Vol. 50, № 6. – Pp. 551–555. 22. Kiselev M.I. The role of the phase and the phase of the methods in the theory and its applications // Math. