Bogdan Gliwa
Anna Zygmunt
Michał Dąbrowski

BUILDING SENTIMENT LEXICONS BASED ON RECOMMENDING SERVICES FOR THE POLISH LANGUAGE

Abstract   *Sentiment analysis has become a prominent area of research in computer science. It has numerous practical applications; e.g., evaluating customer satisfaction, identifying product promoters. Many methods employed in this task require language resources such as sentiment lexicons, which are unavailable for the Polish language. Such lexicons contain words annotated with their emotional polarization, but the manual creation of sentiment lexicons is very tedious. Therefore, this paper addresses this issue and describes a new method of building sentiment lexicons automatically based on recommending services. Next, the built lexicons were used in the task of sentiment classification.*

1. Introduction

Social media has become one of the most important sources of various information and platforms of message exchanging, with a great impact on our lives. This was noticed by big companies a long time ago, so they are especially interested in the study of content on the Internet. Emotions are an integral component of statements in social media. Different groups of users can discuss the same topics in completely different manners, supporting each other or disagreeing. These statements can be classified as objective (expressing factual information) or subjective (containing emotions). and one can value them by expressing an emotional attitude: positive, negative, neutral, objective, or bipolar [26].

A significant increase in interest in the problems of analysis of sentiment can be seen as far back as 2001. Some reasons for such interest in this research area are shown in [21]: the development of advanced methods of analysis of a natural language that was already mature enough so that the methods could be successfully applied in practice, more data and easier availability that were suitable for such analyses (mostly available on the web), and the increasing demand for intelligent applications.

Sentiment analysis has been recognized as a one of the most interesting research areas in computer science [9]. The term "sentiment analysis" (also interchangeably used later with "opinion mining") initially pertained to "automatic analysis of evaluative text and tracking of the predictive judgments" and was closely associated with analyzing market sentiment. Later, the term was rather treated as classifying reviews as to their polarity: either positive or negative. Nowadays, the term refers to "computational treatment of opinion, sentiment, and subjectivity in text" [21].

Automatically discovering the sentiment from entries placed by users in different social media platforms seems to be a difficult task. It is strongly connected with language character (inflected or isolating), language rules, length of entries, context of message, and so on.

2. Related work

2.1. Social media

Internet social media such as online social networking (e.g., Facebook[1], Myspace[2]), blogging (e.g., HuffingtonPost[3]), forums, media sharing systems (e.g., YouTube[4], Flickr[5]), microblogging (e.g., Twitter[6]), wikis, social news (e.g., Digg[7], Slashdot[8]),

---

[1] http://www.facebook.com
[2] http://myspace.com
[3] http://www.huffingtonpost.com
[4] http://www.youtube.com
[5] http://www.flickr.com
[6] https://twitter.com
[7] http://Digg.com
[8] http://slashdot.org

social bookmarking (e.g., Delicious[9]), Opinion, Review, and Ratings Websites (e.g., Epinions[10]) have revolutionized the Internet and the means of communication between people. Users stopped being only consumers of information and became the creators of information. They publish any type of content, ranging from photos, information about their personal lives or news, to reviews and ratings [27]. Product reviews have an impact on building a brand image. Considering the purchase of a product, the potential buyer attaches great importance to the opinions of others [17]. For this reason, it is important for many businesses and organizations to analyze such information. The proper interpretation of opinions allows them to carry out more-effective marketing campaigns and improve product quality, so as to be best-suited to the needs of their consumers.

Most of the content on social media is characterized by emotions; i.e., users express their opinions on the products they have bought. There are various forms of expression of emotions and thoughts beyond textual communication. Some social media sites introduce their own evaluation systems' e.g., marking tweets on Twitter as *favorite* or *retweet*. In social media, the following are the most popular ways of expressing emotions and opinions:

- stars rating – most commonly associated with reviews or products; users can summarize their entry markings by several stars (commonly-used scale of 10 stars),
- emoticons – a pictorial representation of a facial expression, especially popular on Twitter, some services support most popular emoticons and convert their text forms to graphic equivalents,
- graphic signs – specific to any given social media, most popular examples: *thumbs up* or *thumbs down* from *YouTube* or *Like* from *Facebook*.

2.2. Methods of text analysis

Text analysis (text mining) covers many useful techniques to retrieve information from text (in the form of unstructured data). One of the most typical tasks is text classification [1], which involves assigning labels for documents based on their content. An overview of most-commonly-used classifiers in the text mining domain can be found in [10].

An important role in text mining is played by methods of text preprocessing (e.g., stop words removal, words stemming, and lemmatization) and input conversion into structural representation [30]. The Vector space model [12] represents documents as vectors, and each element of the vector usually represents a word (or a group of words) from a collection of documents. Encoding documents into vectors can be conducted using the weighting method (in order to reflect the importance of a word in a specific document), and the most popular weighting method is the algorithm TF-IDF (Term

---

[9]http://delicious.com
[10]http://www.epinions.com

Frequency – Inverted Document Frequency) [24]. This is based on the assumptions that the importance of a word is proportional to the number of occurrences of this word in a document and inversely proportional to the number of documents in which the word occurred. The TF-IDF method can be used to find keywords from the text, but it fails to find a connection between semantically convergent documents that utilize different vocabularies, and more-complex methods need to be applied (such as Topic Modeling [7]).

Topic Modeling [13] is a statistical technique that uncovers hidden, abstract "topics" that occur in a collection of documents. "Topic" is defined as a set of words that co-occur in many documents and, therefore, are presumed to have similar semantics. One of the biggest advantages of this approach is the possibility of finding similar texts even if they incorporate different vocabulary. Moreover, topic modeling can be used to reduce the representation of documents (e.g., using topics as features describing the text instead of all words). There are two main branches of Topic Modeling [7] – algorithms based on Singular Value Decomposition (such as Latent Semantic Indexing [13]) and those based on probabilistic generative processes [5] (such as Latent Dirichlet Allocation [3, 4]).

2.3. Methods of sentiment analysis

Several methods for sentiment analysis have been proposed [16] that can be conducted at different levels of granularity: document (the whole document expresses opinions), sentence (each sentence expresses opinions), entity, and aspect (the most fine-grained analysis). The most general classification of these methods is for supervised machine learning and unsupervised [9].

