




Does pre-processing affect the correlation indicator between Twitter message volume and stock market trading volume?

JOANNA MICHALAK

Nicolaus Copernicus University in Toruń, Faculty of Economic Sciences and Management,
Department of Economics, ul. Gagarina 13a, 87-100 Toruń, Poland

✉ joanna.michalak@umk.pl

 orcid.org/0000-0002-1061-401X

Abstract

Motivation: More and more authors empirically verify the relationship between the volume of tweets and the stock market indicators. The patterns explored from Twitter most often take the form of time series that represent user's activity on different level of granularity (moods, emotions, relevant topic or query-related messages). Sentiment analysis is a technique used to transform text data into information on the mood and related behavioral categories. Supervised machine learning is the most commonly used approach to sentiment analysis. Thus, the results of an empirical analysis of the relationship between social media and stock depend on the quality of results of classification task. The quality of the features used to learn the classifier plays a key role. The feature space is modified using various data pre-processing scenarios that aim to increase accuracy of classification. The impact of pre-processing data on the quality of classification is often discussed in studies. Very few authors discuss the impact of pre-processing on the correlation indicator between Twitter and stock market.

Aim: Analysis of the impact of tweets pre-processing on the Pearson correlation indicator between the mood of Twitter users and stock market trading volume.

Results: The correlation between the volume of stock market trading and the volume of tweets has been empirically confirmed. The effect of pre-processing on the correlation index was noted for the variables 'all_tweets' and 'negative_tweets'. This is because the training set has a significant amount of tweets with negation. However, the results



are not conclusive. The differences between the Pearson correlation index calculated for scenario one and scenario four are not significant. However, this indicates that the effect of noise data may reduce the quality and precision of conclusions. Especially in the case of frequent repetition of a certain category of noise.

Keywords: twitter sentiment analysis; behavioral economy; data mining

JEL: G4I; C38

1. Introduction

Financial market theory is based on assumption of rationality and efficient market hypothesis. However actual processes are full of anomalies which cannot be explained by full rationality assumption, e.g. January effect, herd behaviour, black swan. Behaviorists believe that the impact of social and psychological factors on investor decision making is important in explaining them. Investors feel emotions and are in the mood. Mood affects the decision-making process, cognitive processes and motivation of investors. Decision-makers with a positive mood are optimists, they perceive events as opportunities. Affective perception can affect the expectation of investors, their strategies and risk perception. Emotions are intense, oriented and short-lived feelings. Emotions strongly influence perceptions. Transfer of emotions can be done through activity in online communities. According to Oh & Sheng (2011) there are two types of investors (1) rational — who make their decisions based on fundamental information's, and (2) noise investors. Noise investors operate on the basis of pseudo signals. They are often active on so-called online investors groups. Conversation in the context of noise investors involves discussing alternatives, forecasting, asking, sharing opinions. One of the most used online forums is Twitter. Real-time investor messages are transmitted via Twitter. Some messages can affect investors' emotional processes. Data generated by twitter users can be collected. By processing them, we can infer about moods and emotions for specific groups, such as noise investors. Many authors support this statement (Bollen et al., 2011; Mittal & Goel, 2012; Oh & Sheng, 2011; Rao & Srivastava, 2013).

Sets of text messages are called big social data. According to Ishikawa (2015) and Olshannikova et al. (2017) big social data refers to large data volumes which describe people's behaviour and technology-mediated social interactions in the digital realm. We treat these data as a big data, in term of 3V dimension model. According to 3V model big social data can be treat as predictors for real categories because of high frequency, real-time and volume. In addition, Nisar & Yeung (2018) indicate that social networking sites provide the opportunity to observe the global trend while looking into the individual behaviour. Big social data requires appropriate tools and techniques to transform them into information, like Twitter Sentiment Analysis. Research activity in Twitter Sentiment Analysis field can be divided into the following areas (1) development of research techniques and tools and (2) implementation of explored indicators in various domains. Twitter Sentiment analysis process contains the following

stages: data pre-processing and feature selection. Quality of the features used to learn the classifier plays a key role. The feature space is modified using various data pre-processing scenarios that aim to increase accuracy of classification. The impact of pre-processing data on the quality of classification is often discussed in studies. Very few authors discuss the impact of pre-processing on the correlation indicator between Twitter and stock market. Thus, the aim of the analysis is to verify of the impact of tweets pre-processing on the Pearson correlation indicator between the mood of Twitter users and stock market trading volume. The presented conclusions are the result of a pilot study.

2. Literature review

In behavioral economics early works on knowing investors' sentiment are based on survey. Then, the potential of domain online forums has been noticed, e.g. yahoo.finance (Antweiler & Frank, 2004; Wysocki, 1998). Antweiler & Frank (2004) confirmed that the forum messages reflect current company activity. Later investors turned to social media, which provides them with real-time messages. They start to build so-called online investor communities (Oh & Sheng, 2011). Over time Twitter has become widely accepted as leading platform. Rao & Srivastava (2013) analysed DJIA, NASDAQ-100 and EURO FOREX with information from two online sources: Twitter and SVI Google. They used correlation, Granger causality analysis and ARIMA models. They result indicate a relationship between Twitter/SVI/indexes. Zhang et al. (2011) confirmed the correlation between Twitter and Dow Jones. S&P500 and NASDAQ. Mao et al. (2012) omitted the information about sentiment and as variable they included the volume of tweets. Correlation was confirmed. Bollen et al. (2011) carried out the task of emotion recognition on text data and confirmed correlation and predictive value via neutral network. Mittal & Goel (2012) they studied the relationship by logistic regression, linear regression, SVM and neutral networks. They confirmed the conclusion of previous authors, like Oh & Sheng (2011), Porshnev et al. (2016), Strauß et al. (2018).

The research process is based on a study conducted by Nisar & Yeung (2018). To begin with pairwise correlations, classical Pearson correlation parameter was chosen for interpretation. The volume-based analysis of correlation compares the trade volume with tweets volume (all tweets, positive tweets, negative tweets and neutral tweets). Variables were normalized using z-score, provided by formula:

$$z(xt) = \frac{xt - \mu(x)}{\delta(x)}, \quad (1)$$

where:

- $\mu(x)$ — represents the mean;
- $\delta(x)$ — represents standard deviation.

After the establish the association between variables, the multiple regression is used with 4 independent variables (positive tweets, negative tweets, neutral tweets). By using standardized beta coefficient, the strength of the effect of each independent variable is measure. The higher the absolute value of beta coefficient, the stronger effect. Advantage of using standardized beta it that variables can be easily compared to each other (Freedman, 2009, p. 86).

3. Sentiment analysis methods

3.1. Problem definition

Let's suppose we have a regular opinion defined by Liu (2012): $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$, whereby e_i is an entity of opinion, a_{ij} an entity aspect, s_{ijkl} is the sentiment expressed by person k regarding aspect j of the entity i , h is an opinion holder which send opinion in the time t . By looking at the Liu (2012) quintuple we define the opinion analysis task as: explore all the opinion characteristics in a set of documents d . Thus, we treat sentiment analysis as a sub — area of opinion mining. According to Zobal (2017) sentiment analysis measures people's opinion through natural language processing and computational linguistic. Refining the task, we strive to explore subjectivity (positive, negative, neutral), its direction and intensity (strong-weak) (Haddi et al., 2013; Liu, 2012).

With the sentiment analysis task defined in this way let us pay attention to the special case of tweets level sentiment analysis. Tweet is a 140-character document sent by the registered Twitter users. Access to Twitter resources is possible via Twitter API (rest or streaming) or by scrapping `Twitter.search` web. Each downloaded tweets is stored in JSON format in the dictionary form {'key': 'the value of the key'}. The key is usually information from the following set $K=(\text{text, time, username, replies count, likes count, tweet ID, geolocation})$ (Twitter Developer, 2020).

Denote the set of tweets as D , in which each tweet represent one document $D=(d_1, d_2, d_3, \dots, d_i)$. Due to the length of text, Twitter Sentiment Analysis is treated as sentence level analysis. Therefore, the assumption about regular opinion should be maintained. According to Liu (2012) comparative opinion is impossible to carry out at the sentence level. For each document i information about time and opinion holder is automatically mining and stored as key in JSON file. The proxy variable for entity of opinion is the key word by which Twitter resources are filtered. For stock market relevant data filter is defined as combination cashtag with the company name in format: `Syahoo_finance_abbreviation`. Here the aspect of the entity is omitted, hence the only information about tweet that must be forecast is sentiment (according to Liu's quintuple).

There exist two main approaches towards sentiment analysis: machine learning approach, lexicon-based approach. Second, according to Haddi et al. (2013) depend on a predefined list or corpus of word with a certain polarity,

like WordNet, SentiWordNet or Vader. Machine learning approach (scheme 1) is based on training an algorithm, mostly classification on a selected feature for a specific mission and then test on another set whether it is possible to give a reasonable output (sentiment labels).

Using machine learning approach for the collection of D , lets define C as set of categorical labels $C=(c_1, c_2, c_3, \dots, c_n)$, we are considering a binary classification task so $C=(\text{positive}, \text{negative})$. Let X denote the set of index terms in feature space $X=(x_1, x_2, x_3, \dots, x_j)$, where j denote the number of indexing terms.

3.2. Text representation

Raw text data are incomprehensible for the algorithm, so we are replacing them with the representation that can feed algorithm. Most algorithms expect numerical feature vector of a fixed size. We will use two commonly used representations, n-gram (especially unigram called bag-of-words) and TF-IDF weighing scheme.

Document d_i is represented by a vector $\rightarrow(d_i)$, so D is a set of vectors that are part of common vector space. Each vector consists of j unique components. This space is X -dimensional, documents are identical in terms of features thus information about the word order is lost. By using vectors annotation, we create M matrix representing the full corpus (scheme 2).

