Włodzimierz Gogołek

Dariusz Jaruga

From research on the system of refining the Web
Identifying sentiment words

KEY WORDS

information, internet, Big Data, collecting information, identifying sentiment words, sentiment analysis

ABSTRACT

The available potential of computing power and computer memory have created previously unavailable conditions for the analysis of large information resources – Big Data. In the process of this analysis, the procedures for collecting and analysing information can be used for the purpose of accurate assessment – in terms of emotions (sentiments – good, bad) of the studied phenomena – in the past, in real time, as well as for prediction. The article is a presentation of the key parts of this procedure – the essence of automation of the process of identifying sentiment words.

Refining is the process of purification and refinement of natural substances or industrial products in order to give them an appropriate purity, colour and smell. The definition of the refining process quoted above was taken from *Słownik języka polskiego*[1] (Dictionary of the Polish language). It also reflects the manner in which the substance such as large information resources - the Big Data - is refined. The expected effect of the process is new information hidden in those resources.

[1] M. Szymczak, *Słownik języka polskiego* [Dictionary of the Polish language], Warszawa, 1978.

In the system of refining of information (RI), the substance subjected to processing are the source materials (materials) in a textual or audio form acquired from the Web, or from large collections of data available offline - the Big Data[2]. The final outcome of RI is the result of the statistical analysis of key phrases and the surrounding sentiments, i.e. expressions that convey emotions and are written in the digital information resources. Thanks to RI, it is possible to extract the new, valuable information hidden in the content, e.g. the evaluation of a social phenomenon (support, satisfaction, negation)[3].

When embarking on the examination of a specific social phenomenon (which may be related to business, politics, medicine or other fields), the first step is to determine the words or phrases which are connected with the identification/name of the phenomenon under study. In this method, such key words, or key phrases related to the phenomenon are referred to as "pillars" (e.g. the name of a party, company, surname). The second step is to identify sentiments (a kind of evaluations), the third one - to calculate the frequency with which sentiments are present around pillars. The last step, which was omitted in the article, involves the interpretation of the results (e.g. evaluation of brand popularity, evaluation of the brand).

The above-mentioned stages of refining will be discussed in further parts of this article. They document an important link in the process of refining the information from the Web, i.e. the identification of statistically significant sentiment words[4].

The name of the phenomenon "pillar"

As already mentioned, depending on the object of the study, the pillar can be the name of a brand, product, political party, organisation, city, surname of a person, e.g. of a politician. The pillar can be understood more broadly and does not have to be restricted to a single word or phrase. On the contrary, it may be a whole set of words and phrases that are synonymous or antonymous, or a set of words and phrases concerning a given subject-matter (e.g. the budget, economic result, report, audit, etc.). The pillar can encompass all the possible forms of a word

[2] V. Marx, *The big challenges of Big Data* [in:] "Nature" 2013, vol. 498; W. Gogołek, P. Kuczma, *Rafinacja informacji sieciowych na przykładzie wyborów parlamentarnych* [Refining network information on the example of parliamentary elections], "Studia Medioznawcze" [Media studies] 2013, no. 2 (53).

[3] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, *From Data mining to knowledge discovery in Database*, www.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf [accessed: 11/11/2016].

[4] Ch. Curme et al., *Quantifying the semantics of search behaviour before stock market moves*, "PNAS" 2014, no. 32; J. Smailovič, *Predictive sentiment analysis of Tweets: A stock market application* [in:] *Human-computer interaction and knowledge discovery in complex, unstructured, Big Data* 2013, pp. 77–88.

or expression inflected according to person, tense (future, past or present), moods (conditional, imperative), participles, neologisms, including words containing spelling mistakes, typos and increasingly more common hashtags[5].

For example, the pillar for the Polish word *leczenie* 'treatment' includes 38 words: *leczą*, *lecząc*, *leczący*, *leczcie*, *leczę*, *leczmy*, *leczony*, *leczy*, *leczycie*, *leczyć*, *leczyli*, *leczyliby*, *leczylibyście*, *leczylibyśmy*, *leczyliście*, *leczyliśmy*, *leczył*, *leczyła*, *leczyłaby*, *leczyłabym*, *leczyłabyś*, *leczyłam*, *leczyłaś*, *leczyłby*, *leczyłbym*, *leczyłbyś*, *leczyłem*, *leczyłeś*, *leczyło*, *leczyłoby*, *leczyły*, *leczyłyby*, *leczyłybyście*, *leczyłybyśmy*, *leczyłyście*, *leczyłyśmy*, *leczymy*, *leczysz*. The set includes the forms of the verb inflected according to person in the present, past and compound future tense, the inflected forms of the conditional and imperative mood, as well as the participles.