2.3.1. Supervised methods

In this approach, there is the assumption that classes to which the document should be assigned are known earlier, and training and test sets are used for the classification. A training set consists of an input feature vector with class labels. Then, a classification model is built, and its accuracy is verified based on a test set. The most widely used classifiers [2, 29] for sentiment analysis are Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM). The text features that are most commonly taken into consideration are the frequency of occurrence of the word, part of speech of the word, the presence of the word, and the presence of negation.

In the simplest form, NB classifies texts poorly; but quite intuitive assumptions of the algorithm often make it an initial point for more-advanced solutions. The number of attends was taken to improve the classifier, and the results are often similar to the more-complex ones, such as SVM [25].

Methods based on the SVM classifier perform the best sentiment analysis; measurements of accuracy even show $86-89\%$ [18]. A significant limitation of the supervised methods is the necessity of learning the data set. Usually, this is related to the need for greater computing power. Despite this, they are used successfully; for example, NB achieves very good results, as it is relatively easy to use and understand [19].

2.3.2. Unsupervised methods

Methods based on a sentiment dictionary are representative of an unsupervised class of methods. A special feature of this approach is the lack of need for a learning stage; so for new data, dictionary methods can be used immediately. A sentiment dictionary consists of a set of words with assigned sentiment factors (weight of words).

A typical example of a dictionary method is an approach proposed by Turney in [28], where relations between the words are not considered. An interesting solution is to use the value of sentiment in numerical form with regard to only two adjectives: *excellent* and *poor*. A text is treated as a multiset of words. Using aggregation (sum or average), sentiment factors of each word are used to calculate the overall sentiment of the text. In [8] *SentiWordNet*[11], a dictionary built on the basis of the *WordNet*[12] dictionary, was used to analyze texts written in different languages. The first step was the automatic translation of each text into English, as SentiWordNet consists of only English words. Tests were conducted on German reviews, and the results were slightly weaker than with other dictionary methods.

Another interesting solution to sentiment classification using detection of the parts of speech is presented in [23]. In this approach, texts are analyzed for sentiment content – mostly adjectives. Only sentences in which there is at least one word with sentiment are taken into consideration (and the rest are discarded). For unknown earlier words, an Internet search engine is used to find n most similar words with known sentiment (a sentiment dictionary is used). The value of the sentiment of unknown words is calculated on this basis.

2.3.3. The accuracy of existing methods

Evaluating the quality of methods is not easy, since it is strictly connected with the specificity of the tests and the language of the analyzed texts. Currently, most studies have been conducted for the English and Chinese languages. The characteristics of these languages (both are isolating languages) simplifies sentiment analysis to some extent. Most classifiers identify only two classes: positive and negative. Evaluation of the text neutrality is not a trivial problem, even for people. For this reason, a lot of effort is put into including even some aspects of emotions into classified texts. To create test sets, reviews and opinions from social services are mainly used. In [6], a comparison of supervised and unsupervised methods was made. Tests were performed on 100 movie reviews. Using machine learning, the accuracy of classification was 85%, and with the unsupervised method – 77%. Turney [28] in the approach based on the dictionary method, achieves an accuracy of 65.83%. These tests were conducted by classifying 120 movie reviews.

In [31], the method of building a sentiment dictionary was proposed for the Chinese language.They created three dictionaries, and the calculated *F-measure* for them was 69.23%, 69.93%, and 77.83%, respectively. Using *SentiWordNet* in [8] allows

---

[11]http://sentiwordnet.isti.cnr.it
[12]https://wordnet.princeton.edu/

us to achieve an accuracy of about 66%. Much higher accuracy was obtained using different classifiers [22]: about 78% (NB), about 79% (ME), and about 81% (SVM).

The results show that the more-accurate methods are those that are based on machine learning. But these classes of methods require a time-consuming stage of learning, so they cannot be used ad-hoc. On the other hand, the accuracy of dictionary methods is largely dependent on the quality of the sentiment dictionary as well as the language rules used in the analysis of texts. In addition, the key is to identify the context in which the word was used.

2.4. Building sentiment dictionaries

Several approaches are related to defining some seeds of words (manually defining sentiment values for those words) and propagating information about sentiment to other words by some criteria. One example of this approach is a method proposed by Turney [28], who defined sentiment for other words by assessing co-occurence with words "excellent" and "poor" (to assess the number of co-occurences with those words, he used NEAR operator from the AltaVista search engine).

Hatzivassiloglou and McKeown [11] deduced the polarity of words by considering linguistic links between words or phrases; i.e., they treated adjectives connected by "and" to have the same polarity and those connected by "but" to have the opposite polarity.

Kamps et al [14] proposed a new method for defining sentiment for words. They constructed graphs based on WordNet[13] using synonymy relationship encoded in WordNet synsets. Semantic orientation (sentiment) was calculated by an introduced EVA measure that incorporates geodesic distance using the shortest path in this graph between words (in EVA metrics, they used geodesic distance with words "good" and "bad").

Kim and Hovy [15] developed a model that expands the small list of seed verbs and adjectives (manually annotated) by using relations in WordNet. They used synonym and antonym relations.

Pak and Paroubek [20] built a sentiment lexicon using Twitter[14]. They downloaded tweets and divided them into positive and negative tweets depending on the inclusion of positive or negative emoticons. Words that frequently appeared in the positive dataset and rarely in the negative one have a higher value of polarity (a new measurement introduced by the authors to assess sentiment), and vice versa.

Generally, sentiment analysis methods based on classifiers give better results; but using them is more difficult and laborious (the necessity of preparing and tagging the data set used by classifier to learn). Most studies on sentiment analysis have been done in the English language. The most popular representation of sentiment dictionary is *SentiWordNet*. However, a Polish sentiment dictionary is still lacking.

---

[13] http://wordnet.princeton.edu/
[14] https://twitter.com/

Only a few methods to create a sentiment dictionary in an automatic manner were proposed. One of the main difficulties is to identify the context: there may be different polarizations for the same words used in different contexts.

3. Method description

Developing our method, we had to solve problems related to not only the construction of a sentiment dictionary but also using it to calculate the sentiment of any text. Thus, the process consists of two main parts:

- building a sentiment dictionary,
- analysis of the text sentiment using the sentiment dictionary.

3.1. Building a sentiment dictionary

Creating a sentiment dictionary requires the appropriate data with emotional content as well as information about the strength of this emotional content. An example of data meeting these assumptions are opinions and reviews in social media. They consist of text comments and stars; for example, ("I recommend this book!", 10/10) or ("Waste of time", 1/10). Properly preparing these texts, one can calculate measures of sentiment for all sentences.

