ISSN 1507-3858
e-ISSN 2450-0003

Ewelina Niedzielska
University of Lodz, Faculty of Management
e-mail: ewelina.niedzielska@wz.uni.lodz.pl
ORCID: 0000-0002-2919-3200

USING TEXT ANALYSIS FOR EVALUATING THE BEHAVIOUR OF RATES OF RETURN FROM THE WIG20 INDEX

WYKORZYSTANIE ANALIZY TEKSTU DO OCENY ZACHOWANIA STÓP ZWROTU INDEKSU WIG20

DOI: 10.15611/ie.2020.1.05
JEL Classification: G14

Quote as: Niedzielska, E. (2020). Using text analysis for evaluating the behaviour of rates of return from the WIG20 index. *Informatyka Ekonomiczna. Business Informatics*, (1).

Abstract: The aim of the article is to indicate the possibility of using text analysis for the research of dynamics of the Polish capital market. The first part of the article notes the changes which took place within the data market in recent years and their impact on the discounting of information by stock market investors. The Information Effectiveness Hypothesis and the paradigm of behavioural finance are the basis of theoretical considerations. The second part presents the result of a study, the objective of which was to build an algorithm allowing the prediction of the WIG20 index rates of return based on text data. The test sample consisted of 901 papers published in the "Parkiet" magazine and 748 daily rates of return. The study was conducted using algorithms processing natural language and decision trees used for classification. The results of the study allowed an indication of a model in which the precision and accuracy indicators exceeded a score of 50%.

Keywords: capital market, WIG20, text analysis.

Streszczenie: Celem artykułu jest wskazanie możliwości wykorzystania analizy tekstu do badań nad dynamiką polskiego rynku kapitałowego. W pierwszej części artykułu opisano zmiany, jakie w ostatnich latach nastąpiły na rynku danych, oraz ich wpływ na dyskontowanie informacji przez inwestorów giełdowych. U podstaw teoretycznych rozważań leży klasyczna hipoteza efektywności informacyjnej oraz paradygmat finansów behawioralnych. W drugiej części artykułu przedstawiono wyniki badania, którego celem było zbudowanie algorytmu umożliwiającego przewidywanie stóp zwrotu indeksu WIG20 na podstawie danych tekstowych. Próba badawcza składa się z 901 artykułów czasopisma *Parkiet*, które zawierały w tytułach sformułowanie „WIG20", oraz 748 dziennych stóp zwrotu. Badanie przeprowadzono z wykorzystaniem algorytmów przetwarzania języka naturalnego oraz drzew decyzyjnych. Wskazano model, którego wskaźniki precyzji i dokładności przekraczały 50%.

Słowa kluczowe: rynek kapitałowy, WIG20, analiza tekstu.

1. Introduction

From the point of view of financial markets, the applications and algorithms used for text analysis create an opportunity to use and cope with a ubiquitous flood of information and large amounts of data called 'Big Data'(Chan and Chong, 2017, p. 53). For stock market investors the current status quo means that they have to discount more and more communication flowing into the market and do that increasingly faster, i.e. competently select received content. As a result, decision processes are put in motion in which cognitive simplifications and emotion play a large role. Behavioural financial analysts pay particular attention to this fact. They reduce the knowledge about investor activity from a normative plane to a descriptive plane while trying to examine and explain the real – not optimal – means of conduct of the participants of financial markets.

This paper is an attempt to take part in a discussion regarding the possibility of using text data for analyses within the scope of the capital market. Based on the review of literature made by the author one may conclude that this subject matter has been very popular in recent years. The key stream seems to be the research of the rates of return from stocks of exchange companies and an attempt of a prediction of their changes. The added value of the article consists in conducting an analogous research not on particular companies, but on the entire index. Although the literature in this field can be described as much poorer, it is worth noting that this is not the first such attempt which to be described.