The choice of the pillar is the first step to be undertaken before proceeding to the identification of the words or phrases that are sentiments.

RI methodology assumes that sentiments can be determined by means of three procedures: (1) thanks to the researcher's intuition based on examining a random sample of texts from the source dataset that will be subjected to investigation; (2) available dictionaries of words (dictionaries)[6] that can be regarded as sentiments (which have so far been verified experimentally); (3) based on the frequency analysis of the words from a selected source dataset (AC). The developed tool for calculating frequency enables the identification of atypical words or neologisms which are present in the examined collection of texts with a high frequency. In this way, specific words and phrases are subjected to inflection and ultimately become a sentiment. One advantage of this tool is the ability to analyse the entire source material, rather than only examining its random sample.

At the Faculty of Journalism, Information and Book Studies of the University of Warsaw, an innovative tool was developed which calculates the frequency of all words found in the examined material. In the subsequent stage, the results obtained in this way are analysed according to two procedures in order to identify sentiments: From among the words

[5] A hashtag is a word or phrase written without spaces, preceded by the symbol #, e.g. #dziejesię. Hashtags are used in microblogs, social networks, on websites, etc. to enable the grouping of messages. At the time when the article was written, the entry "hashtag" was not available in e.g. *Encyclopedia Britannica*, or Encyklopedia *PWN* [Encyclopaedia of the Polish Scientific Publishing Company]. Therefore, for the purposes of this study, the definition given in Wikipedia under https://pl.wikipedia.org/wiki/Hasztag was adopted [accessed: 05/10/2016].
[6] W. Gogołek, P. Kuczma, *Rafinacja informacji sieciowych…*, op. cit.

indicated on the basis of AC and/or dictionaries, the researcher selects the suitable words or phrases (sentiments) for the second stage (ACB), or the identification takes place automatically (ACA) based on the verification of statistical significance of the words regarded as sentiments, according to the procedure described below.

Sentiments

The next step involves the search for phrases - sentiments - that carry a particular emotional load (e.g. positive, negative or neutral) and, through the frequency analysis of particular sentiments in the neighbourhood of the pillar within a defined time period, enable the calculation of e.g. the society's attitude to the examined issue.

Sentiments, as it was said, can be determined in several ways, e.g. through manual analysis of randomly selected source materials, by using previously created dictionaries of words and phrases, or by means of the above-mentioned ACA tool developed at the Faculty of Journalism, Information and Book Studies of the University of Warsaw. The tool enables the analysis of the entire source material which will be used for further research, i.e. the identification of sentiments.