The product of the bag-of-word representation is a (j -length) vector containing information about the number of occurrences of unique tokens in the document. Enriching the unigram model with information about neighbour we create n-gram models, where n is the number of words that follow in sequence. The larger the n is we need to use larger training set.

The parameter term frequency — inverse document frequency is calculated according to the formula:

$$tf - idf(t, d) = tf(t, d) \cdot idf(t, d), \quad (2)$$

$$tf(t, d) = \frac{n_{t,j}}{\sum_k n_{k,j}}, \quad (3)$$

$$idf(t, d) = \log \frac{n_d}{1 + df(d, t)}, \quad (4)$$

where:

$tf(t, d)$ — term frequencies obtained are from (3);

$idf(t, d)$ — inverse frequency obtained are from (4);

$n_{t,d}$ — the number of occurrence of term $i(t_i)$ in document $j(d_j)$;

$\sum_k n_{k,j}$ — the number of occurrences of all terms (n_k) in document j ;

n_d — number of documents;

$df(d, t)$ — number of documents that contain word $i(t)$.

Words that appear not only in the document but also dominate the entire corpus D after tf-idf transformation are associated with a low value of weight. These words are not a carrier of useful and distinguishing information. TF-IDF takes the value 0 when feature is absent in the document.

3.3. Pre-processing techniques

Natural language is characterized by a high degree of redundancy so dimensionality reduction to relevant features is first of steps taken for increasing the quality of the feature space. Big social data are more noisy due to (1) Twitter characters (RT, @username, #hashtag, URLs and emoticons), (2) informal nature of communication, words include: slang, spelling errors, abbreviations, neologism, (3) occurrence of stop words or numbers, (4) multilingual tweets and (5) multimodal content.

Pre-processing is the process of normalization of words and removing uninformative tokens. Having a large feature set result in spatial complexity of classifier, also requires a lot of time and RAM to run classifier (Agarwal et al., 2011; Chen & Wójcik, 2016; Symeonidis et al., 2018). Data normalization scenario gives effect in the form of enhancing sentiment analysis (Paudel, 2019; Singh & Kumari, 2016; Uysal & Gunal, 2014). The same authors recommend cautious approach to the choice of pre-processing scenario, there is no universal approach. Scenarios used in the study are listed in table 1. The most commonly used techniques include (Symeonidis et al., 2018):

- deleting and delete numbers;
- detecting and delete repetitions in punctuation;
- detecting and normalize capital letters;
- lowercasing;
- detecting and normalize slang and abbreviation;
- detecting and normalize contraction;
- dealing with negations;
- delete stop words;
- stemming;
- spelling correction;
- removing of punctuation;
- emoticons;
- detecting and deleting twitter characteristics URL, @, #.

3.4. Naïve Bayes Classifiers and its evaluation

The naïve Bayes is a simple probability classifier, often used as a benchmark for another algorithm. Simplicity and computing speed are efficient in verifying the impact of various data processing scenarios. NB is expressed as follow:



$$P(y|x_1, x_2, \dots, x_n) = \frac{P(y)P(x_1, x_2, \dots, x_n|y)}{P(x_1, x_2, \dots, x_n)}, \tag{5}$$

where:

$P(y|x_1, x_2, \dots, x_n)$ — a posteriori probability (conditional probability that y occurs when a specific set of x occurs);

$P(y)$ — a priori probability of class occurrence;

$P(x_1, x_2, \dots, x_n|y)$ — a posteriori probability that x belongs to class y ;

$P(x_1, x_2, \dots, x_n)$ — a priori probability of the occurrence of X , is the same for all classes, hence is omitted.

Thus, the classifier performs the task:

$$\begin{aligned} (y_i) &= \arg \text{MAX}_{y_i} \{P(y|x_1, x_2, \dots, x_n)\} = \\ &= \arg \text{MAX}_{y_i} \{(y)P(x_1, x_2, \dots, x_n|y)\}. \end{aligned} \tag{6}$$

The performance metrics used to evaluate the classifier results are (1) precision, (2) recall and (3) F-measure. In 7–9 formula tp — values of true positive, fp — false positive, tn — true negative, fn — false negative.

$$\text{precision} = \frac{t_p}{t_p + f_p}, \tag{7}$$

$$\text{recall} = \frac{t_p}{t_p + f_n}, \tag{8}$$

$$F\text{-measure} = \frac{2 \cdot \text{precision} + \text{recall}}{\text{precision} + \text{recall}}. \tag{9}$$

4. Results

4.1. Data description

Data from Twitter was scraped via `twitter.search`, for the time window from 01.01.2016 to 31.12.2017. The following keyword queries were used: SFB, SAMZN, SAAPL. As an additional filter a restriction to English tweets was imposed. No geolocation restriction was imposed. The sum of tweet for each company is reported in table 1. In the case of correlation analysis based on big social data, the largest percentage of all tweets should be obtained. By increasing the sample, we aim to better represent the community of online investors.

