ZESZYTY NAUKOWE UNIWERSYTETU PRZYRODNICZO-HUMANISTYCZNEGO W SIEDLCACH Nr 118
Seria: Administracja i Zarządzanie (45) 2018
pl ISSN 2082-5501

EVALUATION OF THE POSSIBILITY OF PREDICTING SOCIAL REACTIONS USING IN-DEPTH ANALYSIS OF INFORMATION PUBLISHED ON WEB-PORTALS

OCENA MOŻLIWOŚCI PRZEWIDYWANIA REAKCJI SPOŁECZNYCH ZA POMOCĄ DOGŁĘBNEJ ANALIZY INFORMACJI PUBLIKOWANYCH NA PORTALACH INTERNETOWYCH

Katarzyna Rostek, Piotr Młodzianowski ✉

Poland, Warsaw University of Technology, Faculty of Management

Abstract. Based on a literature analysis of the applicability of econometric and analytical models in the forecasting of social responses it was concluded that a dominating role is played by conditions and rules deriving from statistical physics. These models examine the straight behaviour and reactions of group members in a specific situation. There are no models where the behaviour of participants is related to phenomena occurring in their environment (closer and further) and dynamically changing states of this environment (in particular the information environment). A model for the study of the web information impact on social responses has been proposed to complement the identified research gap. The model has been evaluated on the basis of investment decisions made at the Warsaw Stock Exchange. The conducted analyzes show the usefulness of the model and the possibility of its further development in the wider context of social applications.
Keywords: forecasting, social responses, text mining, sentymental analysis

Streszczenie. Na podstawie analizy literatury dotyczącej zastosowania modeli ekonometrycznych i analitycznych w prognozowaniu reakcji społecznych stwierdzono, że dominującą rolę odgrywają warunki i reguły wynikające ze statystyki. Modele te badają proste zachowania i reakcje członków grupy w konkretnej sytuacji. Brak jest modeli, w których zachowanie uczestników związane jest ze zjawiskami zachodzącymi w ich otoczeniu (bliższym i dalszym) oraz dynamicznie zmieniającymi się stanami tego środowiska (w szczególności środowiskiem informacyjnym). Celem badania było określenie zależności pomiędzy informacjami pochodzącymi z sieciowych serwisów internetowych, a reakcją społeczną wyrażoną zmianą indeksów GPW w Warszawie. Zaproponowano model badania wpływu informacji internetowych na reakcje społeczne, który następnie wykorzystano w podejmowaniu decyzji inwestycyjnych na Giełdzie Papierów Wartościowych w Warszawie. Przeprowadzone analizy wskazują na przydatność modelu i możliwość jego dalszego rozwoju w szerszym kontekście zastosowań społecznych.
Słowa kluczowe: prognozowanie, reakcje społeczne, text mining, analiza sentymentu

Introduction

The collection, processing and effective use of information is indispensable in the development of civilization. Currently, the fastest, most accessible, and practically unlimited source of information is the Internet, more precisely, all kinds of reference and news portals, social networks, discussion blogs, and other forms of online resources that have now become a feature of the newly emergent social behaviours (Ling, 2012).

Nowadays, information is not only created by humans, but an increasing proportion of the information available is the result of the automated work of machines which can analyze data, for example geological data for the possibility of earthquakes, social data for threats to public safety, or financial data to forecast supply and demand. Taking into account the rate of growth of the availability of infor-

✉ e-mail: Katarzyna.Rostek@pw.edu.pl, Piotr.Mlodzianowski@pw.edu.pl

mation, this can be seen as an information explosion (Hilbert, 2012).

This can be defined as the rapid process of increase in the amount of information available as a result of:

- an increasing rate of production of new information;
- the ease of copying and transmission of data via the Internet;
- an increase in the number of available incoming information channels;
- the large amounts of historical data being collected;
- a lack of an effective way to process and compare different types of information, which is often conflicting, imprecise, and may lead to unnecessary duplication and copying of available data.

Benefits of the information explosion phenomenon include: better and cheaper access to information, the possibility of faster dissemination, and the creation of new roles and jobs related to information processing. Along with the benefits, there are also threats resulting from this phenomenon, which include: growing information processing costs, difficulties in distinguishing between true and false information, the impossibility of being "forgotten" online, and time losses resulting from an increasing number of emails, phone calls, and other pieces of information which employees have to read and understand (Dutta, 2013, s. 48-130).