The article consists of three parts. The first presents the theoretical and practical bases leading the author to research this area. This concerns the dynamic development of the data market, as well as the classic finance problem of information effectiveness and discounting information by stock market investors. In the second part the author describes the conducted study, i.e. sources of data, used tools, the research procedure and the obtained empirical results. The final part is a summary and indication of potential further directions of research.

2. Justification of undertaking research

Recent years have witnessed a previously unseen increase of data. As a result of connecting many everyday use devices to the global network, a constant stream of produced and consumed information is generated (Cavanillas, Curry, and Wolfgang, 2016, p. 3). The dynamics of this market has become so intense that since 2016 its development has been measured at a two-digit pace and its estimated global monetary value in 2019 amounted to 26 billion USD (Prajsna and Sawa, 2018, p. 3). Within the scope of this economic subject, this issue became so noticeable that the term ‘data economics’ was formulated. The term can be defined as the widely understood – both direct and indirect – influence of the data markets and the processes related to it (generating, storage, processing, analyses, distribution etc.) on the economy (Pepato and Micheletti, 2019, p. 88).

In this context the noteworthy fact is that this creates a particular chain within which a person becomes both the supplier and the recipient of data, and this process may influence his or her actions and decisions. At the same time, while observing global literature regarding capital markets, one may note that many researchers are paying increasing attention to the paradigm of behavioural finance. As a result of this shift, the crucial role of non-financial factors as such, which may determine the reactions of investors to the information flowing into the market is indicated. These problems fit into the ‘hypothesis of information effectiveness’ which is a classic concept for finance and within which questions regarding the means and pace of discounting knowledge by stock market participants are raised (Fama, 1991).

The commonness of data is also related to its structural heterogeneousness which is the result of the diversity of the sources of its origin (e.g. social media, data sensors, websites, reports etc.) (Cavanillas et al., 2016, p. 51). In this context the classification of data ranges from structured through semi-structured to unstructured. From the point of view of business as well as the capital market, the fact that more and more data processed in these areas of life are, in spite of appearances unstructured, becomes characteristic (Pepato and Micheletti, 2019, p. 84). This takes the form of images, sounds or finally text (Manyika et al., 2011, p. 33). This is a result of a visible trend of expanding the information duties of stock market companies as they are obliged to more and more extensively report their activity, both through the financial and social prism. Additionally, investors and stockholders have free access to press articles and websites belonging to the trade where the actions and decisions of the boards of companies are being continuously commented on and interpreted.

At the same time, the development of the data market involves the necessity of building new tools allowing data processing. In the context of unstructured data this involves cognitive systems understood as using the technologies of machine learning and artificial intelligence to process natural language (Reinsel, Gantz, and Rydning, 2017, p. 4). The purpose of this is to separate different types of information from text data, allowing its deeper exploration, for example by determining the emotional characterization of the author of the article (Cavanillas et al., 2016, p. 270). The most commonly used term describing such measures is ‘text analysis’ which can be defined as “processing textual language in a digital form on a large scale in order to separate data transformed into useful quantitative or qualitative information” (Reinsel et al., 2017, p. 4). Although extracting data which have a pragmatic value is a multistage and very time-consuming process (Cavanillas et al., 2016, p. 109), this problem – due to its essentiality and topicality – seems to be worth noting since according to the estimates of the International Data Corporation until the year 2025 the amount of data related to cognitive systems is predicted to reach the magnitude of 1.4 zetabyte (Reinsel et al., 2017, p. 4).

2.1. Using text analysis in capital market research

In order to fully understand the problem of using text analysis in research regarding the capital market, it is worth highlighting three issues: the sources of text data, the research subject and the results which may indicate the legitimacy of including this type of information to the decision processes of the investors.

Used sources of text data can be divided into three categories. The first is the data acquired from official documents of the companies, e.g. interim statements (Butler and Kešelj, 2009) or current statements (Groth and Muntermann, 2011). The advantage of this data source is the fact that the authors of these documents usually have the broadest and discerning knowledge about the enterprise (Feng, 2010).