Tweets are published with an ultra-high frequency, for the purpose of this article the number of messages has been aggregate per day. We made a gen-

eral assumption that the tweet dataset is a good representation, if not complete, of the most popular issues related to the companies. Also, from exploratory data analysis (chart 1) we can conclude that the daily activity of investors in social media reflects the opening days of stock market. This can be treated as the first premise for inference about volume-variable correlation.

Despite the filtering the stream with the cashtag, there is a significant low proportion of tweets with one cashtag for each company (e. g. SAAPL has only 65.5% tweets with one cashtag) (table 2, chart 2). These are affected by informative messages, announcements or SPAM. Due to their nature, they do not contain opinion words, mainly ‘S’, ‘URL’, ‘#’. To exclude them from the mood — volume time series dictionary analysis was performed by using the Vader dictionary from the NLTK module. However, they have a certain level of informativeness, hence the ‘neutral tweets’ time series was included in correlation analysis¹ (chart 3).

Stock market data was obtained from yahoo.finance for the same time period. To facilitate analysis and capture of correlations, stock market data has been filled with weekends and holidays by counting the average by $day_indicator = (day_indicator(t+1) + day_indicator(t-1)) / 2$, where t — time.

4.2. Volume-based analysis result

Algorithm was run for each of the feature transformation. Each transformation was compared with the raw features space. Table 1 presents the scenarios (1–5) which were adopted arbitrary. The higher scenario tends to reduce dimensionality of feature space. Actions have been taken for the BOF and TF-IDF representations.

The training dataset were obtained from (Michailidis, 2017) presented by Go et al. (2009). ‘Sent140’ is a large data set with ‘noisy labels’. It means that they used the Twitter Search API to collect tweets by using keyword search: and as a proxy for sentiment. Tweets with positive emoticons were treated as positive, and tweets with negative emoticons were treated as negative. ‘Noisy’ means that there was no manual data label.

Table 3 shows the effect of the evaluation of the classifier. Parameter values are similar, the difference between 1 and 4 scenarios is visible but the difference is low. This means that a large proportion of the data noise are the characteristics of Twitter from first scenario. Thus, only aggregation of all data cleansing activities results in a change of parameters. This is a conclusion consistent with most references in the literature and training guides. Short messages forced to focus on reducing the impact of ‘#’, URL and ‘S’. Alternatively, to capture the differences, scenarios which are related to one technique should be designed. Thus, the differences clearly will indicate that the actions that improve the quality of the feature space are correct. Another interesting approach

¹ Examples of neutral tweets: ‘there are today’s #block trades #options ...’, ‘Apple watch bands set to...’.

to capturing effects of pre-processing would be use of manually tagged training data set in which high noise of selected problem characteristics would be found.

Table 4 confirms expectation of a positive correlation between volumes variable. Moreover, the supposition that this value would be the largest for apple was confirmed. Although, Pearson correlation coefficient is not truly conclusive for this type of analysis, chart 3 shows a clear violation of the linearity rule which is an assumption of the test. But in most works Pearson coefficient is considered as the initial one. The low difference between the evaluation of the classifiers for scenarios affects the overall conclusions on correlation. Differences in correlation were tried to be shown between the data pre-processing scenarios 1 and 4 (table 4). Actions aimed at improving the quality of features had a positive effect on the value of the Pearson correlation coefficient for a variable that represents negative emotions. This is the result of a significant number of negations in the training set. This is an important application in the domain of behavioral economics/behavioral finance. By increasing the accuracy of the classification (especially due to the multi-label classification), we increase the quality of time series. Thus, we aim to organize the complex space of big social data.

Unfortunately, according to the author, the difference in the value of the correlation coefficient are too low to study the differences in multi regression with standardized beta. Table 5 presents the results of regression analysis (with time series obtained with first pre-processing scenario). In case of SAAPL and SAMZN negative messages play a significant role. As expected, set of neutral messages is not significant. The parameter value for the negative variable is greater than for positive messages — companies must be wary of negative comments on social media.

5. Conclusion

Correlation between social media sentiment and stock market via trading volume variable was captured. Results are strong enough to recommend research to forecast stock market by using Twitter variables. This idea is based on approach of behavioural economics, with profoundly link sentiment with individual decision making and behaviour. The presence of uncertainty in stock market, overall effect of sentiment on decision-making nowadays tends to be even stronger. Long lasting positive sentiment might lead to overvaluation and even to market bubbles with market pessimism makes stocks undervaluation, with might open space for purchase at advantages prices. That is why it is important to look for variables that will help forecast changes and react in real time. These criteria meet big social data. This study proposes a method of combining conclusions from the area of research on the development of Twitter Sentiment Analysis with analyses in the domain of behavioral economics. Attempts were made to point out that the quality of conclusions for behavioral economics depends on the quality of the analyses in Twitter Sentiment Analysis. As a result of the analysis, certain signals were received that confirm this hypothesis. In-

creasing the quality of features affects the quality of applications in research in which variables from Twitter are used. However, this is a signal, this study should be developed.

References

- Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. In M. Nagarajan, & M. Gamon (Eds.), *LSM'11: proceedings of the workshop on languages in social media*. Stroudsburg: ACL.
- 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). doi:10.1111/j.1540-6261.2004.00662.x.
- Bollen, J., Mao, H., & Pepe A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In N. Nicolov, & J.G. Shanahan (Eds.), *Proceedings of the fifth international AAAI Conference on weblogs and social media*. Barcelona: AAAI.
- Chen, E.E., & Wojcik, S.P. (2016). A practical guide to big data research in psychology. *Psychological Methods*, 21(4). doi:10.1037/met0000111.
- Freedman, D.A. (2009). *Statistical models: theory and practice*. Leiden: Cambridge University Press.
- Go, A., Bhayani, R., & Huang, L. (2009). *Twitter sentiment classification using distant supervision*. Retrieved 01.04.2020 from <https://www-cs.stanford.edu>.
- Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science*. 17. doi:10.1016/j.procs.2013.05.005.
- Ishikawa, H. (2015). *Social big data mining*. Boca Raton: CRC Press. doi:10.1201/b18223.
- Liu, B. (2012). *Sentiment analysis and opinion mining*. San Rafael: Morgan & Claypool Publishers.
- Mao, Y., Wei, W., Wang, B., & Liu, B. (2012). Correlating S&P 500 stocks with Twitter data. In X. Fu, P. Gloor, & J. Tang (Eds.), *Proceedings of the first ACM international workshop on hot topics on interdisciplinary social networks research*. New York: ACM. doi:10.1145/2392622.2392634.
- Mittal, A., & Goel, A. (2012). *Stock prediction using twitter sentiment analysis*. Retrieved 01.04.2020 from <http://cs229.stanford.edu>.
- 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(2). doi:10.1016/j.jfds.2017.11.002.
- Oh, C., & Sheng, O. (2011). Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. In D.F. Galletta & T.P. Liang (Eds.), *Proceedings of the international conference on information systems*. Atlanta: AIS.
- Olshannikova, E., Olsson, T., Huhtakamäki, J., & Kärkkäinen, H. (2017). Conceptualizing big social data. *Journal of Big Data*, 4(1).doi:10.1186/s40537-017-0063-x.

- Paudel, S., Prasad, P.W.C., Alsadoon, A., Islam, M.R., & Elchouemi, A. (2019). Feature selection approach for Twitter sentiment analysis and text classification based on Chi-Square and Naïve Bayes. In J. Abawajy, K.R. Choo, R. Islam, Z. Xu, & M. Atiquzzaman (Eds.), *International conference on applications and techniques in cyber security and intelligence ATCI 2018: applications and techniques in cyber security and intelligence*. Cham: Springer. doi:10.1007/978-3-319-98776-7_30.
- Porshnev, A., Lakshina, V., & Redkin, I. (2016). Could emotional markers in Twitter posts add information to the stock market ARMAX–GARCH Model. *Higher School of Economics Research Paper*, 54/FE/2016. doi:10.2139/ssrn.2763583.
- Rao, T., & Srivastava, S. (2013). Modeling movements in oil, gold, forex and market indices using search volume index and Twitter sentiments. In H. Davis, H. Halpin, & A. Pentland (Eds.), *WebSci'13: Proceedings of the 5th annual ACM web science conference*. New York: ACM. doi:10.1145/2464464.2464521.
- Singh, T., & Kumari, M. (2016). Role of text pre-processing in Twitter sentiment analysis. *Procedia Computer Science*, 89, 549. doi:10.1016/j.procs.2016.06.095.
- Strauß, N., Vliegenthart, R., & Verhoeven, P. (2018). Intraday news trading: the reciprocal relationships between the stock market and economic news. *Communication Research*, 45(7). doi:10.1177/0093650217705528.
- Symeonidis, S., Effrosynidis, D., & Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications*, 110. doi:10.1016/j.eswa.2018.06.022.
- Uysal, A.K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management*, 50(1). doi:10.1016/j.ipm.2013.08.006.
- Wysocki, P.D. (1999). Cheap talk on the web: the determinants of postings on stock message boards. *University of Michigan Business School Working Paper*, 98025. doi:10.2139/ssrn.160170.
- Zhang, X., Fuehres, H., & Gloor, P.A. (2011). Predicting stock market indicators through Twitter ‘I hope it is not as bad as I fear’. *Procedia: Social and Behavioral Sciences*, 26. doi:10.1016/j.sbspro.2011.10.562.
- Zobal, V. (2017). *Sentiment analysis of social media and its relation to stock market*. Unpublished bachelor thesis, Charles University, Prague. Retrieved 01.04.2020 from <https://is.cuni.cz>.
- Tweeter Developer. (2020). Retrieved 01.04.2020 from <https://developer.twitter.com>.
- Michailidis, M. (2017). *Sentiment 140 dataset with 1.6 million tweets*. Retrieved 01.04.2020 from <https://www.kaggle.com>.



Acknowledgements

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

Funding: this research was funded by the Nicolaus Copernicus University in Torun, Faculty of Economic Sciences and Management statutory sources.



Appendix 1: framework for sentiment analysis

Table 1.
Pre-processing scenarios

Cleaning scenario	Action on dataset
1	lowercase, stopwords, stemming, remove punctuation, Twitter characteristic
2	1 + delete numbers
3	2 + slang and abbreviation, spelling correction
4	3 + contraction, negation, emoticons (all actions)

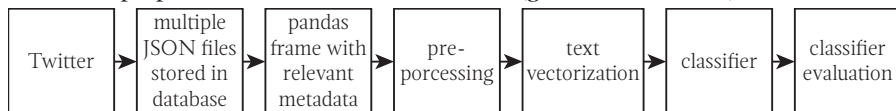
Notes:

All scenarios were carried out for bag-of-words and TF-IDF model.

Source: Own preparation.

Scheme 1.

Process of preparation documents for classifiers (general framework)



Source: Own preparation.

Scheme 1.

Document–Term matrix

$$\begin{array}{c}
 \text{documents} \\
 M = \begin{array}{ccc}
 t_{11} & \dots & t_{1j} \\
 \vdots & \dots & \vdots \\
 t_{i1} & \dots & t_{ij}
 \end{array} \\
 \text{indexing terms}
 \end{array}$$

Source: Own preparation.

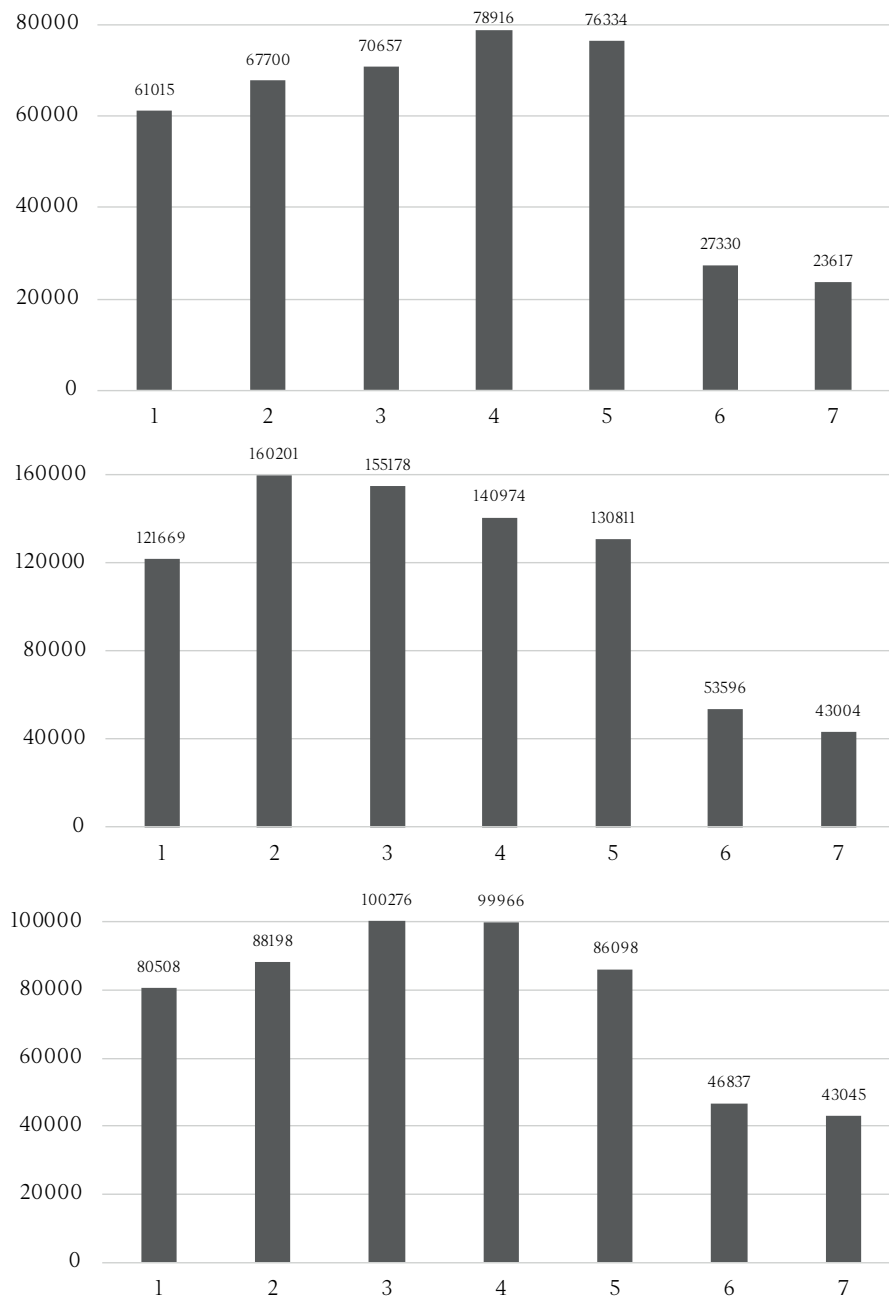
Appendix 2: exploratory data analysis

Table 2.
The number of tweets in the dataset for companies and percentage of noise in the cashtag

Variable	SAAPL	SFB	SAMZN
number of tweets	808218.00	5474190.00	405758.00
daily average tweets	1109.40	750.59	557.11
cashtag=1.0 (in %)	65.60	52.31	50.50
cashtag<5.0 (in %)	84.11	80.83	77.07

Source: Own preparation.