Information can have a positive or negative impact on the user/person who receives it. In a positive sense, received information can reduce the recipient's information gap, enabling better decision making, confirming the accuracy of previously obtained information, and provide a competitive edge. Manipulation of information occurs frequently, and is an important factor in the modern information environment. The scale of this phenomenon is becoming a problem, together with associated aggressiveness, and increasingly sophisticated methods of manipulation, especially using spoken language and images (Babik, 2011). Information is able to affect the behaviour of crowds during mass events, as well as the behaviour of entire communities in response to an existing or induced phenomenon or situation. One example could be the run on Greek banks in 2015, during which, after news of the fiasco of talks between the European Union and Greece, 1 billion EUR was withdrawn. Therefore, the ability to effectively anticipate social reactions becomes a necessity, resulting also from the requirements for crisis management, and from the duty to ensure public safety.

Typical econometric and analytical models used in this field are based mainly on laws and principles derived from statistical physics. There are, however, no models in which the participants' behaviour can be directly related to the phenomena occurring around them and in the rapidly changing environment, particularly in the information environment. In order to fill this gap, a model of the impact of network information on social reactions will be proposed, presented on the basis of an analysis of investment decisions made on the Warsaw Stock Exchange (WSE). Thus, the main goal of the study will be formulated as follows:

MG: Determining the relationship between information from Internet websites and social responses manifested as changes in the WSE indices.

In order to achieve the main goal stated above, a study was prepared and carried out, with the following specific objectives:

G1: identifying Polish financial information websites that can illustrate customer reactions to changes occurring on the WSE;

G2: identifying and selecting keywords for analysis, and categorization of the keywords into classes;

G3: an examination of the impact of selected classes of words on social reactions and on the resulting changes in the value of stock exchange indices.

In relation to the defined goal, it should be noted that the majority of Internet resources exist in the form of unstructured data (e.g. documents, texts, images, recordings), which makes automatic processing difficult. Intelligent text mining systems and sentiment analysis help in searches in that giant repository, allowing for searching, classifying, summarizing and correctly interpreting information. This article will present the opportunities afforded by this type of analysis, and the results of preliminary research carried out on data published on websites connected to WSE customers.

Material and methods

Text mining is a method for exploring unstructured text documents and posts published on the Internet. A paper by H.P. Luhn published in 1958 on the automatic creation of literature abstracts and describing the role of keywords in the source text

(Luhn, 2008, s. 159-165), can be considered the first mention of text mining. The principles of text mining were developed in 1960, when the first computer systems processing unstructured text were constructed. Further development of tools for exploratory text analysis took place in the 1990's, with the appearance of new research areas, including natural language processing (NLP) and artificial intelligence (AI), on which today's text mining is based. Research on methods of exploring unstructured data seems to be extremely important, as it saves time and money that would otherwise be spent on human reading, understanding, and processing huge repositories of text documents.

Text mining is nowadays often enriched with sentiment analysis. Sentiment analysis is a method for analyzing qualitative data on the basis of the occurrence of emotionally charged words. Sentiment analysis is based on two assumptions. Firstly, some uttered words express emotions. Secondly, uttering some words may evoke certain emotions (Pang, Lee, 2008, s. 1-135). Thus, sentiment analysis indicates the emotional states of the speaker/writer, and also allows for determining the emotional effect that a given utterance may have. The term sentiment analysis was introduced in this sense by S. Das and M. Chen and R.M. Tong (Das, Chen, 2001; Tong, 2001).

Opinion analysis, exemplified by sentiment analysis, uses solutions developed in the field of natural language processing(Tong, 2001). Its practical application was accompanied by the rapid development of dictionaries for the analysis of utterances and documents (Nielsen, 2011, s. 93-98). These dictionaries allow both for simple classifications (positive / negative), and also for more complex classifications (anxiety / awe / aggression / sadness / love). Mixed dictionaries combining both ideas have also been created. An example of such a tool is the dictionary created by T. Loughran and B. McDonald (Loughran, McDonald, 2011, s. 35-65), which classifies statements referring to economics and finance on the basis of their emotional charge.