The process of creating a semantic dictionary consists of several steps, such as:

- selecting the appropriate social media,
- developing a common model of reviews in the database,
- crawling data from selected social media and storing in the database,
- language preprocessing of data,
- calculating a sentiment factor of the words,
- creating a sentiment dictionary,
- verifying the effectiveness of the generated sentiment analysis.

Processing data with reviews

The goal of this step is to count the occurrences of each word in reviews marked by the specified number of stars. Intuitively, it can be said that the words that are more common in the opinions of a large number of stars (positive) have a positive sentiment. The content of each opinion is divided into smaller chunks called tokens, based on white-space characters. Tokens are words that are the smallest components in the sentiment analysis. Punctuation marks are not deleted, since they will be used for the detection of negations. Then, the words are converted to base forms (words stemming). This is done to reduce the size of the dictionary; this is particularly important in the case of the Polish language (which is an inflected language). At this stage, the *Morfologik*[15] library was used. This library also enables the improvement of typos, which is extremely useful since the content on social media often has a large

---

[15]http://www.morfologik.blogspot.com

number of typos. Then, linguistic rules (for example, the detection of negations) are applied to the data, which results in a special marking of selected words. The next step is to count the occurrences of each word with regard to the number of stars of the opinion. At the same time, using *Morfologik*, the part of speech of each word is detected. In our study, we considered adjectives, adverbs, and verbs. Nouns are ignored because most of them are neutral or highly contextual (i.e., they are positive or negative depending on the domain). Stop words are omitted since they do not express any emotion. A list of stop words for the Polish language was derived from *Wikipedia*[16].

Sentiment dictionary creation

A sentiment dictionary consists of the triples: {word, sentiment factor, part of speech}, where a *sentiment factor* specifies an emotional charge and is expressed by the following numerical values:

- if sentiment factor equals 0 – the word is neutral,
- if sentiment factor is less than 0 – the word is negative,
- if sentiment factor is greater than 0 – the word is positive.

It seems that, having calculated the number of occurrences of a word in the breakdown of the number of stars of reviews, one can determine its sentiment (e.g., a word that more frequently occurs in negative reviews is assumed to be negative). It turned out, however, that drawing conclusions only on the basis of the frequency of occurrences of the word is not sufficient (mostly due to the imbalance of opinions with different polarity), and therefore, a new metric – *the relative ratio of the word occurrences*, was introduced (1).

$$r_x = \frac{n_x}{s_x}, \tag{1}$$

where

r_x – relative ratio of word occurrences in x-stars opinions,

n_x – number of word occurrences in x-stars opinions,

s_x – number of occurrences of all words in x-stars opinions.

It defines correctly if the word is positive or negative but is not suitable to determine how great the emotional charge is. Therefore, *the comparative ratio of the word occurrences* was defined (2).

$$w_x = \frac{r_x}{\sum\limits_{i=1}^{k} r_i}, \tag{2}$$

where

w_x – comparative ratio of word occurrences in x-stars opinions,

r_x – relative ratio of word occurrences in x-stars opinions,

---

[16]https://pl.wikipedia.org/wiki/Wikipedia:Stopwords

k – max number of stars in opinions,

$\sum_{i=1}^{k} r_i$ – sum of the relative ratio of word occurrences.

The next step is to obtain a single value that specifies the sentiment of the word based on *the comparative ratio of the word occurrences* (2) and *weight vector* (4) for each value of review rating stars. *Weight vector* is defined:

$$\overrightarrow{v} = [\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_k], \alpha_i = -\frac{k-1}{2} + (i-1) \tag{3}$$

We used the scale of ratings in the range 1–10 (for datasets with different scaling, we rescale to this range), so the values of *weight vector* have the following form:

$$\overrightarrow{v} = [\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_{10}], \alpha_1 = -4.5, \alpha_2 = -3.5, \alpha_3 = -2.5, \ldots, \alpha_{10} = 4.5 \tag{4}$$

Sentiment factor of the word (5) is the sum of the products of the *comparative ratio of the word occurrences* and the corresponding weights from *weight vector*.

$$s_w = \sum_{i=1}^{k} (w_i \times \alpha_i) \tag{5}$$

If the word has a *sentiment factor* value equal to 0 ($s_w = 0$), this means that the word is neutral or has neutral polarity; if $s_w < 0$, then polarity is negative; otherwise, ($s_w > 0$) is positive. As it is a rare situation, it is convenient to establish upper and lower limitations (epsilon) within which the word will be treated as neutral. In our research, we adopted this solution to each part of speech:

$$s_w >= \epsilon_1 \wedge s_w <= \epsilon_2 \quad \text{– neutral polarity}$$
$$\epsilon_1 \text{ – lower epsilon,} \qquad \epsilon_2 \text{ – upper epsilon}$$

For example: in our experiments, we assumed that epsilon for verbs is: $\epsilon_1 = -0.8$ and $\epsilon_2 = +0.8$ and for adjectives and adverbs: -0.3 and $+0.3$ (those values were determined empirically).

3.2. Using a sentiment dictionary

Having built a sentiment dictionary, we can use it to determine the sentiment content of any text. In the beginning, all negations are detected in the text and marked properly (negation rule is quite simple – we are looking in a sentence for the word *nie* (English: *not*) and all words in the distance of 4 words around it (or less, if we found the end of the sentence or its part separated by comma) are negated, and such negations are added to lexicon with suffix *_NEG*). Then, the words in the text are converted to base forms (lemmatization). The text is analyzed word by word, and using the sentiment dictionary, a list of words (positive, negative, and neutral) is created. The list stores information about the name of the word, part of speech, and

its value of sentiment. The determination of the overall sentiment for the text is based on the calculated sum of the positive and negative sentiment factors. On the basis of the two values, it is determined what percentage of the statement is positive and to what extent negative. The absolute value is calculated as the difference between the two sums of factors. If it falls between the lower and upper epsilon boundaries, the text is marked as neutral. Otherwise, both sums are compared: if the sum of factors of the positive words is greater than the sum of factors of the negative words, the text is labeled as positive; otherwise – negative. If the text does not contain words with sentiment charge (i.e., not found in the selected sentiment dictionary), it is marked as neutral. Lower and upper epsilon boundaries are set to values −0.3 and 0.3 (lower epsilon differentiate between negative and neutral polarity, upper epsilon – between neutral and positive).

4. Results

This section provides the description of datasets as well as the generated sentiment lexicons and obtained results using them.