The second category comprises press articles and information originating from websites. The value of these sources is their amount and accessibility which has its own meaning both from the point of view of investors and researchers. At the same time, there are difficulties related to the discernment of the opinions of experts and the facts which they present. It must be stressed that press articles are relatively quite popular in this research stream (Tetlock, Saar-Tsechansky, and Macskassy, 2008).

The third category is social media. The usefulness of such data results from the fact that the number of persons who post on the Internet, also in regard to the capital market is uncommonly high (Kearney and Liu, 2014, p. 174). It should be pointed out that such activity is characteristic of most of all individual investors (Bukovina, 2016, p. 20). One cannot assume that data extracted from this source will be representative of the reactions of the entire market. Undoubtedly, however, in this context the mentioned interaction between a person and data presents itself in the strongest way.

The de facto type of the analyzed market is accepted by the author as the research subject. The techniques based on working with large data sets found a broad application in the scope of financial analysis (Chun and Kim, 2004, p. 131). Numerous researchers concurrently point out that text analysis is a promising trend in this area. This type of data has proved useful in predictions of prices of particular companies (Boudoukh, Feldman, Kogan, and Richardson, 2013; Chen, De, Hu, and Hwang, 2014; Heston and Sinha, 2016) as well as a broader market (Tetlock, 2007; Dougal, Engelberg, García, and Parsons, 2012; Dzielinski and Hasseltoft, 2012).

Based on a review of literature, one may conclude that analyses regarding stock market companies in the short-term scope are of the greatest interest. A much less popular area of using text data is considering the rates of return from market indices.

This problem may nevertheless be treated as essential considering the existence of financial instruments based on the dynamics of broad portfolios. This entails derivatives (futures contracts and options) and investment funds, especially ETF (Exchange Traded Fund), among others.

One of the first studies in this scope was published as early as 1998 by Wutrich et al. The authors managed to build a model which allowed an automatic drawing, classifying and prediction of the dynamics of the following indices: Dow Jones, FTSE100, Nikkei225, Hang Seng and Singapore Straits. Finally, the model was to recommend buying, selling or no action in regard to each of the analysed indices. However, the decision accuracy of the model was below 50% (from 40% to 46.7% depending on the index) (Wuthrich, Cho, Leung, and Zhang, 1998). Another important study was published by Tetlock in 2007, concentrating on one index, the Dow Jones Industrial Average. The author managed to reject the hypothesis of no relation between press content (the data was drawn from The Wall Street Journal) and the capital market. He also suggested that a crucial factor which may determine the rates of return from the index a day after the announcement is the pessimism in them. The author also proved that a high level of this factor prognosticates a down-trading in the market prices and in the next phase their reversion to the fundamental price. The change in the level of pessimism by one standard deviation influenced the index to change by 8.1 base points on the following day (Tetlock, 2007). Zubair and Cios drew analogous conclusions in 2015. They verified the hypothesis of the existence of a correlation between text sentiment (the data was drawn from the Reuters archive) and the results of the S&P500 index. The cited authors noticed that this relation exists mainly in the case of negative sentiment. An analogous property was not found for positive sentiment (Zubair and Cios, 2015).

Similar analyses were conducted on European markets. In 2018 Feuerriegel and Gordon presented the results of research conducted on the DAX, CDAX and STOXX indices. The text data, differently from the abovementioned research, did not originate from journals but from the financial statements of the companies (the period 1997-2015). The aim of their research was predictive. The authors verified that integrating unstructured data to the model will allow for a higher level of prognosis accuracy. The results of this research cannot be deemed conclusive but the authors managed to indicate situations in which such action would lower the prognosis errors in the long-term (Feuerriegel and Gordon, 2018).

Text analysis was also used in Poland. In 2017 Rostek and Młodzianowski attempted to use text data to prognose the movement of the WIG, WIG20, mWIG40 and sWIG80 indices. The texts were classified by searching for keywords defined by the authors. As an example, for the positive sentiment these were: boom, green, profit, and for the negative sentiment: fall, red, loss. The results of the research indicated a classification accuracy of the movement of indices (rise or fall) ranging from 61% and 68%. In the case of a prognostic model, the highest result did not exceed 60% (Rostek and Młodzianowski, 2017) . Searching for keywords, due to its simplicity nevertheless seems to be an interesting concept, possibly for a relatively simple practical application. A similar conclusion was drawn by Kavšek in 2017, however assuming a reverse logic than Rostek and Młodzianowski. The aim of his research was to detect words in press articles which would be correlated with the movement of the Dow Jones Industrial Average index. As a result, a list of expressions related to rises and falls of the index was built (Kavšek, 2017).

Nevertheless, it is worth noting that there is a wide variety of tools and methods supporting the solution of problems related with working on unstructured data sets. Nowadays, one of the more important fields regards machine learning algorithms and artificial intelligence, and therefore further exploration of the research apparatus in this field should be considered. The research conducted in 2010 by Kara et al., indicates that following this path may bring a result in the form of a significant improvement of prediction abilities of prognosis models. The subject of their research was the National100 index of the Istanbul stock market. For comparative purposes two classification algorithms were used: ANN (Artificial Neural Networks) and SVM (Support Vector Machines). The results of the research allowed them to obtain results exceeding 70%. It was also noted that using the ANN classifier had a higher predictive accuracy (75.74%) than in the case of SVM (71.52%) (Kara, Boyacioglu, and Baykan, 2011). Analogous research was conducted in later years, indicating the legitimacy of using an algorithmic approach to predictions of the dynamics of the capital market, including stock market indices (Guresen, Kayakutlu, and Daim, 2011; de Oliveira, Nobre, and Zárate, 2013). Unfortunately, this analyses did not include text variables into their models. Nonetheless, an attempt will be made of using such data as well as machine learning algorithms for predictions of the rate of return from the stock market index.

3. Research description and results

The aim of the conducted research was to build an algorithm allowing to make a prediction of the rates of return from the WIG20 index. Thus the research was based on the following hypothesis: text data is a carrier of information regarding the changes of the rates of return of the studied index.

The schematics of the research were divided into the stages below:

1) collecting text and financial data,
2) cleaning text data,
3) calculating the daily rates of return from the WIG20 index,
4) transforming text data into input data of the machine learning algorithm,
5) integrating text data with financial data,
6) building machine learning algorithms,
7) obtaining the research results and their evaluation.

3.1. Collecting text and financial data

Two types of data were used in the research: text and financial, covering the period from 1 January 2016 to 31 December 2018. The text data consisted of 901 press articles published in the "Parkiet" magazine[1]. The articles were downloaded from the EMIS database (Emerging Markets Information Service).

Table 1. Description of text data

Specification		Data
Number of articles		901
Number of days on which the following number of articles were published	two articles	206
	three articles	48
	four articles	8
Number of days without trading session on which articles were published		97

Source: own study.

The financial data contained the opening and closing prices of the WIG20 index. Within the studied period of time, 749 exchange listings and 347 days took place without a trading session. The articles from the trading session days and days without a trading session were considered.

3.2. Cleaning text data and calculating daily rates of return

The process of cleaning data consisted of eliminating the so-called stop words which are the most common Polish words that do not have information value in the context of the functioning of algorithms (e.g. that, and, or) as well as the combination of dyadic expressions into single expressions (e.g. PKN Orlen into PKNOrlen). Additionally, using the *Słownik gramatyczny języka polskiego* (Polish Grammar Dictionary) (Saloni, Woliński, Wołosz, Gruszczyński, and Skowrońska, 2012) the process of lemmatization was conducted, consisting in a reduction of particular words into their basic form (e.g. all verbs into infinitive forms). This action was intended at reducing the multidimensionality of qualities and is a common measure while working with text data. The collected research sample allowed a calculation of 748 daily rates of return. Their basic statistics are presented in Table 2 below.

Table 2. Descriptive statistics for daily rates of return for the WIG20

Specification	Average	Min.	Max.	Standard deviation	Variance
Data	0.03%	-2.79%	5.58%	0.0101	0.0001

Source: own study.

[1] According to the study conducted in 2018 by (Polish: Stowarzyszenie Inwestorów Indywidualnych (Association of Individual Investors), "Parkiet" is in fourth place (of eighteen) within Internet websites which are the source of information regarding company activity.

3.3. Transforming text data into input data of the machine learning algorithm

In the next stage of the study the corpus of the text data was transformed into two matrices being the input data for the machine learning algorithm. The first matrix (M1) was built by an automatic calculation of the frequency of appearance of each word in the core. The tf-idf (Term Frequency – Inverse Document Frequency) index was used for the second matrix (M2) allowing to present the frequency of appearance of a particular word in individual documents in one numerical value, as well as the number of documents containing this word in the entire corpus. Owing to this, the importance of particular words for the entire collection of the analysed texts could be determined. In the case of days on which a number of articles were published, they were treated as one document.

3.4. Integration of text data with financial data

After rectifying the data, a reduction of their multidimensionality and their conversion, the integration of both types of data was performed by assigning corresponding rows of both matrices and the results of the rates of return of the WIG20 index from the previous quotation to particular rows.

3.5. Building machine learning algorithms

Two decision tree algorithms were built, differentiated by the input data matrix (M1 and M2). The entire sample was successively divided into the training part (80%) and the test part (20%). Based on the training data a learning process was conducted, the aim of which was the choice of variables by the algorithm allowing the correct assignment of the rates of return to one of two classes: positive or negative. The prediction ability of both models was verified on test data. The algorithms were written in the Python language and the XGBoost library was used for their creation.

3.6. Obtaining the research results and their evaluation

The results of the analysis indicated a higher classification ability of the model in which the input data was built by calculating the number of words, since the M2 model manifested high errors of the first order both in the context of positive and negative rates of return. Based on the achieved results of the recall index, one may conclude that it was burdened with a high level of error of the second order. The results of the precision and recall indexes both for negative and positive rates of return, oscillated within 50%. From the point of view of making investment decisions which could be potentially made based on such analyses, this model can be described as undependable.

The M1 model presented in turn a relatively high level of errors of the first order in the case of negative rates of return and an error of the second order in the case

Table 3. Results of the predictions of the classification algorithms

Input	Class	Precision	Recall	f1-score
M1	0	0.56	0.95	0.71
	1	0.83	0.26	0.40
	accuracy	0.61	0.61	0.61
M2	0	0.51	0.57	0.54
	1	0.49	0.43	0.45
	accuracy	0.50	0.50	0.50

Precision – truly positive / (truly positive + falsely positive); recall – truly positive / (truly positive + falsely negative); f1-score – 2 * (precision * recall) / (precision + recall); accuracy – (truly positive + truly negative) / (truly positive + truly negative + falsely positive + falsely negative).

Source: own study.

of positive rates of return. Both errors negatively influenced the level of the f1-score and the accuracy indexed. However, in the case of this model the recall indexed for negative rates of return and precision for positive rates of return amounted to 0.95 and 0.83, respectively. This means that while using it, a relatively small number of rates of return were classified as positive, which considerably discerns both models. The presented results may be the premise to conclude that using text data may allow a formation of models used for predictions of the rates of return from the WIG20 index. At the same time, one may conclude that using simple input data, i.e. matrices of the frequency of appearance of words in the text corpus allows for achieving more satisfying results than using the tf-idf index for this purpose.