One of the first researchers to draw attention to the possibilities of using the described tools for financial market analysis was E. Lupiani-Ruiz. He developed a financial news search engine (Lupiani-Ruiz, et al., 2011, s. 15565-15572). It was limited to searching the text for numerical values. The possibility of using financial news for forecasting the direction of changes in stock exchange indices has been intensively researched since the beginning of the 21st century, with varying results (Hagenau, Liebmann, Neumann, 2013, s. 685-697; Mittermayer, 2004; Mlodzianowski, 2018; Rostek, Mlodzianowski, 2017; Schumaker, Chen, 2009, s. 1-19; Tetlock, Saar-Tsechansky, Macskassy, 2008, s. 1437-1467). The research has also covered currency markets (Peramunetilleke, Wong, 2002, s. 131-139; Nassirtoussi, Aghabozorgi, Ying Wah, Chek Ling Ngo, 2015, s. 306-324). These studies focused on searching for relationships between emerging information and news and market changes.

The most commonly used method is the so-called "bag of words", which treats the frequencies of individual words in the document as attributes, and then identifies relationships between them and the changes occurring on the market. The specific location and word order are disregarded. The multidimensionality of the attribute space created using this method is a serious problem. Typical texts contain from a few to many thousand different words. For this reason, methods for choosing words, or groups of words most semantically significant for a given set of documents are sought; or words are preliminarily categorized into classes. The classes represent words with related meanings or expressing similar emotions. This method also has disadvantages: words with the same spelling may have a different meaning, in particular if diacritics are removed during text processing, thus distorting the words containing Polish letters *ą, ć, ę, ł, ń, ó, ś, ż, ź.* The meaning of any Niven word also changes depending on the preceding words, or the context of the utterance.

The study utilized information from the most popular websites covering Business, Finance and Law. According to a survey conducted in January 2015 by Megapanel PBI / Gemiu, there are 20 Polish websites that meet this criterion, with 6 of these websites responsible for as many as 68% of the total number of users. These are: wp.pl (Money.pl), onet.pl, gazetaprawna.pl, bankier.pl, gazeta.pl, and interia.pl. Those websites were chosen as the source of the research material.

The study covered information on the home page and on subpages, one level down. The content was downloaded from the pages, but in order to ensure the objectivity of the research, the users' comments under the articles were omitted. The next stage consisted in decomposing the downloaded content into single words. Then, the frequencies of individual words were counted.

The analysis of the websites was carried out every day at 08:50, before the opening of Warsaw Stock Exchange and after its closing at 17:30. Each test lasted about 5 minutes. The results thus collected were the basis for predicting the social reaction, defined as the direction of changes in stock exchange indices. It should be noted that all the information available at 08:50 was included, regardless of the time of its publication. The 17:30 data was used to verify the accuracy of the selected key words. The complete set of observations covered the period between 01-06-2016 and 31-12-2016, totaling 149 trading sessions.

To convert a stream of characters into individual words, the author's own software implemented in MS Excel was used. Then, the aggregate database of words was searched for key words. All analyses were carried out with SAS Institute tools: SAS Enterprise Miner and SAS Enterprise Guide. Thanks to the adopted form of word identification, there was no need to bring a word to the lexeme (basic form) level, a significant issue with a heavily inflected language such as Polish (where, for example, words czytać, czytam, czytali, przeczytasz - meaning to read, I am reading, they were reading, you will read, are all versions of the same lexeme). The searched for words were subjected to morphological synthesis, i.e. generating a suitable form of a part of speech based on the basic form and attributes describing the form. This allowed searching for the same words, differing only in their grammatical form (e.g. case forms as in kryzys (nominative), kryzysu (genitive), kryzysie (locative etc.)), then counting them, and categorizing them into two classes: positive and negative. Due to the test consisting in searching for predetermined content, the issues arising from the appropriate interpretation of punctuation and from disambiguation of homographs (words with the same spelling but different meanings e.g. a match can mean a sports competition, a state of equivalence, a marriage, or a small wooden stick for making fire) were disregarded. The disadvantage of this approach was the inability to take into account the context-dependent meaning of words.

The next stage of the test is the construction of a frequency matrix, which transforms the set of found and classified words into a quantitative format. The matrix rows are words appearing on a given day on the tested information websites. The columns represent positive (Kp) and negative (Kn) word classes. A cell of the frequency matrix can be defined as follows:

Occurence_matrixij = f (the number of occurrences of word i on day j)

The last column of the frequency matrix contains the evaluation of the Information Environment Disposition (IED) before the trading opens (IED_j), which is the difference between the frequencies of words in the positive and negative classes. It is calculated as follows:

$$IED_j = Kp_j - Kn_j$$

if:

$IED_j > 0,\ forecast\ index\ change\ direction\ on\ day\ j\ is\ increase$
$IED_j < 0,\ forecast\ index\ change\ direction\ on\ day\ j\ is\ decrease$
$IED_j = 0, no\ forecast\ on\ day\ j$

where:
IED_j – Information Environment Disposition before the start of trading on day j,
Kp_j – number of positive class words on day j,
Kn_j – number of negative class words on day j.
IED is compared with the value of stock exchange index changes on the same day.

