social data*. https://manovich.net/content/old/03-articles/64-article-2011/64-article-2011.pdf.

Mao, Y., Wei, W. & Liu, B. (2012). *Correlating S&P 500 stocks with Twitter data*. In Proceedings of the ACM SIGKDD Workshop on Sentiment Analysis in Social Media (pp. 69–72). DOI: 10.1145/2392622.2392634.

Michalak, J. (2021). *Opportunities and challenges of big social data analytics based on examples from psychology*. In: K.S. Soliman (Ed.), Innovation management and information technology impact on global economy in the era of pandemic: Proceedings of the 37th International Business Information Management Association Conference (IBIMA), 30–31 May 2021, Cordoba, Spain (pp. 6374–6381).

Michalak, J. & Kruszewski, T. (2021). *Pandemia a reakcje inwestorów: analiza big data komunikatorów z Twittera wobec zjawiska niepewności*. Toruń: Towarzystwo Naukowe Organizacji i Kierownictwa.

Möbert, J. (2009). *Dispersion in beliefs among speculators as a determinant of crude oil prices*. Deutsche Bank Research, Research Notes, 32.

Nofsinger, J.R. (2005). *Social mood and financial economics*. Journal of Behavioral Finance, 6(3), 144–160.

Pagolu, V.S., Challa, K.N.R., Panda, G. & Majhi, B. (2016). *Sentiment analysis of Twitter data for predicting stock market movements*. CoRR, abs/1610.09225.

Plutchik, R. (1980). *Chapter 1 – A GENERAL PSYCHOEVOLUTIONARY THEORY OF EMOTION*. In: R. Plutchik & H. Kellerman (Eds.), Theories of Emotion (pp. 3–33). Academic Press. doi:10.1016/B978-0-12-558701-3.50007-7.

Porshnev, A.V., Lakshina, V.V. & Redkin, I.E. (2016). *Using emotional markers' frequencies in stock market ARMAX-GARCH model* (MPRA Paper No. 82875). University Library of Munich, Germany.

Rao, T. & Srivastava, S. (2012). *Analyzing stock market movements using Twitter sentiment analysis*. In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012) (pp. 119–123). https://doi.org/10.1109/ASONAM.2012.30.

Reinsel, D., Gantz, J. & Rydning, J. (2017, April). *Data Age 2025: The Evolution of Data to Life-Critical*. [White paper]. Seagate Technology LLC. Retrieved from https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf.

Ross, S.A. (1989). *Information and volatility: The no-arbitrage martingale approach to timing and resolution irrelevancy*. Journal of Finance, 44(1), 1–17.

Tyszka, T. (2010). *Decyzje. Perspektywa psychologiczna i ekonomiczna.* W: J. Brzeziński (red.), Wykłady z psychologii (t. 16, s. 299–327). Warszawa: Wydawnictwo Naukowe Scholar.

Xiong, X., Li, Y., Qiao, S., Han, N., Wu, Y., Peng, J. & Li, B. (2018). *An emotional contagion model for heterogeneous social media with multiple behaviors*. Physica A: Statistical Mechanics and its Applications, 490, 185–202. https://doi.org/10.1016/j.physa.2017.08.025.

Zhang, X., Fuehres, H. & Gloor, P. (2011). *Predicting Stock Market Indicators Through Twitter – "I Hope it is Not as Bad as I Fear*. Procedia - Social and Behavioral Sciences, 26, 55–62. https://doi.org/10.1016/j.sbspro.2011.10.562.

Zobal, V. (2017). *Sentiment analysis of social media and its relation to stock market* (Bachelor's thesis). Prague: Charles University, Faculty of Social Sciences, Institute of Economic Studies.

Acknowledgements

Author contributions: author has given an approval to the final version of the article.

Funding: this research was funded by a Faculty of Economics Sciences and Management, Nicolaus Copernicus University in Torun.

Supplementary information: author/authors acknowledge following people and institution for help with the preparation of the article: (name and last name or institution name).

Note: the results of this study were presented in another form, such as a poster/abstract at a conference.