4. Conclusion

This paper fits into a relatively new research stream within which attempts of analysis and predictions of the dynamics of the capital market based on a specific type of unstructured data which is a text, are made. The paper provided information indicating the topicality and essentiality of the handled problems. On the one hand the intense development of the data market can be included in them, on the other the notable interest of researchers in the paradigm of behavioural finance. The described issues can be included in the context of the hypothesis of information effectiveness within which the means of discounting information by the market participants are being reflected upon.

The study indicate a model allowing the achievement of the results of prediction exceeding a classification accuracy of 50% in most cases. Within the analysis, two models were built differentiated by a different means of processing the text data, which were separate input data. In the first model (M1), the information about the frequency of appearance of particular words in the corpus was used, whereas the other (M2) was based on the results of the tf-tdf index. The text data, along with the calculated rates of return from the WIG20 index, were divided into the training and test group (80% and 29%, respectively). Next the models were trained in classification of the rates of return one day in advance, and divided into positive and negative by using the decision tree algorithm. The evaluation of the results was made by estimating the level of the precision, recall, f1-score and accuracy indexes. Based on this, one may conclude that from the point of view of making investment decisions, the M1 model allowed more satisfying results than the M2 model. The main premise for such a conclusion is the fact that by using this model a small number of negative rates of return were classified as positive.

It is noteworthy that one classification algorithm was used in the study. A review of literature indicates that other classifiers, such as the naïve Bayes classifier, SVM or neural networks can be used for the same purpose. Undoubtedly, conducting an analogous research using other classification methods will bring a more complex answer to the question of the possibility of using text data to predict the dynamics of the Polish capital market. Additionally, although the chosen source of data (i.e. "Parkiet") can be considered as one of the more important for stock market investors, including other magazines or branch websites would enable to increase the training group leading to a presentation of more decision variables to the models. Nonetheless, the review of literature and the results of the author's own study seem to justify the conclusion that from a cognitive point of view, it is worth continuing research in the indicated field.

References

Boudoukh, J., Feldman, R., Kogan, S., and Richardson, M. (2013). *Which news moves stock prices? A Textual Analysis* (NBER Working Paper Series, 18725), 1-45.

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

Butler, M., and Kešelj, V. (2009). Financial forecasting using character N-Gram analysis and readability scores of annual reports. In Y. Gao, and N. Japkowicz (Eds.), *Advances in Artificial Intelligence* (pp. 39-51). Canadian AI 2009. Lecture Notes in Computer Science, 5549. Berlin, Heidelberg: Springer.

Chan, S. W. K., and Chong, M. W. C. (2017). Sentiment analysis in financial texts. *Decision Support Systems*, (94), 53-64.

Cavanillas, J. M., Curry E., and Wolfgang W. (Eds.). (2016). *New horizons for a data-driven economy. A roadmap for usage and exploitation of Big Data in Europe.* Springer Open, Switzerland.

Chen, H., De, P., Hu, Y., and Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media*. Review of Financial Studies*, *27*(5), 1367-1403.

Chun, S. H., and Kim S. H. (2004). Data mining for financial prediction and trading: Application to single and multiple markets*. Expert Systems with Applications*, *26*(2), 131-39.

Das, S. R. (2014). *Text and context: Language analytics in Finance.* Foundations and Trends® in Finance, (8), 145-261.

De Oliveira, F.A., Nobre, C. N., and Zárate, L. E. (2013). *Applying Artificial Neural Networks to prediction of stock price and improvement of the directional prediction index – Case study of PETR4, Petrobras, Brazil*, Expert Systems with Applications, *40*(18), 7596-7606.

Dougal, C., Engelberg, J., García, D., and Parsons, C. A. (2012). Journalists and the stock market. *Review of Financial Studies*, *25*(3), 640-679.

Dzielinski, M., and Hasseltoft, H. (2012). Aggregate news tone, stock returns, and volatility. *SSRN Electronic Journal*. (Working Paper), University of Zurich.