The combination of text data with the time series consists in assigning the quantitative value of the index change, calculated on the day of the analysis, to the forecasted direction of the change. It is worth considering after how long a period the value of indices will reflect the state of knowledge from the analysis period, i.e. after what period that knowledge becomes a component of the price. Taking into account the fact that the analyzed information is available to all users without any barriers, the time of its "absorption" by the market should be close to zero. The study adopted 10 periods for index analysis, including the two main ones: 09:00 – the value of the analyzed indices at opening (the period immediately after the online information analysis) and 17:05 – the value of indices at the close of trading.

The effect of the information's influence was measured by the value of the index change expressed in points. If the forecast direction was consistent with the actual index change direction, then the change in value was treated as a profit, otherwise it was treated as a loss. The index change values were calculated at 09:00 (IndexChangeOpen) and at 17:05 (IndexChangeClose) in the following way:

$$IndexChangeOpen_{w,j} = OpeningValue_{w,j} - ClosingValue_{w,j-1}$$
$$IndexChangeClose_{w,j} = ClosingValue_{w,j} - OpeningValue_{w,j}$$

where:
w – index name,
j – test day.

For all indices, additional analysis of the impact of information on the index change was carried out at 10:00, 11:00, 12:00, 13:00, 14:00, 15:00, 16:00, 17:00, calculated as follows:

$$IndexChange_{w,j,h} = IndexValue_{w,j,h} - OpeningValue_{w,j}$$

where:
w – index name,
j – test day,
h – test hour.

The news published after the end of the trading session could only be considered by investors at the opening the following day.

As was the case for developing the frequency matrix, it was also necessary to transform the set of words, found and classified at 17:30 on a given day, into quantitative form. The frequency matrix cell was defined as follows:

Frequency_matrix'$_{i'j'}$ = f' (the number of occurrences of word 'i on day 'j)

The last column of the frequency matrix contains the evaluation of the Information Environment Disposition (IED) before trading opens (IED$_j$′), which is the difference in the frequencies of words in the positive and negative classes. This is calculated as follows:

$$IED_j{}' = Kp_j{}' - Kn_j{}'$$

if:

$IED_j{}' > 0,\ forecast\ index\ change\ direction\ on\ day\ j\ is\ increase$

$IED_j{}' < 0,\ forecast\ index\ change\ direction\ on\ day\ j\ is\ decrease$

$IED_j{}' = 0, no\ forecast\ on\ day\ j$

where:
IED$_j$′ – Information Environment Disposition after the close of trading on day j,
Kp$_j$′ – number of positive class words on day j after the close of trading,
Kn$_j$′ – number of negative class words on day j after the close of trading.

The IED′ value was compared with the value of stock exchange index changes (IndexChangeClose$_{w,j}$′) on the same day. This is calculated as follows:

$$IndexChangeClose_{w,j}{}' = OpeningValue_{w,j} - ClosingValue_{w,j-1}$$

where: w – index name, j – test day.

If the IED at the close of trading was consistent with the index change direction, the change value on that day was treated as a value accurately reflected using selected words and created classes. If the IED$_j$′ value adopted a different direction of change than the stock market index, the value of index change was qualified as a value that could not be accurately reflected. In this manner, it was possible to determine whether the words selected described the changes in stock indices with a sufficiently high probability (greater than the coin toss = 50%) and whether they could be used to forecast changes in indices. Other analyzed variables included the optimal transaction time (using the example of the WIG index), the impact of the number of individual words on the size of the index change, and the possibility of the occurrence of critical phenomena such as a stock market crash (large drop in the index) or euphoria (large increase in the index).

Based on the completed analysis and comparisons, conclusions have been developed regarding the possibility of using online information from websites to predict social reactions, and the resulting changes in stock exchange indices. In order to achieve the main goal and the specific goals, the analysis focused on answering the following research questions:

Q1: Which financial websites are the most popular among Polish stock market investors? The answer to Q1 would allow the specific goal G1 to be achieved.

Q2: Do the selected positive and negative words predict social reactions, and consequently, the changes in stock market indices? The answer to Q2 would allow the specific goal G2 to be achieved.

Q3: To what extent do the selected classes of words correspond to social reactions and related changes of stock exchange indices? The answer to Q3 would allow the specific goal G3 to be achieved.

Results and discussion

The analysis of websites covering Business, Finance, and Law allowed for the identification of the most popular websites among Polish stock exchange investors. It showed that 68% of the total number of users focused on six of the websites included in the study (Figure 1). In this way, the answer to the formulated research questions Q1 was obtained, which translates into the achievement of the specific objective G1.

The study utilized the values of the following Warsaw Stock Exchange indices: WIG, WIG20, mWIG40 and sWIG80. This choice was guided by the format of the study, in which selected words were searched for, whilst omitting the context and correlation with the names of individual companies. Among the analyzed content, the following words were searched for: bessa, hossa, spada, rośnie, niedźwiedź, byk, zielony, czerwony, zysk, strata (traci), ożywienie, kryzys [bear market, bull market, rises, falls, bear, bull, green, red, profit, loss (loses), recovery, crisis]. They had been selected ex ante at the author's discretion. The choice was motivated by the words being able to reflect the mood and emotions prevailing on the equity market. They were compared with tests carried out at 17:30 to determine the accuracy of the description of the index change direction on a given day. Next, the words were grouped into two classes using sentiment analysis (Table 1), dividing words into those causing negative (decreases) and positive (increases) emotions.

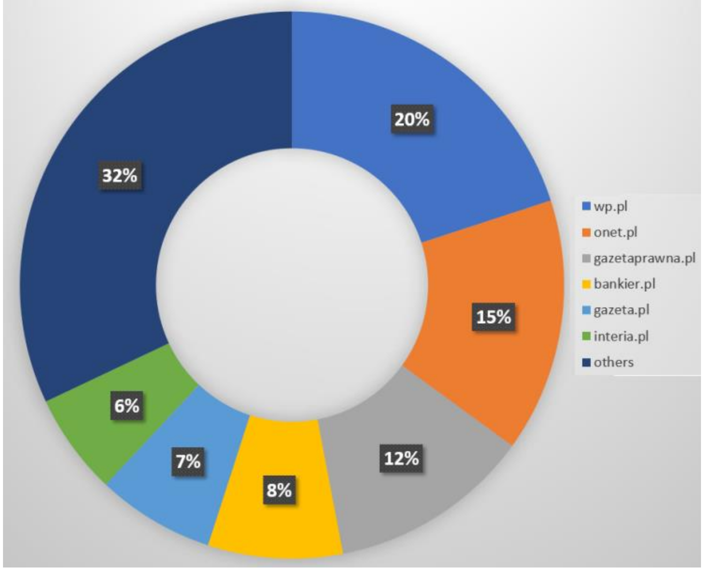

Figure 1. Business, Finance, and Law websites in the sample by the number of users
Source: own elaboration.

Table 1. Categorization into positive and negative classes

Negative class (Kn)	Positive class (Kp)
bessa, spada, niedźwiedź, czerwony, strata, kryzys [bear market, falls, bear, red, crisis, loss (loses)]	hossa, rośnie, byk, zielony, zysk, ożywienie [bull market, rises, bull, green, profit, recovery]

Source: own elaboration.

Answering the Q2 research question and achieving the specific research goal G2 consisted in determining whether the occurrence of selected keywords at 17:30 corresponded to index changes on the analyzed day. Test results for a selected 5-day period and the change in the WIG index in relation to IED$_j$ are shown in Table 2 (scores are in points). The results of the complete study covering 149 days and WIG, WIG20, mWIG40, and sWIG80 are presented in Table 3.

Table 2. W Test results: IED$_j$′ and the changes in WIG between 22-09-2016 and 28-09-2016

Test day	2015-09-22	2015-09-23	2015-09-26	2015-09-27	2015-09-28
IndexChangeClose’$_{wig,j}$	529,65	-289,1	-417,67	-169,45	39,06
IED$_j$′	increase	decrease	decrease	decrease	increase
The value that was could or couldn't be reflected	529,65	289,1	417,67	169,45	39,06
The value that was could or couldn't be reflected ascending	529,65	818,75	1236,42	1405,87	1444,93

Source: own elaboration.

Table 3. Test results: IED_j' for 17:05 covering 149 days and WIG, WIG20, mWIG40 and sWIG80 indices

Index name	WIG	WIG20	mWIG40	sWIG80
$IndexChangeClose'_{wig,j}$	37950,45	1769,38	2881,11	5330,65
Total volatility of $IndexChangeClose'_{wig,j}$	48415,16	2313,11	3881.06	7542,08
Success rate	(78%)	(76%)	(74%)	(71%)

Source: own elaboration.

On the basis of this analysis, it can be concluded that the frequency of key words and the proposed categorization allows for describing changes in the direction of the analyzed stock market indices with a better than random (coin toss) probability. For each of the analyzed indices, the analysis indicated a better than 50% effectiveness in reflecting the index change. This may indicate that the selected positive and negative words and their proposed division into classes satisfactorily reflects investor behaviour and changes of the stock exchange indices caused by this behaviour.

In order to answer the Q3 research question and achieve the specific objective G3, online information that became available before the opening of WSE was converted into quantitative format. Then, the results of the class tests were compared with the directions of stock index changes. The results for WIG, WIG20, mWIG40, and sWIG80 across the whole period of the study, for 09:00 – opening of trading, and for 17:05 – close of trading, are presented in Tables 4 and 5.

Table 4. Test results: IED_j for 09:00 covering 149 days and WIG, WIG20, mWIG40 and sWIG80 indices

Index name	WIG	WIG20	mWIG40	sWIG80
$IndexChangeOpen_{w,j}$ (correctly predicted)	15616,22	454,41	659,36	1796,98
Total volatility of $IndexChangeOpen_{w,j}$	27105,94	750,91	1518,43	3671,15
Success rate	(58%)	(61%)	(43%)	(49%)

Source: own elaboration.

Table 5. Test results: IED_j for 17:05 covering 149 days and WIG, WIG20, mWIG40 and sWIG80 indices

Index name	WIG	WIG20	mWIG40	sWIG80
$IndexChangeClose_{wig,j}$ (correctly predicted)	24776,97	1219,93	2074,7	4004,42
Total volatility of $IndexChangeClose_{wig,j}$	41228,85	2097,54	3214,13	6733,79
Success rate	(60%)	(58%)	(65%)	(59%)

Source: own elaboration.

Based on the analysis, it can be seen that the forecast success coefficient for 09:00 is above 50% for WIG and WIG20, while for mWIG40 and sWIG80 the values are respectively 43% and 49%. The results from examining the IED and the changes in the value of stock exchange indices for 09:00 suggest that the Information Environment Disposition before opening significantly affects WIG and WIG20 at the start of trading. In the 17:05 test, the forecast success coefficient for WIG, mWIG40 and sWIG80 significantly exceeded 50%, and for each of these, the coefficient at 17:05 was higher than at 09:00, as illustrated in Figure 2.

This shows that the Information Environment Disposition before opening has a stronger impact on the change of WIG, mWIG40 and sWIG80 indexes at

17:05 than at 09:00. Therefore, it can be concluded that investors making a purchase / sale decision take into account the available information to a greater extent at 17:05 than at 09:00, even though the information is available in their information environment on online information portals before the opening of trading. The success coefficient for the 17:05 forecast is higher than 50% for all indices, which shows that predicting investors' reactions on the basis of occurrences of defined word classes is more accurate than a random prediction (based on a coin toss) would be.

The next stage in answering the research question Q3 was determining how quickly online information was "absorbed" from the investor's information environment. The study demonstrated that online information is taken into account by investors at the opening of trading at 09:00 to a lesser extent than at the close of trading at 17:05. Table 6 summarizes the forecast success for the analyzed indices during the trading period.

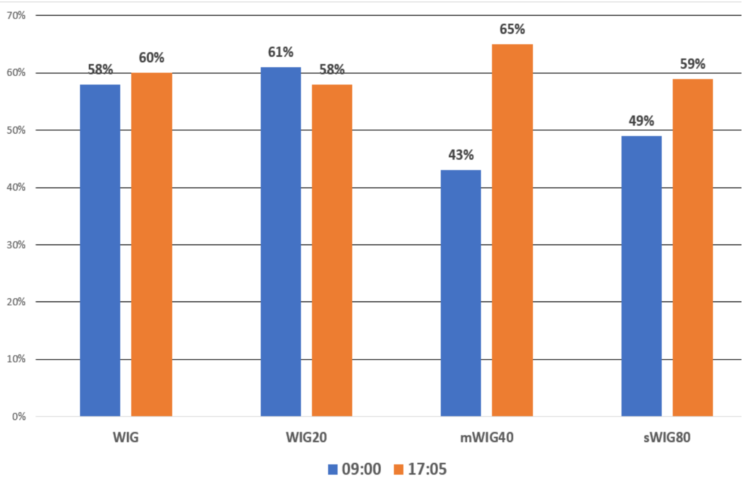

Figure 2. Forecast success rates for 09:00 and 17:05
Source: own elaboration.