4.1. Datasets

Experiments were conducted on three crawled datasets from the following websites: *lubimyczytac.pl*, *opineo.pl*, and *rankinglekarzy.pl*. The datasets contain opinions (in Polish) of people on various subjects – each dataset is dedicated to a different area.

4.1.1. lubimyczytac.pl

lubimyczytac.pl is a website consisting of reviews of books. This site was built in 2010 and now is one of the most popular Polish sites regarding books opinions. Users can express their opinions about books in the form of comments and reviews; they can also make friends with other book readers. Reviews can be marked using stars from 1 (very bad) to 10 (great). In this dataset, there are also opinions with 0 stars; but this means no mark, so we do not utilize such reviews in our experiments. The prepared dataset contains 445,134 reviews. Figure 1 shows the distribution of opinions in this dataset.

4.1.2. Opineo.pl

Opineo.pl is an independent service comprised of opinions about products, online stores, and companies. Products are divided into categories such as computers, cars, or home. The prepared dataset contains 593,914 opinions. Reviews are marked from 0 stars (very bad) to 5 (great). Figure 2 presents the distribution of marks for reviews.

4.1.3. RankingLekarzy.pl

RankingLekarzy.pl is a website with opinions about doctors. Physicians can create their own profiles and share information about their specializations and ranges of diseases they treat, whereas patients can express their feelings about visits at the

doctor's office and have the possibility to make an appointment with a doctor. Reviews are marked from 1 star (very bad) to 5 (very good) in several categories, such as doctor competency, explanation of treatment, kindness, complexity of the treatment's approach, the amount of time spent with the doctor, and whether or not the patient would recommend that doctor. The final mark for a doctor is an average of all marks from patients; for our experiments, we rounded those average mark to have the form of a whole number. Figure 3 contains the distribution of doctors marks in this dataset. The dataset is comprised of 251,724 reviews.
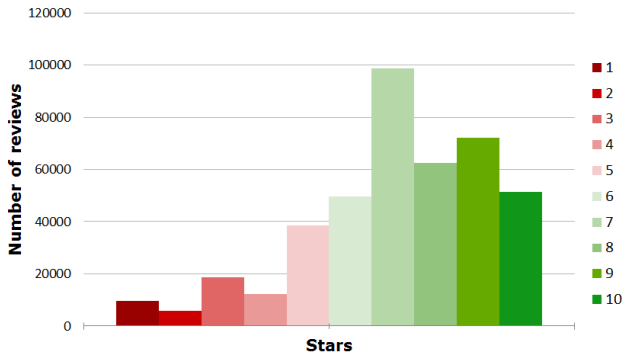

Figure 1. Distribution of opinion marks in *lubimyczytac.pl* dataset.
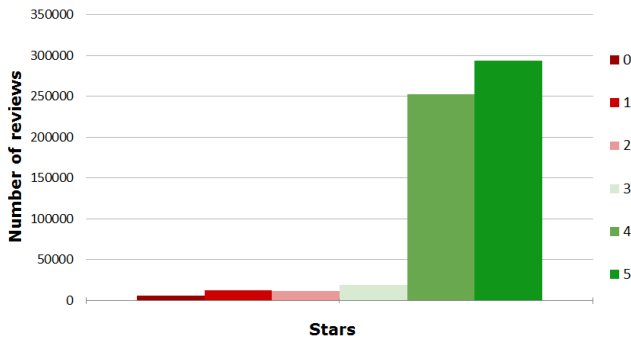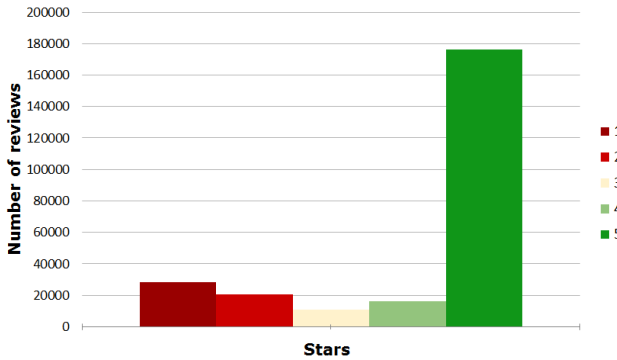

Figure 2. Distribution of opinion marks in *opineo.pl* dataset.

4.1.4. Preparing datasets to experiments

Each dataset covers different domains (books, products, or doctors), but the majority of opinions in them are positive (as we notice in figures 1, 2, and 3). Datasets have also different systems of marks, so we needed to unify it. We decided to use a scale from 1 to 10 (as in *lubimyczytac.pl*), and other systems of marks are scaled proportionally.

Figure 3. Distribution of opinion marks in *RankingLekarzy.pl* dataset.

4.2. Comparison of dictionaries

Using the datasets described earlier, three sentiment dictionaries were built (each dictionary was created based on a single dataset). We used 20,000 reviews (10,000 positive reviews and 10,000 negative ones) from each dataset to conduct experiments, and we divide it into two parts by stratified sampling: 14,000 reviews used to generate the dictionary (7000 positive reviews and 7000 negative ones) and 6000 reviews used to assess the quality of method of sentiment calculation (half of the reviews are positive and half are negative). A single word from these datasets should be used at least 30 times to be incorporated into the built dictionaries. A comparison of the generated dictionaries can be found in Table 1. Negation words presented in this table are words with negation found in the process of building a sentiment lexicon. As we can notice, the biggest is the dictionary created on the *lubimyczytac.pl* dataset. It is related with the fact that reviews on this portal are the longest from tested datasets. Another interesting fact is that, on *opineo.pl* and *rankinglekarzy.pl*, the highest number of words in the generated dictionaries have negative sentiments; but on *lubimyczytac.pl*, the situation is different – the positive words constitute the highest part of all words in that dictionary.