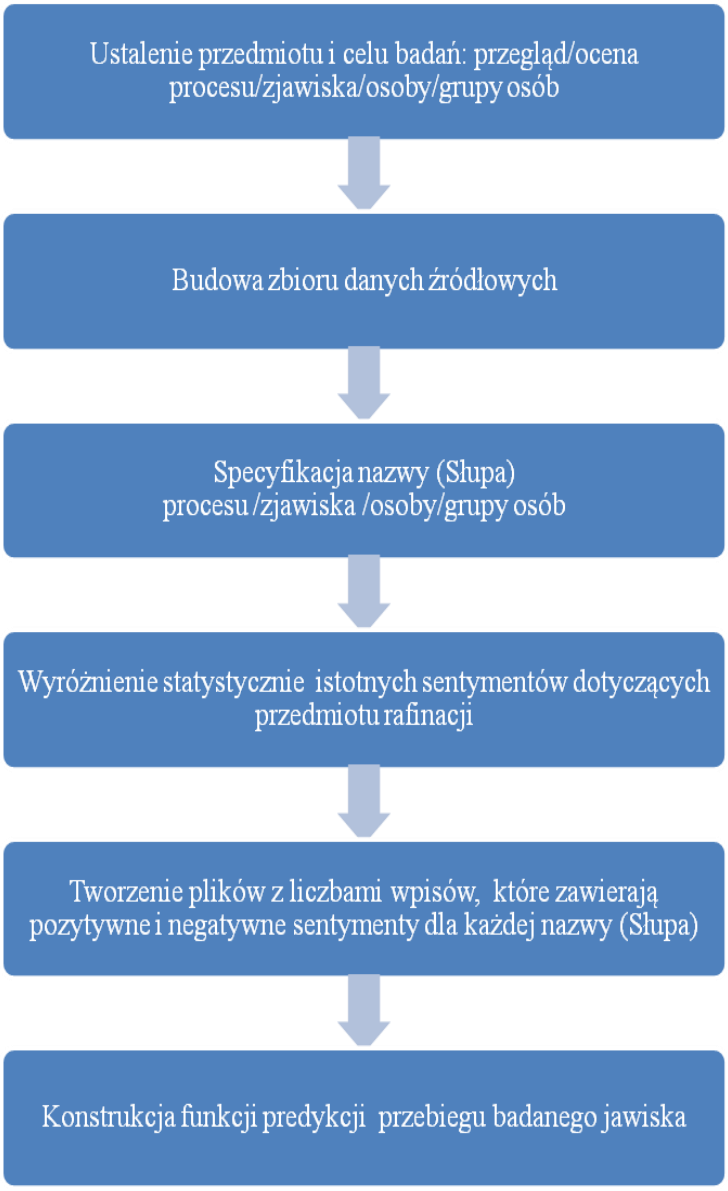


Fig. 1. Sequence of refining network resources
Source: own work

Source material

Prior to launching the procedure of sentiment identification, the source material collected by an innovative BigData robot is subjected to preparatory processing. The processing consists of three stages: filtering the content to be examined from the robot's database; purification of the data and transforming it into a normalised form required by the software; performance of relevant calculations. The materials available online can be divided into four basic groups based on the manner of presentation: textual materials, images, sound and video.

At the present stage of research, the method for identifying sentiments in the neighbourhood of pillars comes down to the analysis of purely textual materials. This does not preclude the use of audio materials examined by means of available tools for analysis of human speech sounds. Graphic materials containing texts are analysed using OCR (Optical Character Recognition) technology.

In line with the subsequently described stages of operation of the programme for calculating sentiment words, the first step consists of filtering the records to be examined from the BigData robot's database. Filtering mainly involves the selection of a specific subset of content that has been collected by BigData. The records to be examined can be selected based on several criteria, such as the periods of time (data/hour from -to), sources of information, words or phrases present in the content, or in the title. The data selected in this way is passed on to the next stage where it is purified and transformed into a normalised form required by the software. In practice, the purification of a website means deleting all markers from its source. As a result of this operation, only useful content of the document is left of the source material.

The textual material obtained from a source document has to be given a normalised form before being subjected to further analysis. Normalisation means that the redundant space, tab and end-of-line characters are removed from the text. As a result of normalisation (Fig. 2), a single line of text is obtained which includes the entire content of the document and where particular words are separated by no more than a single space. Punctuation marks, brackets, etc. are also left in the normalised text.

The normalised text form is required by the software in order to calculate the frequency of words in the neighbourhood of the previously described *Pillar*.

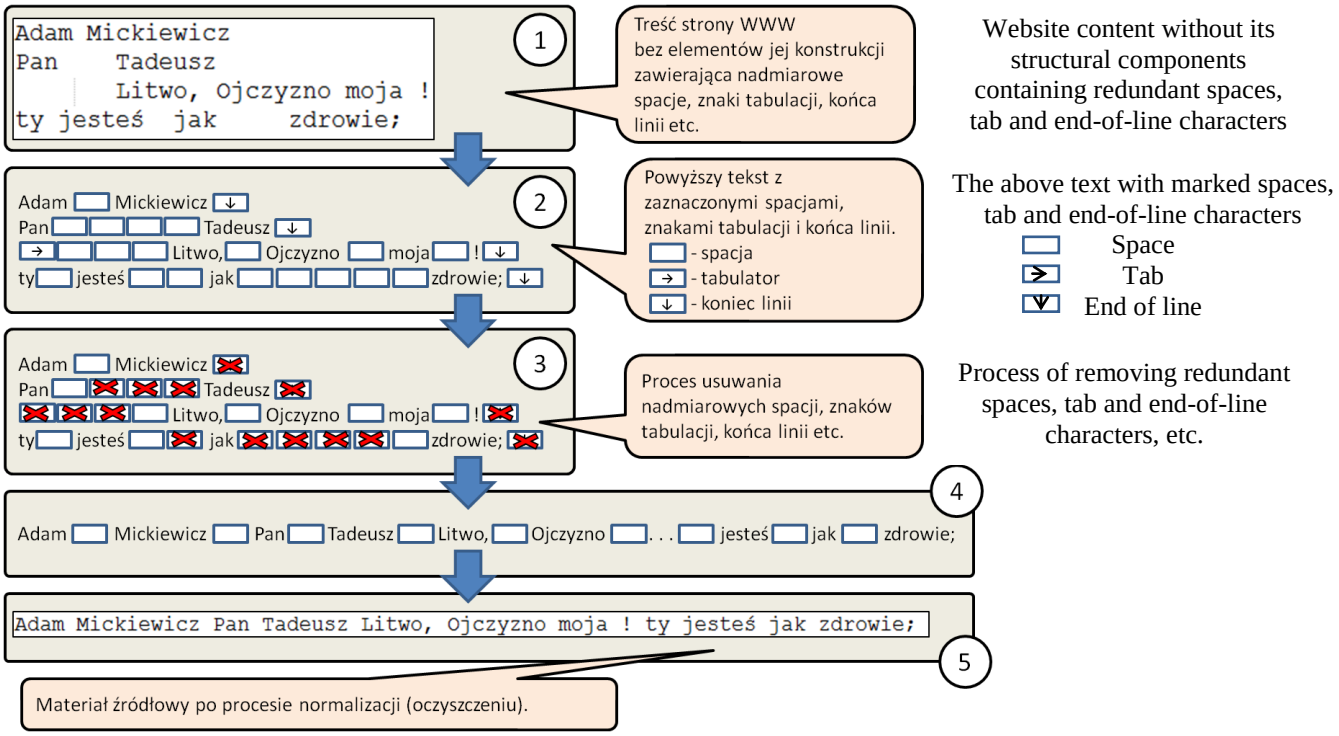


Fig. 2. Process of source text normalisation
Source: own work

Frequencies

The next step after text normalisation are the procedures for calculating the frequency of words in the neighbourhood of the pillar, i.e. words separated from the pillar by a specified "n" number of characters[7] which does not exceed the value defined as the input parameter. The greater the value of the distance, the more words will be qualified into the group of words in the neighbourhood of the pillar. This was presented diagrammatically in Figure 3.

[7] For the purposes of this study, the neighbourhood of a pillar was expressed through the number of characters and marked with the letter "n", or, if necessary, with the letter "n" with subscript, e.g.. n_1, n_2 etc.

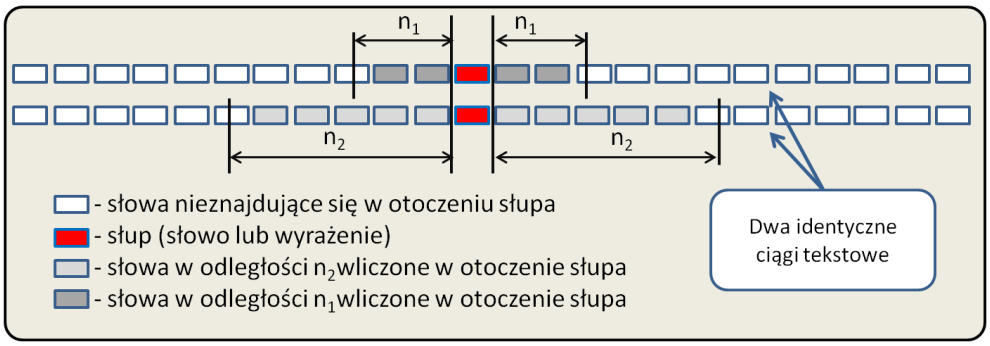


Fig. 3. The number of words qualified as the neighbourhood of the pillar, depending on the value of parameter "n".
Source: own work

In the first case, for "n_1" there were four words in the neighbourhood of the pillar, in the second case for "n_2" there were 10 words in the pillar's neighbourhood. If the limit of "n" characters occurs in the middle of a word rather than on its boundary, the word is not counted as the neighbourhood of the pillar. It does not matter between which letters of a given word the limit of the area defined by parameter "n" passes. The boundary of a word is understood as the contact point between the space and the first letter of the word, or the last letter and the space that follows.