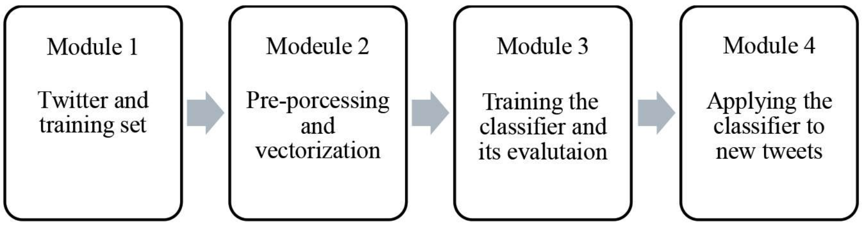
Appendix

Table 1. Basic Statistics for Apple Sentiment Indicators

Indicator	Mean	Minimum	Maximum	Standard Deviation	Coefficient of Variation
Anger	14.83	0	129	16.59	111%
positive (NRC)	2.94	0	56	4.37	149%
Fear	20.96	0	188	24.62	117%
negative (NRC)	54.02	1	692	56.11	104%
Sadness	12.89	0	426	21.60	167%
negative (VADER)	170.29	10	1 401	157.91	93%
positive (VADER)	353.52	18	1 789	230.66	65%
neutral (VADER)	558.32	31	3 771	415.12	74%
negative (ML)	499.15	30	2 593	326.07	65%
positive (ML)	863.99	35	4 086	495.20	57%

Source: own preparation.

Scheme 1. Sentiment analysis pipeline (SML)



Source: own preparation.

Table 2. Accuracy results of MNB and BNB classifiers for different input data variants

Input	MNB	BNB
CountVectorizer n_gram (1,1)	0.76	0.76
TF-IDF n_gram (1,1)	0.75	0.76
CountVectorizer n_gram (1,2)	*0.77*	*0.77*
CountVectorizer n_gram (2,2)	0.71	0.73
CountVectorizer n_gram (1,3)	0.77	0.77
CountVectorizer n_gram (3,3)	0.57	0.59
CountVectorizer max_df = 80%	0.76	0.76
TF-IDF n_gram (1,2)	*0.77*	*0.77*

Input	MNB	BNB
TF-IDF n_gram (2,2)	0.72	0.73
TF-IDF n_gram (1,3)	0.77	0.77
TF-IDF n_gram (3,3)	0.58	0.59
TF-IDF max_df = 80%	0.75	0.76

Source: own preparation.

Table 3. Accuracy results of MNB using 10-fold cross-validation

Fold	Input	
	CountVectorizer n_gram (1,2)	TF-IDF n_gram (1,2)
1	0.7717625	0.77205000
2	0.7711750	0.77194000
3	0.7701250	0.77100652
4	0.7724125	0.77345625
5	0.7715565	0.77281250
6	0.7713437	0.77298120
7	0.7722875	0.77248125
8	0.7715187	0.77284375
9	0.7725250	0.77263750
10	0.7699000	0.77134375

Source: own preparation.

Table 4. Forecast results from the ARX-GARCH(1,1) model for the entire sample period

Variable	MSE	RMSE
Log_returns	0.0001165	0.0107946
Anger	0.005669918489112514	0.07529886114087327
Fear	0.005699481277622778	0.07549490895168215
Negative NRC	0.005928843960052896	0.07699898674692347
Positive NRC	0.005692587271408601	0.07544923638718024
Sadness	0.005692467771616053	0.07544844446120842
Positive VADER	0.005703967614034624	0.07552461594761423
Negative VADER	0.00592364749401755	0.07696523561984041
Neutral VADER	0.005692861412134283	0.07545105308830542
Negative ML	0.0056820356504783485	0.07537927865453707
Positive ML	0.00572882863650787	0.0756890258657612
Tim Cook	0.005705412260441544	0.07553417941860191

Variable	MSE	RMSE
Stock market	0.005744863528669907	0.07579487798439884
Iphone7	0.005700893300462075	0.07550426014776965
Bullish Index	0.005802472404850668	0.07617396146223897

Source: own preparation.

Table 5. Significance of sentiment variables for ARX-GARCH(1,1) models in the sample 2017-01-01:2017-07-29

Variable	*t*	*t-1*	*t-2*	*t-3*	R-Square
Anger					0.010
Fear			X	X	0.025
Negative NRC				X	0.018
Positive NRC				X	0.011
Sadness	X			X	0.010
Positive VADER				X	0.025
Negative VADER				X	0.050
Neutral VADER				X	0.025
Negative ML					0.019
Positive ML					0.007
Tim Cook	X		X	X	0.028
Stock Market		X			0.007
Iphone 7				X	0.033
Bullish Index	X			X	0.050

Source: own work.

Table 6. Results for ARX-GARCH(1,1) analysis

Variable	Equation
Log_returns R-square: 0.000	$\boldsymbol{y_t} = \mathbf{0{,}001155}\ (0{,}07173) + \boldsymbol{\varepsilon_t}$ $\sigma^2_t = 0{,}00\ (0{,}00) + 0{,}0\mathbf{100}\boldsymbol{\varepsilon}^2_{t\text{-}1}\ (\mathbf{0{,}1154})$ $+ 0{,}9700\ (0{,}00)\ \sigma^2_{\boldsymbol{t\text{-}1}}$
Fear NRC R-square: 0.22	$y_t = 0{,}5260\ (0{,}000) - 0{,}0181 * \boldsymbol{X_t}\ (0{,}708) -$ $0{,}0235\boldsymbol{X_{t-1}}\ (0{,}674) - 0{,}1006 * \boldsymbol{X_{t-2}}\ (0{,}049) + 0{,}1161 *$ $\boldsymbol{X_{t-3}}\ (0{,}014) + \boldsymbol{\varepsilon_t}$ $\sigma^2_t = 1{,}8693\ (0{,}116) + 0{,}0100\boldsymbol{\varepsilon}^2_{t\text{-}1}\ (0{,}\mathbf{405})$ $+ 0{,}9700\ (0{,}00)\ \sigma^2_{\boldsymbol{t\text{-}1}}$

Variable	Equation
Anger R-square: 0.003 NRC	$y_t = 0{,}5225\ (< 0{,}001) - 0{,}0173 * \boldsymbol{X_t}\ (0{,}759) + 0{,}469\boldsymbol{X_{t-1}}\ (0{,}642) - 0{,}0344 * \boldsymbol{X_{t-2}}\ (0{,}548) + 0{,}0466 * \boldsymbol{X_{t-3}}\ (0{,}362) + \boldsymbol{\varepsilon_t}$ $\sigma_t^2 = 1{,}1399 * \mathbf{10^{-4}} + 2{,}1015 * \mathbf{10^{-10}}\boldsymbol{\varepsilon}^2_{t-1}\ (\mathbf{1{,}00}) + 0{,}9832\ (<0{,}001)\ \sigma^2_{t-1}$
Negative NRC R-square=0.10	$y_t = 0{,}5298\ (0{,}000) - 0{,}0662 * \boldsymbol{X_t}\ (0{,}835) - 0{,}0306\boldsymbol{X_{t-1}}\ (0{,}868) - 0{,}0810 * \boldsymbol{X_{t-2}}\ (0{,}796) + 0{,}1044 * \boldsymbol{X_{t-3}}\ (0{,}290) + \boldsymbol{\varepsilon_t}$ $\sigma_t^2 = 0{,}000114\ (0{,}504) + 0{,}000\boldsymbol{\varepsilon}^2_{t-1}\ (1{,}000) + 0{,}9832\ (0{,}00018)\ \sigma^2_{t-1}$
Positive NRC R-square=0.14	$y_t = 0{,}5275\ (0{,}000) - 0{,}1016 * \boldsymbol{X_t}\ (0{,}054) - 0{,}0730\boldsymbol{X_{t-1}}\ (0{,}476) - 0{,}1250 * \boldsymbol{X_{t-2}}\ (0{,}041) + 0{,}0817 * \boldsymbol{X_{t-3}}\ (0{,}107) + \boldsymbol{\varepsilon_t}$ $\sigma_t^2 = 0{,}000188\ (0{,}138) + 0{,}0100\boldsymbol{\varepsilon}^2_{t-1}\ (0{,}397) + 0{,}9700\ (0{,}000)\ \sigma^2_{t-1}$
Sadness R-square=0.15	$y_t = 0{,}5317\ (0{,}000) - 0{,}1772 * \boldsymbol{X_t}\ (0{,}095) - 0{,}0595\boldsymbol{X_{t-1}}\ (0{,}610) - 0{,}0462 * \boldsymbol{X_{t-2}}\ (0{,}571) + 0{,}1249 * \boldsymbol{X_{t-3}}\ (0{,}027) + \boldsymbol{\varepsilon_t}$ $\sigma_t^2 = 1{,}8815 * \mathbf{10^{-4}}\ (0{,}147) + 9{,}9998 * \mathbf{10^{-3}}\boldsymbol{\varepsilon}^2_{t-1}\ (0{,}396) + 0{,}9700\ (0{,}000)\ \sigma^2_{t-1}$
Positive VADER R-square = 0.029	$y_t = 0{,}5237\ (< 0{,}000) - 0{,}0761 * \boldsymbol{X_t}\ (0{,}252) - 0{,}0714\boldsymbol{X_{t-1}}\ (0{,}349) - 0{,}1010 * \boldsymbol{X_{t-2}}\ (0{,}190) + 0{,}1184 * \boldsymbol{X_{t-3}}\ (0{,}009) + \boldsymbol{\varepsilon_t}$ $\sigma_t^2 = 1{,}8645 * \mathbf{10^{-4}}\ (0{,}136) + 9{,}9999 * \mathbf{10^{-3}}\boldsymbol{\varepsilon}^2_{t-1}\ (0{,}446) + 0{,}9700\ (0{,}000)\ \sigma^2_{t-1}$
Negative VADER R-square = 0.033	$y_t = 0{,}5340\ (< 0{,}000) - 0{,}1072 * \boldsymbol{X_t}\ (0{,}729) - 0{,}0138\boldsymbol{X_{t-1}}\ (0{,}974) - 0{,}0982 * \boldsymbol{X_{t-2}}\ (0{,}830) + 0{,}1494 * \boldsymbol{X_{t-3}}\ (0{,}127) + \boldsymbol{\varepsilon_t}$ $\sigma_t^2 = 1{,}2174 * \mathbf{10^{-4}}\ (0{,}117) + 1{,}2294\boldsymbol{\varepsilon}^2_{t-1}\ (1{,}000) + 0{,}9821\ (0{,}002)\ \sigma^2_{t-1}$
Neutral VADER R-square = 0.035	$y_t = 0{,}5260\ (< 0{,}000) + 0{,}0207 * \boldsymbol{X_t}\ (0{,}750) - 0{,}0033\boldsymbol{X_{t-1}}\ (0{,}974) - 0{,}1850 * \boldsymbol{X_{t-2}}\ (0{,}080) + 0{,}1684 * \boldsymbol{X_{t-3}}\ (0{,}009) + \boldsymbol{\varepsilon_t}$ $\sigma_t^2 = 1{,}1274 * \mathbf{10^{-4}}\ (0{,}117) + 1{,}2294\boldsymbol{\varepsilon}^2_{t-1}\ (1{,}000) + 0{,}9821\ (0{,}002)\ \sigma^2_{t-1}$

Variable	Equation
Negative ML R-square: 0.012	$y_t = 0{,}5369\ (0{,}000) + 0{,}0737 * \boldsymbol{X_t}\ (0{,}273) - 0{,}0419\boldsymbol{X_{t-1}}\ (0{,}683) - 0{,}0532 * \boldsymbol{X_{t-2}}\ (0{,}557) + 0{,}0270 * \boldsymbol{X_{t-3}}\ (0{,}725) + \boldsymbol{\varepsilon_t}$ $\sigma_t^2 = 1{,}8868 * \mathbf{10}^{-4}\ (0{,}143) + 0{,}0100\boldsymbol{\varepsilon}^2_{t-1}\ (0{,}354) + 0{,}9700\ (0{,}000)\ \sigma^2_{t-1}$
Positive ML R-square: 0.002	$y_t = 0{,}5200\ (0{,}000) + 0{,}0284 * \boldsymbol{X_t}\ (0{,}631) - 0{,}0608\boldsymbol{X_{t-1}}\ (0{,}488) - 0{,}0127 * \boldsymbol{X_{t-2}}\ (0{,}895) + 0{,}0351 * \boldsymbol{X_{t-3}}\ (0{,}659) + \boldsymbol{\varepsilon_t}$ $\sigma_t^2 = 1{,}8936 * \mathbf{10}^{-4}\ (0{,}117) + 9{,}9896\boldsymbol{\varepsilon}^2_{t-1}\ (0{,}458) + 0{,}9690\ (0{,}000)\ \sigma^2_{t-1}$
Tim Cook R-square: 0.027	$y_t = 0{,}5259\ (0{,}000) + 0{,}1037 * \boldsymbol{X_t}\ (0{,}014) - 0{,}0415\boldsymbol{X_{t-1}}\ (0{,}221) - 0{,}0864 * \boldsymbol{X_{t-2}}\ (0{,}012) + 0{,}0858 * \boldsymbol{X_{t-3}}\ (0{,}012) + \boldsymbol{\varepsilon_t}$ $\sigma_t^2 = 1{,}8589 * \mathbf{10}^{-4}\ (0{,}133) + 9{,}9896 * \mathbf{10}^{-3}\boldsymbol{\varepsilon}^2_{t-1}\ (0{,}456) + 0{,}9700\ (0{,}000)\ \sigma^2_{t-1}$

In parentheses, the p-value is provided.
Source: own preparation.