Table 1

Comparison of dictionaries.

	lubimyczytac.pl	*opineo.pl*	*rankinglekarzy.pl*
all words	6500	1297	979
common words		632	
positive words	2647	282	321
neutral words	1606	177	155
negative words	2247	838	503
negation words	1122	181	178

Figure 4 depicts some words in different lexicons. Words that are green have positive polarity; in red – negative; and gray – lexicons do not agree on the polarity of the analyzed word. The size of the fonts maps the intensity of the polarity. Due to the large number of words in the dictionaries, the presented diagram contains only those words with the highest absolute sentiment polarity. Words that are common in all lexicons and have the same polarity include (among others) *odradzać* (in English: *dissuade, discourage*), *uwielbiać* (English: *adore, admire*), *polecać* (English: *recommend*), and *świetnie* (English: *great*). The *lubimyczytac.pl* lexicon contains the most-sophisticated words among all lexicons, such as *syzyfowy* (English: *Sisyphean*) and *bezosobowy* (English: *impersonal*). *Opineo.pl* includes many words describing features or defects of products; e.g., *nieszczelny* (English: *leaky*), and *pojemny* (English: *capacious*). *RankingLekarzy.pl* have mostly words describing attitude or behavior of people; e.g., *niedelikatny* (English: *indelicate*) and *profesjonalnie* (English: *professionally*). However, the most interesting are words that have different polarity in different lexicons, such as *chiński* (English: *Chinese*) and *płakać* (English: *cry*). For products, *Chinese* is often a synonym for cheap, low-quality products (negative polarity); but in books, the polarity of this word is positive – it may be connected with *Chinese* fairy tales or stories (which are rated highly). Similarly, *cry* in the context of doctor and disease is not a positive thing; but in the context of books, it may mean that the story is touching and moving that makes the reader cry (which means that the book is very good). These examples indicate that it is very hard to prepare a lexicon for general use in an automatic way because the context matters in many cases.

4.3. Results of sentiment classification

Experiments were conducted for each sentiment lexicon (each dictionary was built on a single dataset), and each lexicon was tested on three test datasets covering unseen reviews (6000 messages from each dataset).

4.3.1. Generated lexicon based on *lubimyczytac.pl* dataset

Table 2 covers results using the lexicon built on the *lubimyczytac.pl* dataset. The best results are achieved on *lubimyczytac.pl* dataset, and the worst – *opineo.pl* (classification had the lowest recall on this dataset).

Table 2

Results of sentiment classification using sentiment lexicon built on *lubimyczytac.pl* dataset.

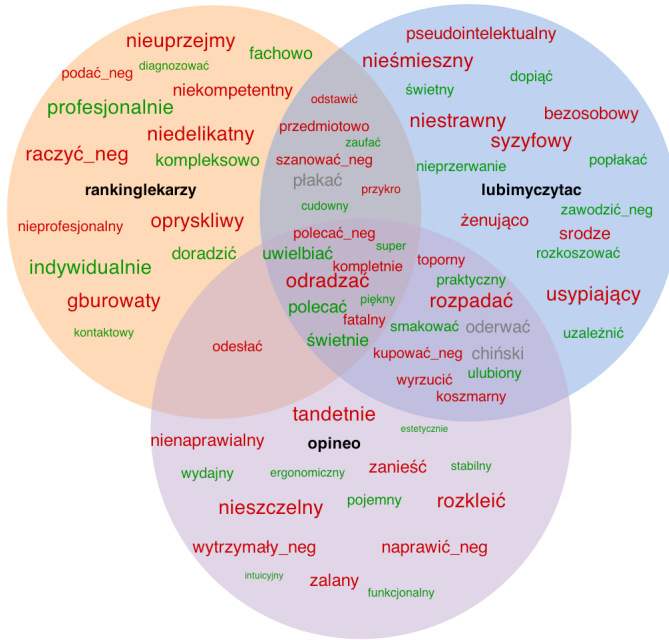| Test data	Accuracy	Precision	Recall	F-measure
lubimyczytac.pl	87%	0.84	0.91	0.87
opineo.pl	62%	0.67	0.48	0.56
rankinglekarzy.pl	78%	0.76	0.81	0.78

Figure 4. Comparison of words in dictionaries using Venn diagram.

4.3.2. Generated lexicon based on *opineo.pl* dataset

Table 3 presents results using the lexicon built on the *opineo.pl* dataset. The best results are obtained using test data from *opineo.pl*, as we expected. The worst are on the *lubimyczytac.pl* dataset. This suggests that these datasets differ the most from each other.

Table 3

Results of sentiment classification using sentiment lexicon built on *opineo.pl* dataset.

Test data	Accuracy	Precision	Recall	F-measure
lubimyczytac.pl	58%	0.65	0.36	0.46
opineo.pl	84%	0.87	0.79	0.83
rankinglekarzy.pl	79%	0.84	0.72	0.78

4.3.3. Generated lexicon based on *RankingLekarzy.pl* dataset

Table 4 shows results using the lexicon built on the *RankingLekarzy.pl* dataset. It is not surprising that the best results are accomplished on the *RankingLekarzy.pl* dataset and the worst are on the *lubimyczytac.pl* dataset.

Results of sentiment classification using sentiment lexicon built on *rankinglekarzy.pl* dataset.

Test data	Accuracy	Precision	Recall	F-measure
lubimyczytac.pl	69%	0.75	0.57	0.65
opineo.pl	75%	0.77	0.7	0.73
rankinglekarzy.pl	89%	0.89	0.9	0.89

4.3.4. Summary of results

As we notice in Tables 2, 3, and 4, the best results are obtained when we test a generated lexicon on the same dataset that was used to create it (we always tested on reviews not used to generate a lexicon). This means that the vocabulary used in datasets is of paramount importance. The lexicon based on *RankingLekarzy.pl* achieved good results on all tested datasets; when we look at Table 1, one can see that common words in this dictionary constitute the largest part of the whole dictionary (as compared to the others). This may suggest that this lexicon is the most universal from those created; other dictionaries, however, contain many words specific to the domains for which they were built.

5. Sentimeter – application for analyzing sentiment in texts

This section provides a description of the implemented application useful to visualizing the results of sentiment classification as well as the details of use cases on a few messages retrieved from Twitter.

5.1. Sentimeter's features

Sentimeter is a web application (*single-page application*) built to interactively test sentiment in texts. Its main visual interface is presented in Figure 5. For typed input text, the user can choose a lexicon used to sentiment classification and visualize the results in the form of a chart representing the proportion of positiveness and negativeness of an analyzed sentence. Below the chart, we can see words from an analyzed sentence that transfer polarity (according to the selected sentiment lexicon). For each polarity word, the application also shows information about the factor (see Section 3) and part of speech.

As an example of input text showed in that figure, we used the following sentence in Polish: *Film ma wiele negatywnych recenzji, ale ja uważam, że był pouczający* (English: *The movie has a lot of negative reviews, but I thought it was illuminating*). We can notice that the sentiment classification was conducted using a lexicon built based on *lubimyczytac.pl*, and this sentence was classified as negative. Two words were found in the sentiment dictionary – the first is positive and the second is negative (look at values of factors). But overall, the results are negative.