Fama, E. (1991). Efficient capital markets: II. *The Journal of Finance*, *46*(5), 1575-1617.

Feng, L. (2010). Textual analysis of corporate disclosures: A survey of the literature. *Journal of Accounting Literature*, (29), 143-165.

Feuerriegel, S., and Gordon, J. (2018). Long-term stock index forecasting based on text mining of regulatory disclosures. *Decision Support Systems*, (112), 88-97.

Groth, S. S., and Muntermann, J. (2011). An intraday market risk management approach based on textual analysis. *Decision Support Systems*, *50*(4), 680-691.

Guresen, E., Kayakutlu, G., and Daim, T. U. (2011). Using artificial neural network models in stock market index prediction. *Expert Systems with Applications*, *38*(8), 10389-10397.

Heston, S. L., and Sinha N. R. (2016). News versus sentiment: Predicting stock returns from news stories. *Finance and Economics Discussion Series*, 1-35.

ISI Emerging Markets. (n.d.). *Emerging Markets Information Service*. Retrieved August 8, 2019 from www.emis.com/pl

Kara, Y., Boyacioglu, M., and Baykan, Ö. (2011). Predicting direction of stock price index movement using artificial neural networks and support vector machines: The sample of the Istanbul Stock Exchange. *Expert Systems with Applications*, *38*(5), 5311-5319.

Kavšek, B. (2017). Using words from daily news headlines to predict the movement of stock market indices. *Managing Global Transitions*, *15*(2), 109-121.

Kearney, C., and Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, (33), 171-185.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity.* McKinsey Global Institute.

Pepato, C., and Micheletti, G. (2019). Second Interim Report The European Data Market Monitoring Tool: Key Facts & Figures, First Policy Conclusions, Data Landscape and Quantified Stories, IDC Italia srl, The Lisbon Council. Retrieved from https://datalandscape.eu/sites/default/files/report/D2.6_EDM_Second_Interim_Report_28.06.2019.pdf

Prajsna, P., and Sawa, M. (2018). *Global Data Market Size 2017-2019*, OnAudience. Retrieved from https://www.onaudience.com/files/OnAudience.com_Global_Data_Market_Size_2017-2019.pdf

Reinsel, D., Gantz, J., and Rydning, J. (2017). Data Age 2025: The evolution of data to life-critical don't focus on Big Data. Focus on the data that's big. *IDC White Paper*.

Rostek, K., and Młodzianowski, P. (2017). Współzależność informacji sieciowych oraz zmian indeksów zachodzących na Giełdzie Papierów Wartościowych w Warszawie. *Zeszyty Naukowe Uniwersytetu Przyrodniczo-Humanistycznego w Siedlcach*, *42*(15), 249-263.

Saloni, Z., Woliński, M., Wołosz, R., Gruszczyński, W., and Skowrońska, D. (2012). *Słownik gramatyczny języka polskiego.* Retrieved August 8, 2019 from http://sgjp.pl

Stowarzyszenie Inwestorów Indywidualnych. (2018). *Czy polscy inwestorzy zapiszą się do PPK? Wyniki Ogólnopolskiego Badania inwestorów 2018.* Retrieved August 8, 2019 from https://www.sii.org.pl

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, *62*(3), 1139-1168.

Tetlock, P. C., Saar-Tsechansky, M., and Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *Journal of Finance*, *LXIII*(3), 1437-1467.

Wuthrich, B., Cho, V., Leung, S., and Zhang, J. (1998). Daily stock market forecast from textual web data. *IEEE International Conference on Systems, Man, and Cybernetics, Conference Proceedings*, 1-6.

Zubair, S., and Cios, K. J. (2015). Extracting news sentiment and establishing its relationship with the S&P500 Index. *48th Hawaii International Conference on System Sciences, Conference Proceedings*, 969-975.