The neighbourhood of the pillar includes those words that are completely contained in the "n" range, or are on the boundary of that area - the boundary of a word. The greater the value of parameter "n", the more words will be qualified as the neighbourhood of the pillar. In practice, the value of parameter "n" falls within the range of 10 to 20 characters. Therefore, parameter "n" significantly affects the set of words in the neighbourhood of the pillar, but is not the only parameter that decides which words will be regarded as the neighbourhood of the pillar. Another factor that decides which words will be counted as the neighbourhood of the pillar is the mutual position of two pillars in the examined text. The most important aspect to be considered is the mutual distance between two pillars which, in this study, will be described by parameter "m". In other words, "m" is the distance, measured in characters, between one pillar and the other, and more precisely, between the left-hand and right-hand

boundary of a pillar which is determined in a way similar to that of a word, as depicted in Figure 4.

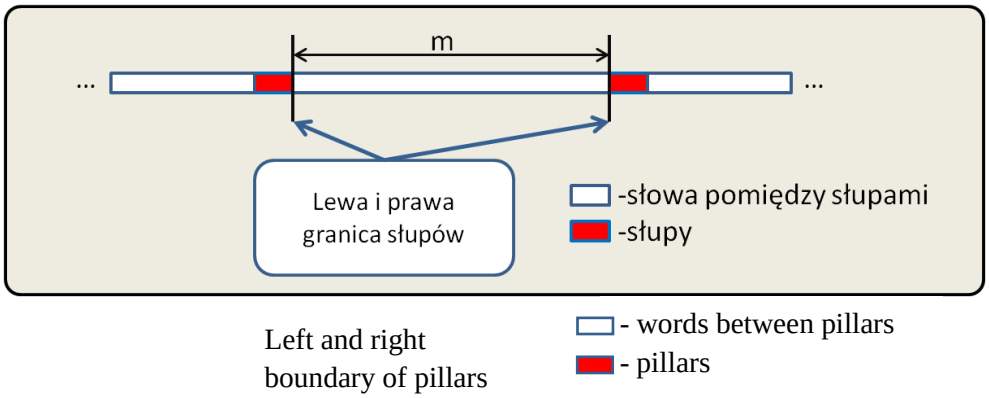


Fig. 4. Illustration of left and right boundary of two pillars and the distance between them defined by parameter "m".
Source: own work

In the examined source text that has already been normalised, the number of pillars may range from zero to a certain finite natural number. Even if a pillar occurs only once in a given text, it is necessary to consider its position in relation to the beginning and the end of the text as a whole. As a reminder, a normalised text is a single line with unambiguously specified single beginning and ending. For two or more pillars, their mutual position needs to be additionally analysed, because it determines the manner of counting the words in their neighbourhood. Researchers have identified six combinations of mutual positions of various pillars. They exhaust all the possible situations that may occur, and consequently provide the basis for developing a complete algorithm for calculating the frequency of words in the neighbourhood of a pillar.