Apart from conducting sentiment classification and visualization of its results, *Sentimeter* may be used to browse and search words in lexicons as presented in Figure 6.
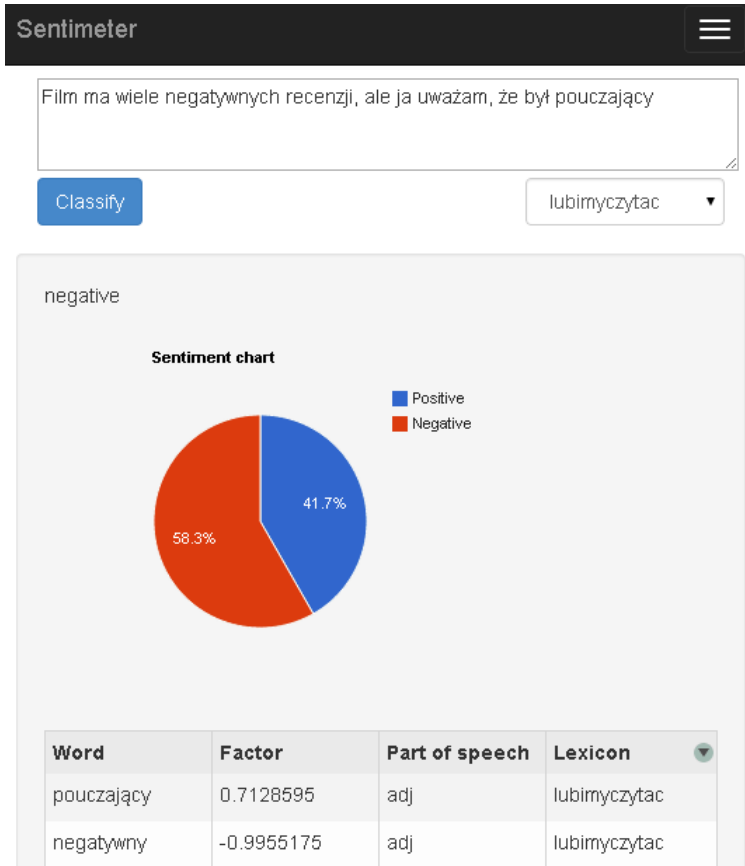

Figure 5. Visual interface of classification results in *Sentimeter* application.

5.2. Case studies for Twitter messages

Apart from testing on previously-mentioned datasets (*lubimyczytac.pl*, *opineo.pl*, and *rankinglekarzy.pl*), we also decided to test results on Twitter messages[17] concerning different topics. For Twitter messages, we do not have any values (such as the number of stars) describing sentiment of its content. Therefore, such messages have to be verified manually, and we present the results of the sentiment classification for a few messages with some interpretation.

---

[17]www.twitter.com

Figure 6. Visual interface of browsing sentiment lexicon in *Sentimeter* application.

5.2.1. Case 1

Polish: *Licytowanie liczby ofiar w Tunezji to kolejny dowód na nieodpowiedzialność obecnej władzy. Kompletny brak profesjonalizmu.*
English: *Bidding on the number of victims in Tunisia is further proof of the irresponsibility of the current government. A complete lack of professionalism.*

Figure 7 presents the results of sentiment classification for methods utilizing different lexicons. The meaning of the analyzed message has a negative connotation, and all methods based on the various lexicons agree with that. In all sentiment dictionaries, the word *kompletny* (English: *complete*) has a strong negative polarity, and this word determines the result. Another word that was recognized in all sentiment dictionaries was *kolejny* (English: *another*); but in the *lubimyczytac* lexicon, this word is positive (yet it is negative in the others). It is worth noting that the presented approach misses discovering the sentiment for nouns during the creation of the polarity lexicon (sentiment polarity can only be assigned for adjectives, adverbs, and verbs). But in this particular case, nouns transfer the majority of sentiment polarity. Despite this fact, the lexicons include some words that could determine the result.
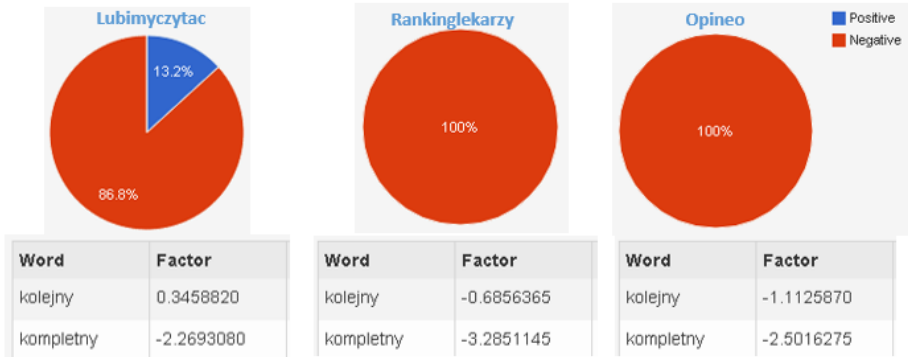
Word	Factor
kolejny	0.3458820
kompletny	-2.2693080

Word	Factor
kolejny	-0.6856365
kompletny	-3.2851145

Word	Factor
kolejny	-1.1125870
kompletny	-2.5016275

Figure 7. Case 1 for sentiment classification based on different lexicons.

5.2.2. Case 2

Polish: *Moim zdaniem urządzenia Apple mają po prostu fatalny współczynnik cena/jakość, dlatego drugi raz ich nie kupię.*
English: *In my opinion, Apple has a terrible price/quality ratio, so I simply will not buy their products again.*

Figure 8 shows the results for the second case. The overall sentiment for the message is negative, and all lexicons concurred. The word *fatalny* (English: *terrible*) has strong negative sentiment in all sentiment dictionaries. *Lubimyczytac* and *Opineo* also contain the word *kupić_neg* (English: negation of *buy*), which also has a negative connotation. Moreover, *rankinglekarzy* also includes the word *prosty* (English: *simple*). In this case, we can see that transformation of word *prostu* to *prosty* is wrong (whole expression *po prostu* means *simply*). The *Opineo* lexicon also contains the word *drugi* (English: *second*) with negative polarity. This can be explained by the fact that some expressions like *z drugiej ręki* (English: *second-hand*, meaning something not new), which may be related to items having worse quality that new ones.