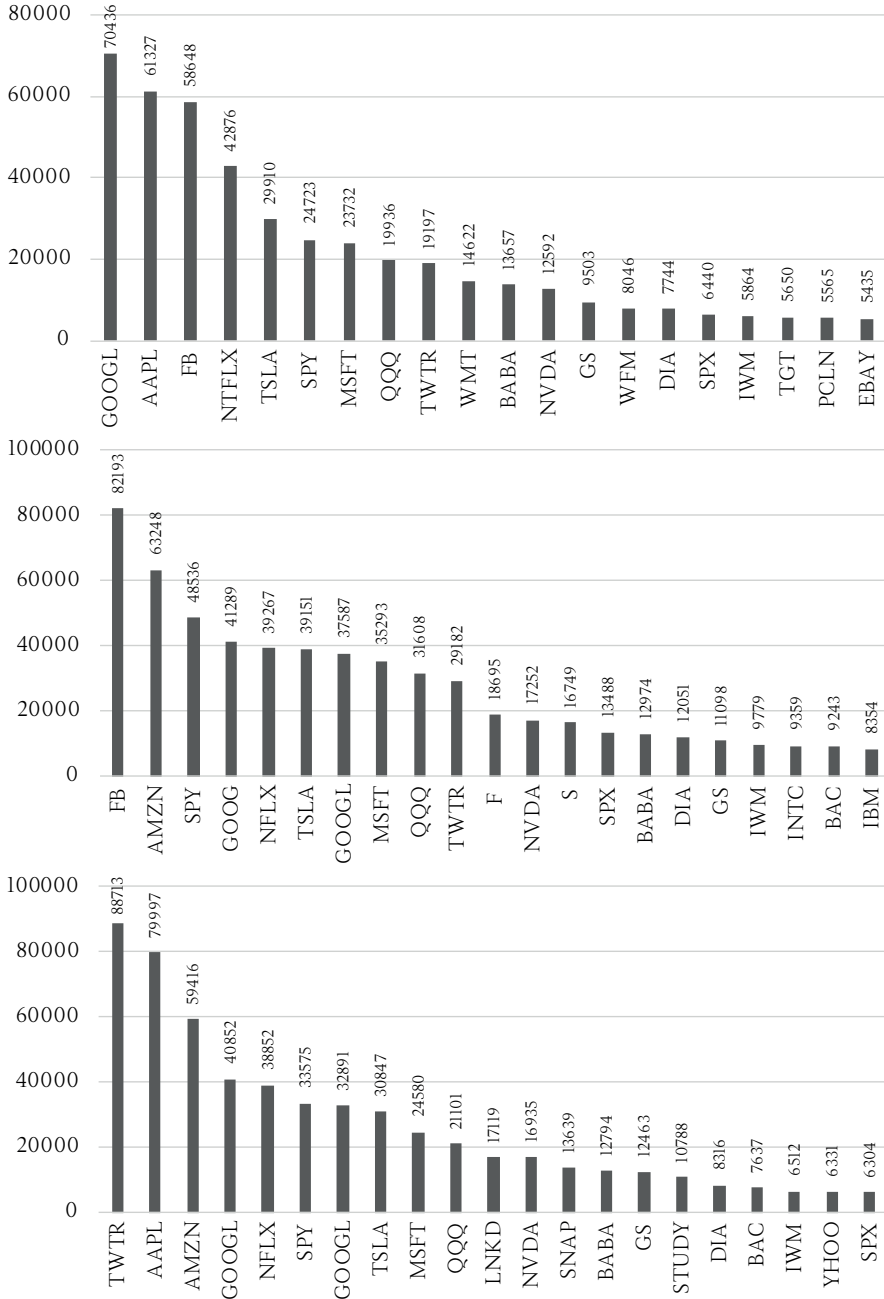
Chart 1.
Investor activity broken down by weekdays Amazon, Apple and Facebook



Source: Own preparation in Python.



Chart 2.
Co-occurrence of cashtags for Amazon, Apple and Facebook



Source: Own preparation in Python.