Table 6. Forecast success rates for WIG, WIG20, mWIG40 and sWIG80 indices

Test hour	Index Name (%)			
	WIG	WIG20	mWIG40	sWIG80
10:00	58	54	49	49
11:00	55	52	48	44
12:00	56	53	50	45
13:00	56	53	52	45
14:00	58	56	54	46
15:00	58	55	55	47
16:00	59	56	57	49
17:00	57	56	60	53
17:05	60	58	65	59

Source: own elaboration

The analysis indicates that the forecast success coefficient for all indices achieves the maximum value at the close of trading, i.e. at 17:05 (green). It is surprising that the minimum success rate for all indices (with the index value at 09:00 being the reference point) also occurs at the same time, i.e. 11:00 (orange). For mWIG40 and sWIG80, the forecast success coefficient for 11:00 was below 50%. This means that the optimal time for predicting the change in these indices is not 09:00, shortly after the IED, test, but 11:00. Therefore, one should consider the case in which, for mWIG40 and sWIG80, $\text{OpeningValue}_{w,j}$ in the formula:

$$IndexChangeClose_{w,j} = ClosingValue_{w,j} - OpeningValue_{w,j}$$

is replaced with IndexValue$_{w,j,11:00}$

where:
w – index name,
j – test day,
h – test hour.

In such a case, the success coefficients for mWIG40 and sWIG80 forecasts are respectively 68% and 72%.

The next issue that was analyzed in order to obtain answers to the Q3 research questions was the relationship between the frequency of words within individual classes and the forecast success coefficient. Figure 3 presents the result of the IED tests and the corresponding changes in WIG.

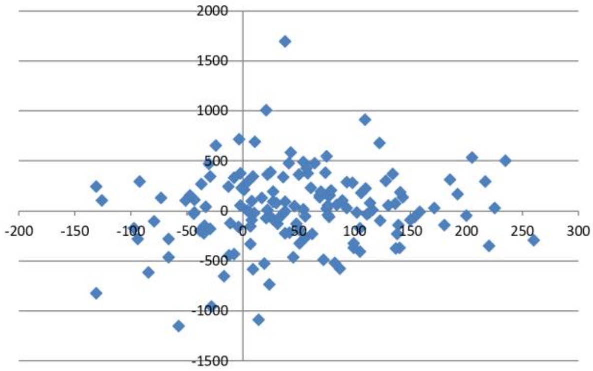

Figure 3. The value of WIG index change depending on IED
Source: own elaboration.

Quadrant I and III of the coordinate system contain elements corresponding to index changes that were correctly predicted. In the first quadrant, the index increase was consistent with the growth forecast, while in the third quarter the fall forecast was consistent with the index decrease. The second and fourth quadrants of the coordinate system contain inaccurate forecasts.

Table 7 presents the asymmetry in the distribution of the forecast success (quadrants I and III) and its failure (quadrants II and IV). In the analyzed indices, the success of the forecast is primarily due to the correctly predicted increases (quadrant I). And in turn, incorrectly predicted increase is the main factor in the unsuccessful forecast (quadrant IV). This was due to the fact that IED tests resulted in index growth being the most common forecast.

The pro-growth Information Market Disposition was a result of the bullish market at the time of the study. Clearly, as the market situation changes, so will the disposition.

Table 7. Forecast success and failure rates for WIG, WIG20, mWIG40 and sWIG80 depicted in quadrants of the coordinate system

Quadrant	Index name (%)			
	WIG	WIG20	mWIG40	sWIG80
I	42	42	47	37
II	12	13	9	12
III	18	16	18	23
IV	28	29	26	28

Source: own elaboration.