5.2.3. Case 3

Polish: *Z każdym kolejnym meczem rozumiem bardziej czemu Lewandowski jest tyle wart, instynkt napastnika niesamowity :)*
English: *With each subsequent game the more I understand why Lewandowski is so much worth, his striker instinct is amazing :)*

Figure 9 depicts the results of sentiment classification for this case. This message has positive polarity, and all lexicons agree with this result. The word *niesamowity* (English: *amazing*) has strong positive polarity in all dictionaries. Another word present in all of them is *kolejny* (English: *subsequent*), and it has various polarity in those lexicons (actually, this word should not transfer any sentiment polarity whatsoever). *Lubimyczytac* and *Opineo* also contain the word *rozumieć* (English: *understand*) with

negative polarity. Moreover, in the *opineo* lexicon, we additionally can find the word *wart* (English: *worth*) with a very strong positive sentiment value.
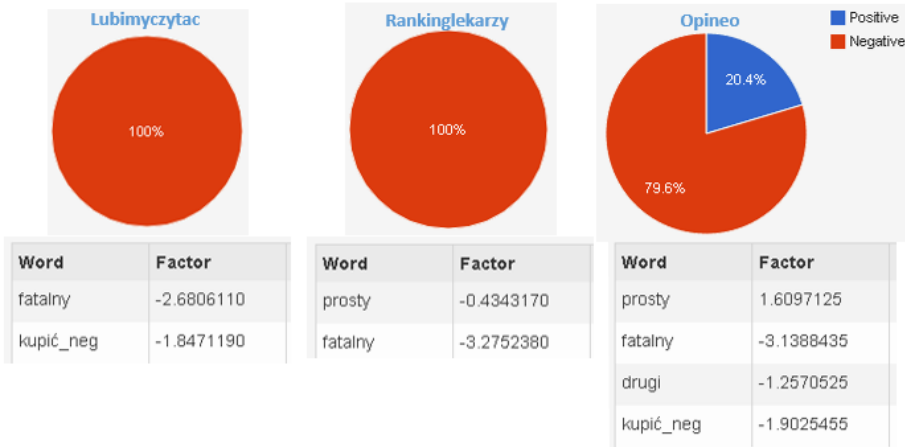

Word	Factor
fatalny	-2.6806110
kupić_neg	-1.8471190

Word	Factor
prosty	-0.4343170
fatalny	-3.2752380

Word	Factor
prosty	1.6097125
fatalny	-3.1388435
drugi	-1.2570525
kupić_neg	-1.9025455

Figure 8. Case 2 for sentiment classification based on different lexicons.

Word	Factor
kolejny	0.3458820
niesamowity	2.1492445
rozumieć	-0.8453850

Word	Factor
niesamowity	2.6739110
kolejny	-0.6856365

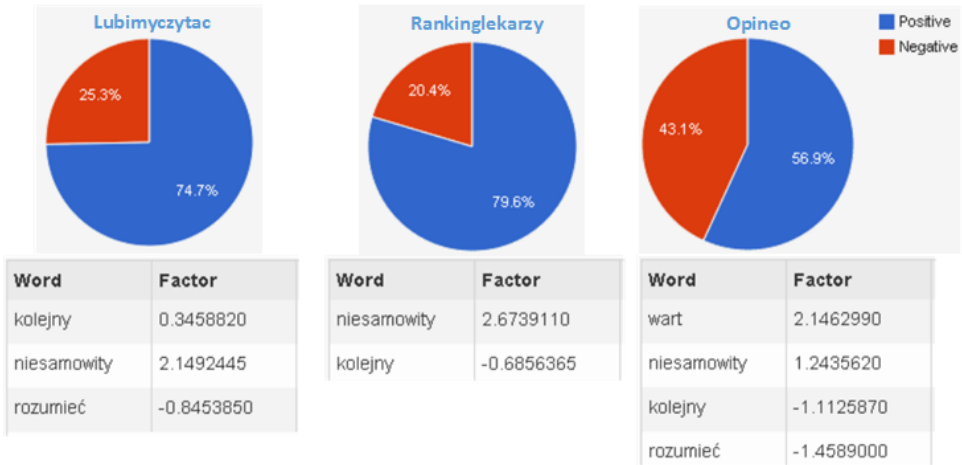| Word	Factor
wart	2.1462990
niesamowity	1.2435620
kolejny	-1.1125870
rozumieć	-1.4589000

Figure 9. Case 3 for sentiment classification based on different lexicons.

6. Conclusion

In this paper, we presented a new method of building sentiment lexicons for datasets containing tagged messages by numbers of stars (expressing sentiment polarity for the message). Rating messages in some scale is typical for recommender systems; in this

study, we used some real-world datasets containing user recommendations on some aspects – *lubimyczytac.pl* (book reviews), *rankinglekarzy.pl* (patient recommendations about doctors), and *opineo.pl* (product reviews). The built polarity lexicons were tested on a prepared sample of reviews from those datasets. The achieved results were quite high, but experiments showed that the word transferring sentiment can have different polarization in different domains. Furthermore, we also assessed the quality of sentiment classification using created lexicons on some messages published in Twitter.

To sum up, the proposed method can generate good quality sentiment lexicons automatically, which is important (especially in those applications where such lexicons do not exist). This method also has some limitations. It is very hard to create universal sentiment dictionary that can be used in different domains – we do not consider the context of words to calculate a sentiment value, and we have to remember that polarities of words may differ in various contexts. Another limitation can be the consideration of single words when some phrases often have different meanings and different polarities than the individual words.

Future directions of research concern building sentiment lexicons for the English language and their comparison with some polarity lexicons available for that language. We are also planning to test different methods of sentiment classification using polarity lexicons.

Acknowledgements

References

[1] Aggarwal C.C., Zhai C.: *A Survey of Text Classification Algorithms*, pp. 163–222. Springer US, Boston, MA, 2012, ISBN 978-1-4614-3223-4, `http://dx.doi.org/10.1007/978-1-4614-3223-4_6`.

[2] Anjaria M., Guddeti R.M.R.: A novel sentiment analysis of social networks using supervised learning. *Social Network Analysis and Mining*, vol. 4(1), pp. 1–15, 2014, ISSN 1869-5469, `http://dx.doi.org/10.1007/s13278-014-0181-9`.

[3] Blei D., Lafferty J.: Dynamic topic models. In: *Proceedings of the 23rd international conference on Machine learning*, pp. 113–120, 2006.