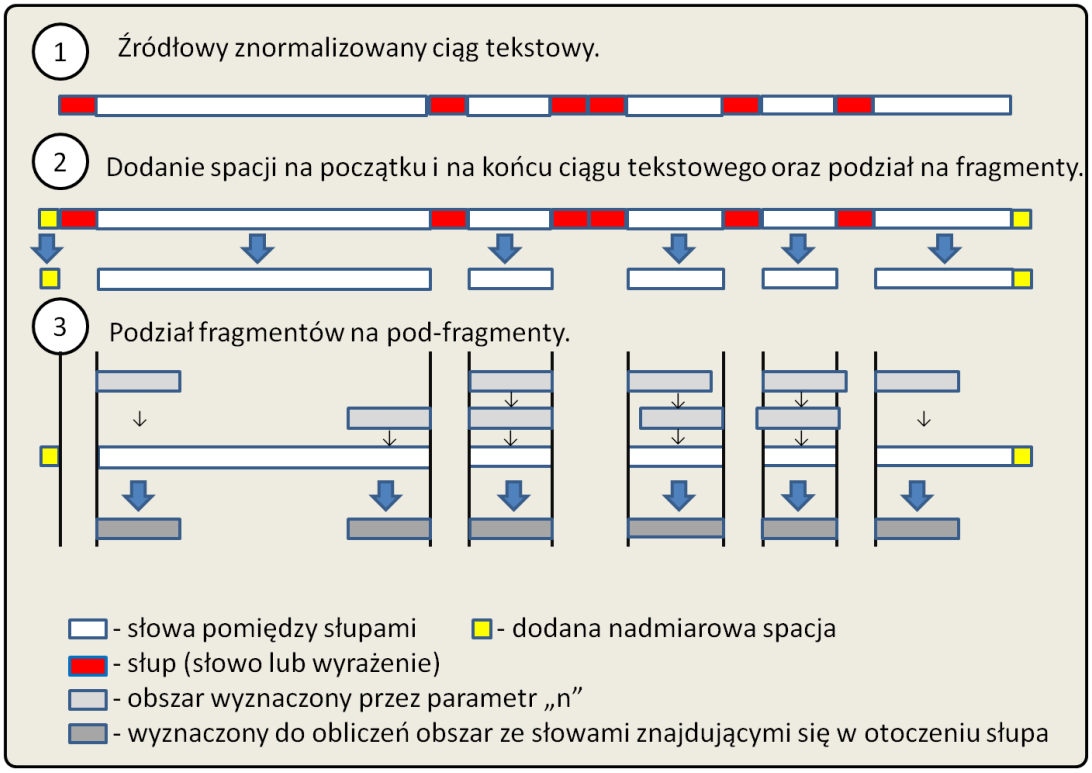
One of such cases involves a situation where the two examined areas of pillars' neighbourhoods overlap. In that case, the words found both in the common section and on its both boundaries have to be counted once. This means that all the words found between the two pillars are counted as the neighbourhood.

The mutual position of two or more pillars can always be analysed as one of the six possible cases of mutual position of two pillars. The procedure is simplified by the algorithm of the computer software for calculating the frequency of words in the neighbourhood of a pillar. The algorithm for calculating the frequency of words in the neighbourhood of a pillar

consists of several steps and assumes that the source material submitted for examination has already been purified and normalised.

In the first step, a single space was added at the beginning and the end of the normalised text string. This ensures that all pillars are in the middle of the analysed text string and allows us to eliminate two special cases, i.e. when the pillar is exactly at the beginning or at the end of the examined text string, which would otherwise complicate the calculation of frequency for the first and the last fragment of the text.

In the second step, the normalised text string with spaces added at the beginning and at the end should be divided into fragments, with the pillar being the dividing line. In this way, the analysed string is divided into fragments. If the number of pillars in a text equals zero, the algorithm terminates its work at this stage.



1. Normalised source text string
2. Adding a space at the beginning and end of the text string and division into fragments
3. Division of fragments into sub-fragments

- words between pillars
- added redundant space
- pillar (word or phrase)
- area defined by parameter “n”
- area selected for calculations, containing words belonging to pillar’s neighbourhood

Fig. 5. Division of the examined text into fragments
Source: own work

In the third step, each fragment is examined in terms of its length, i.e. it shows the actual distance between the pillars in a given fragment of text. Next, the frequencies of occurrence of particular words in fragments are calculated and added up. Partial frequencies calculated for each fragment are aggregated and added up. By adding together the information from each fragment we obtain complete information about the frequency of all words in the pillar's neighbourhood defined by parameter "n". The work of the algorithm was diagrammatically presented in Figure 5.

The procedure resulted in the creation of a set of frequencies of all words found in the specified neighbourhood of each pillar. This data provides the basis for distinguishing the most frequently occurring words and among them - those words which are significant (most frequently occurring) sentiments[8].

Identifying sentiments

As already mentioned, the standard approach is to identify sentiments based on the examination of subjective evaluations of a specific group of people and the available dictionaries. The method does not exploit the full potential of the BigData resources, i.e. the usage of sentiments identified in real time. This goal is achieved by means of ACA, i.e. the procedure of self-learning identification of sentiments. This is a totally innovative product and, as such, requires basic testing (there are no publications or information concerning this procedure), collecting experience (working on large datasets), to determine the accuracy/significance of sentiment identification.

In the first step of the ACA procedure, the sets of the words – presentiments (from the database in relation to time) - occurring most frequently around the pillar are identified . For each of the k presentiments, the frequency is calculated -Wk(t) variable - in time intervals (t_1, $t_{2,\dots}$ t_n). As a result, a set of Wk(t) variables will be singled out.