Conclusions

The presented research results allowed for the achievement of the set goals and answers to the formulated research questions. The main conclusion resulting from the study is the existence of a significant relationship between online information and the changes of Warsaw Stock Exchange indices. The achieved forecast accuracy, at the level of not less than 58% (for WIG20) allows obtaining financial benefits on the equity market. For this reason, further research in a market environment is required. If the accuracy of the forecast remains at a similar level, the tool could be used as the basis for the construction of an algorithmic trading system or to support decision making by stockbrokers. The research results presented in this paper suggest that the proposed solution may be used for assessing the sentiments prevailing in the investors' online information environment, which is an alternative to the Index of

Investors' Moods maintained by the Individual Investors Association. It should also be stated that this analytical tool needs further development and one of the fundamental issues is the identification of key words and their categorization into appropriate and well targeted classes. This is an extremely difficult task which requires thorough research among equity market participants in terms of word selection and categorization into representative classes, as well as determining their weighted importance for the forecast. Dissemination of tools enabling stock exchange investors to analyze large datasets in near real time will have a positive impact on reducing barriers in access to information. As a consequence, there will be an increase in the speed at which emerging information that is extracted from the analyzed data streams is reflected in the stock prices. Selection, reading, understanding and interpreting the information by a single investor using traditional visualization methods is more and more time-consuming. It is hypothesized that the implementation of analytical support solutions would improve the information efficiency of the Warsaw Stock Exchange.

References

Babik, W. (2011). O manipulowaniu informacją w prywatnej i publicznej przestrzeni informacyjnej, Kraków: Uniwersytet Jagielloński.

Das, S., Chen, M. (2001). Yahoo! for Amazon: Extracting market sentiment from stock message boards. Proceedings of the Asia Pacific finance association annual conference, 2001, vol. 35.

Dutta, S. (2013). Business Communications, Dheli: PHI Lerning Private Limited.

Hagenau, M., Liebmann, M., Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features, Decision Support Systems, no. 55, 2013, 685-697.

Hilbert, M. (2012). How much information is there in the "information society"?, Significance, vol. 9, no. 4, 2012, 8-12.

Ling, R. (2012). Taken for Grantedness: The Embedding of Mobile Communication into Society, The MIT Press.

Loughran, T., McDonald, B. (2011). When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks., The Journal of Finance, vol. 66, no. 1, 2011, 35-65.

Luhn, H.P. (1958). The automatic creation of literature abstracts, IBM Journal of Reasearch and Development, 159-165.

Lupiani-Ruiz, E., García-Manotas, I., Valencia-García, R., García-Sánchez, F., Castellanos-Nieves, D. Fernández-Breis J.T., et al. (2011). Financial news semantic serach engine. Expert Systems with Applications, no. 38, 2011, 15565-15572.

Mittermayer, M.A. (2004). Forecasting intraday stock price trends with text mining techniques. Proceedings of the 37th annual Hawaii international conference on system sciences.

Młodzianowski, P. (2018). A Study of the Influence of Online Information on the Changes in the Warsaw Stock Exchange Indexes. Acta Universitatis Lodziensis Folia Oeconomica, vol. 3, nr 335, 2018, 123-138.

Nassirtoussi, A.K., Aghabozorgi, S., Ying Wah, T., Chek Ling Ngo D. (2015). Text mining of news-headlines for FOREX market prediction. A Multi-layer Dimension Reduction Algorithm with semantics and sentiment, Expert Systems with Applications, no. 42, 2015, 306-324.

Nasukawa, T., Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. Proceedings of the Conference on Knowledge Capture, 70-77.

Nielsen, F.Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblog, Proceedings of the ESWC2011 Workshop on „Making Sense of Microposts": Big things come in small packages 718 in CEUR Workshop Proceedings, Heraklion, 93-98.

Pang, B., Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, vol. 2, no 1-2, 2008, 1–135.

Peramunetilleke, D., Wong, R.K. (2002). Currency exchange rate forecasting from news headlines, Australian Computer Science Communications, no. 24, 2002, 131-139.

Rostek, K., Młodzianowski, P. (2017). Współzależność informacji sieciowych oraz zmian indeksów zachodzących na Giełdzie Papierów Wartościowych w Warszawie. Zeszyty Naukowe Uniwersytetu Przyrodniczo-Humanistycznego w Siedlcach nr 115. Seria: Administracja i Zarządzanie (42) 2017, 249-263.

Schumaker, R.P., Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZF in text system, ACM Transactions on Information Systems, no. 27, 2009, 1-19.

Tetlock, P.C., Saar-Tsechansky, M., Macskassy, S. (2008). More than words: Quantifying language to measure firms fundamentals. The Journal of Finance, no. 63, 2008, 1437–1467.

Tong, R.M. (2001). An operational system for detecting and tracking opinions in on-line discussion. Working Notes of the SIGIR Workshop on Operational Text Classification. New York: ACM, 2001, 1-6.