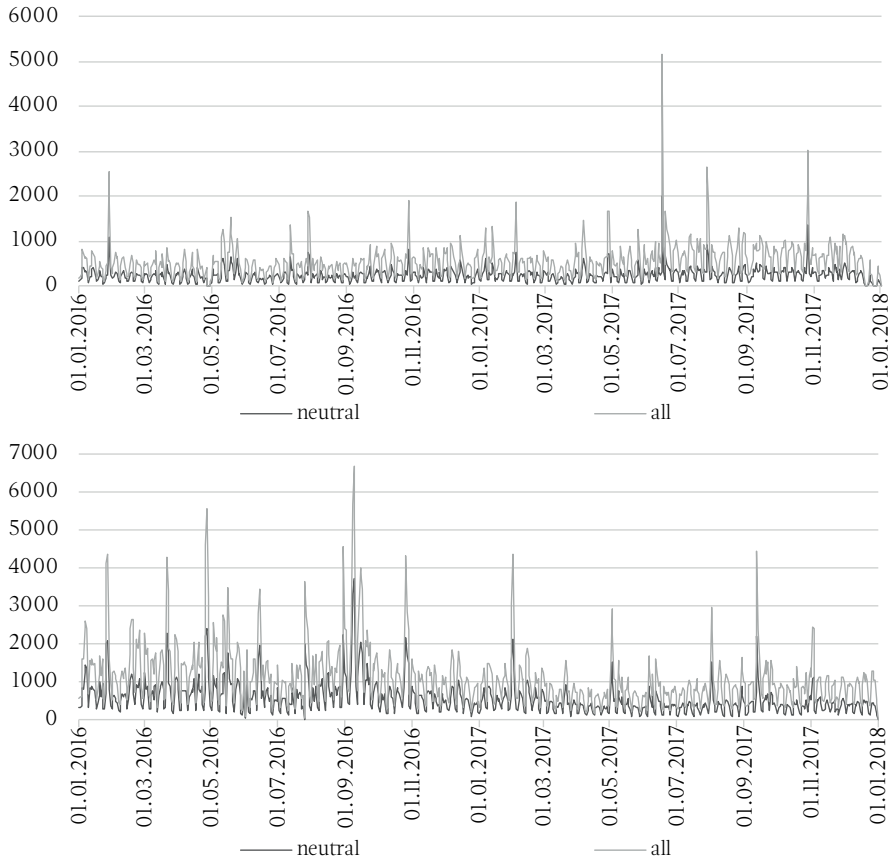
Appendix 3: Twitter time series and classification results

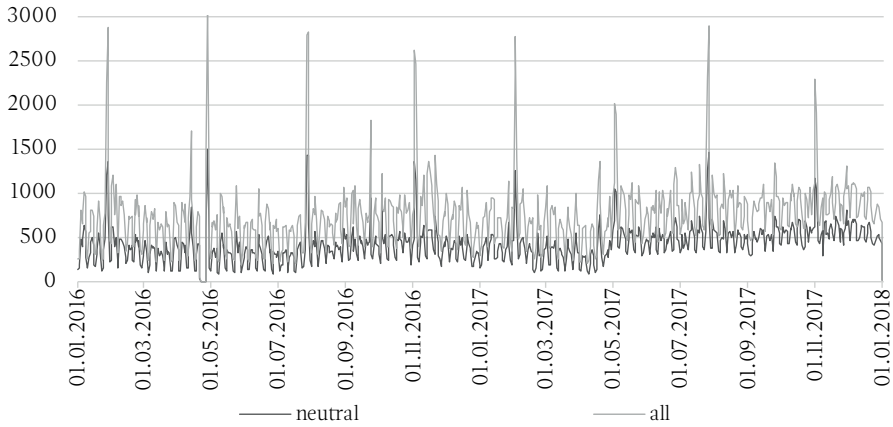
Table 3.
Evaluation of the Multinomial Naïve Bayes classifier

Scenarios	TF-IDF				BOW			
	accuracy	precision	recall	f-measure	accuracy	precision	recall	f-measure
1	0.761	0.769	0.750	0.758	0.775	0.782	0.762	0.772
2	0.768	0.773	0.758	0.766	0.776	0.783	0.762	0.772
3	0.764	0.690	0.740	0.700	0.785	0.766	0.759	0.779
4	0.871	0.960	0.894	0.876	0.871	0.752	0.910	0.780

Source: Own preparation.

Chart 3.
All daily tweets and neutral tweets for Amazon, Apple and Facebook (calculated using the lexicon-based approach by Vader in NLTK module)





Source: Own preparation in Python.

Appendix 4: correlation between volume of tweets and trading volume

Table 4.
Correlation indicator for companies for scenario 1 and scenario 4

Variable	Scenario 1				Scenario 4			
	<i>all_day</i>	<i>neutral</i>	<i>negative</i>	<i>positive</i>	<i>all_day</i>	<i>neutral</i>	<i>negative</i>	<i>positive</i>
Apple volumen	0.4998	0.4943	0.5310	0.4820	0.5203	–	0.5447	0.4823
Facebook volumen	0.3600	0.3175	0.4244	0.4062	0.3754	–	0.4277	0.4062
Amazon volumen	0.3760	0.3738	0.3710	0.3495	0.3776	–	0.3821	0.3495

Source: Own preparation.

Table 5.
Regression result for volume-based analysis

Model	R	neutral	negative	positive
Apple	0.3080	–0.0945	0.4774***	0.1977***
Amazon	0.1510	0.1534	0.2380*	0.0088
Facebook	0.1094	–0.3980***	0.2592***	0.2937***

Source: Own preparation.