[8] Y. Hongliang et al., *Identifying sentiment words using an optimization-based model without Seed Words*, https://www.google.pl/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&cad=rja&uact=8&ved=0ahUKEwjEl5W966XQAhUDjSwKHU_OAT4QFgguMAI&url=http%3A%2F%2Fieeexplore.ieee.org%2Fiel7%2F7222492%2F7222827%2F07222836.pdf&usg=AFQjCNHT85uTAy_yYKjle7fngNTF67jctw [accessed: 11/11/2016].

Next, a similar method is used to obtain/calculate the quantitative evaluations (from e.g. market research results, results of surveys by Polish Public Opinion Research Centres such as OBOP and CBOS) for all of j pillars (Sj) it makes points of reference - the one quantitative evaluation Sj (t) (j- name of pillar, value and direction +/– and t - time) in the same time, like Wk (t) intervals. The quantitative evaluation - (Sj (t)) represents the second variable. The value of statistical significance of the relationship between Sj(t) and all variables in the Wk(t) set indicate the validity of choice of the sought-after (real) sentiments - Wf(t). The result of the ACA procedure consists of the most statistically significant (real) sentiments for the selected pillar (j) in a sample limited by the size of the analysed sets (in practice, those are millions of entries). To put it differently, a word (presentiment) which is not popularly regarded as carrying an emotional load may turn out to be a sentiment. For example, the preliminary studies connected with sentiments around the pillar =REFUGEES were the following results Wf(t): Poland, Germany, Europe, EU, Merkel and Ukraine; pillar = HOLIDAYS: photos, oblivion, computer and smartphone; pillar = HEALTH: income, diet, food, smoking, stress and alcoholism.

The result of analysing of the changes of the sentiment*s* Wf(t) of the phenomenon in time intervals (t_1, t_2…., t_n) provides the possibility of predicting the change of the phenomenon in time $t_n + 1$ on the basis of Wf(t) sentiments (predictors). The calculations employ the multiple regression analysis which is used to build a model that is possibly the best adapted to empirical data in time before t_n and enables the evaluation of the status of a phenomenon in time $t_n + 1$. This is due to the fact the regression model is more suitable (statistically significant) for the obtained data (in time $t_n + 1$) than random data.

Conclusion

For obvious reasons, the article does not describe the details of the presented procedure that enables the evaluation of the course of phenomena in the past, in real time and their prediction, mainly thanks to the identification of segments and their correct synthesis[9].

[9] M. Huberty, *Awaiting the second revolution: From digital noise to value creation*, http://eds-1a-1ebscohost-1com-1ebsco.han.buw.uw.edu.pl/eds/detail/detail?vid=3&sid=16298e68-0990-4883-9fc20e0e51a4dab5%40sessionmgr4004&hid=4205&bdata=Jmxhbmc9cGwmc2l0ZT1lZHMtbGl2ZSZzY29wZT1zaXRl#db=edb&AN=101516526, [accessed: 06.05.2015]; Y. Liu, "Big Data and predictive business analytics", "Journal of Business Forecasting" 2015, http://eds-1a-1ebscohost-1com-1ebsco.han.buw.uw.edu.pl/eds/pdfviewer/pdfviewer?sid=16298e68-0990-4883-9fc2-0e0e51a4dab5%40sessionmgr4004&vid=6&hid=4205 [accessed: 05/05/2015].

The research concerning this subject faces a number of related research problems, e.g. difficulties involved in defining unequivocally the name of the object of study, which was the case e.g. during electoral campaign research where a problem arose (which was solved) concerning the multiple terms used to refer to two Polish parties - Citizens' Platform (PO) and Law and Justice (PIS)[10].

The key condition for the success of the RI procedure is having sufficient information to exceed the threshold of no confidence towards the results achieved which would allow drawing reliable conclusions, e.g. predicting a phenomenon. The system has to "learn" - to achieve the assumed statistical significance. One solution (beside the size of the source dataset) would be to intensify the work on using the multiple regression function and take into account more than one parameter - the frequency of sentiments - for making predictions. The idea here is to combine the phenomena/objects of research which are studied independently, but exhibit mutual relationships.

Some of the phenomena/objects of study are too random and cannot be subjected to statistical analysis for the purposes of prediction, using the tools/methods available to date. The problem can be solved by defining the areas for which RI can become the tool of broadly understood diagnosis (e.g. identification of threats).

[10] W. Gogołek, P. Kuczma, *Rafinacja informacji sieciowych…*, op. cit.