[4] Blei D., Ng A., Jordan M.: Latent Dirichlet allocation. *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[5] Blei D.M.: Probabilistic Topic Models. *Communications of the ACM*, vol. 55(4), pp. 77–84, 2012, ISSN 0001-0782, `http://doi.acm.org/10.1145/2133806.2133826`.

[6] Chaovalit P., Zhou L.: Movie Review Mining: A Comparison Between Supervised and Unsupervised Classification Approaches. In: *Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05)*, HICSS '05, IEEE Computer Society, 2005.

[7] Crain S.P., Zhou K., Yang S.H., Zha H.: *Dimensionality Reduction and Topic Modeling: From Latent Semantic Indexing to Latent Dirichlet Allocation and Beyond*, pp. 129–161. Springer US, Boston, MA, 2012, ISBN 978-1-4614-3223-4, `http://dx.doi.org/10.1007/978-1-4614-3223-4_5`.

[8] Denecke K.: Using SentiWordNet for multilingual sentiment analysis. In: *Proceedings of the 24th International Conference on Data Engineering Workshops*, pp. 507–512, IEEE Computer Society, 2008.

[9] Feldman R.: Techniques and Applications for Sentiment Analysis. *Commun. ACM*, vol. 56(4), pp. 82–89, 2013, ISSN 0001-0782, `http://doi.acm.org/10.1145/2436256.2436274`.

[10] Harish B., Guru D., Manjunath S.: Representation and classification of text documents: A brief review. *IJCA, Special Issue on RTIPPR (2)*, pp. 110–119, 2010.

[11] Hatzivassiloglou V., McKeown K.R.: Predicting the Semantic Orientation of Adjectives. In: *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, EACL'97, pp. 174–181, Association for Computational Linguistics, Stroudsburg, PA, USA, 1997, `http://dx.doi.org/10.3115/979617.979640`.

[12] Hotho A., Nürnberger A., Paaß G.: A Brief Survey of Text Mining. *LDV Forum – GLDV Journal for Computational Linguistics and Language Technology*, vol. 20, pp. 19–62, 2005.

[13] Huang Y.: Support vector machines for text categorization based on latent semantic indexing. Tech. rep., Electrical and Computer Engineering Department, The Johns Hopkins University, 2003.

[14] Kamps J., Marx M., Mokken R.J., Rijke M.D.: Using wordnet to measure semantic orientation of adjectives. In: *National Institute for*, pp. 1115–1118, 2004.

[15] Kim S.M., Hovy E.: Determining the Sentiment of Opinions. In: *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Association for Computational Linguistics, Stroudsburg, PA, USA, 2004, `http://dx.doi.org/10.3115/1220355.1220555`.

[16] Liu B.: *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 2012, `http://dx.doi.org/10.2200/S00416ED1V01Y201204HLT016`.

[17] Mudambi S.M., Schuff D.: What makes a helpful online review? A study of customer reviews on Amazon.com. In: *MIS Quarterly*, pp. 185–200, 2010.

[18] Mullen T., Collier N.: Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 412–418, 2004.

[19] Niu Z., Yin Z., Kong X.: Sentiment Classification for Microblog by Machine Le-arning. In: *Proceedings of the 2012 Fourth International Conference on Compu-tational and Information Sciences*, pp. 286–289, IEEE Computer Society, 2012.

[20] Pak A., Paroubek P.: Twitter for Sentiment Analysis: When Language Resources are Not Available. In: *Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop on*, pp. 111–115, 2011, ISSN 1529-4188.

[21] Pang B., Lee L., Pang B., Lee L.: Opinion Mining and Sentiment Analysis. *Foun-dations and Trends in Information Retrieval*, vol. 2(1–2), pp. 1–135, 2008, ISSN 1554-0669, `http://dx.doi.org/10.1561/1500000011`.

[22] Pang B., Lee L., Vaithyanathan S.: Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. In: *Proceedings of the ACL-02 Conference on Em-pirical Methods in Natural Language Processing*, *EMNLP '02*, vol. 10, pp. 79–86, Association for Computational Linguistics, Stroudsburg, PA, USA, 2002, `http://dx.doi.org/10.3115/1118693.1118704`.

[23] Peng T.C., Shih C.C.: An Unsupervised Snippet-Based Sentiment Classification Method for Chinese Unknown Phrases without Using Reference Word Pairs. In: *Web Intelligence/IAT Workshops*, pp. 243–248, IEEE, 2010.

[24] Ramos J.: Using tf-idf to determine word relevance in document queries. In: *Proceedings of the first instructional conference on machine learning*, 2003.

[25] Rennie J.D., Shih L., Teevan J., Karger D.R.: Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In: *ICML*, pp. 616–623, AAAI Press, 2003.

[26] Tromp E., Pechenizkiy M.: SentiCorr: Multilingual Sentiment Analysis of Per-sonal Correspondence. In: *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 1247–1250, IEEE Press, 2011, ISSN 2375-9232.

[27] Turetken O., Olfman L.: Introduction to the Special Issue on Human-Computer Interaction in the Web 2.0 Era. In: *AIS Transactions on Human-Computer Inte-raction*, pp. 1–5, 2013.

[28] Turney P.D.: Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. In: *Proceedings of the 40th Annu-al Meeting on Association for Computational Linguistics*, ACL '02, pp. 417–424, Association for Computational Linguistics, Stroudsburg, PA, USA, 2002, `http://dx.doi.org/10.3115/1073083.1073153`.

[29] Vinodhini G., Chandrasekaran R.M.: Sentiment Analysis and Opinion Mining: A Survey. *International Journal*, vol. 2(6), 2012.

[30] X. Hu H.L.: Text analytics in social media. In: C.C. Aggarwal, C. Zhai, eds., *Mining Text Data*, pp. 385–414, Springer, 2012.

[31] Zhang H., Yu Z., Xu M., Shi Y.: An Improved Method to Building a Score Lexicon for Chinese Sentiment Analysis. In: *SKG*, pp. 241–244, IEEE Computer Society, 2012.

Affiliations

Bogdan Gliwa

> AGH University of Science and Technology, Department of Computer Science, Kraków, Poland, `bgliwa@agh.edu.pl`

Anna Zygmunt

> AGH University of Science and Technology, Department of Computer Science, Kraków, Poland, `azygmunt@agh.edu.pl`

Michał Dąbrowski

> AGH University of Science and Technology, Department of Computer Science, Kraków, Poland, `midabrow@gmail.com`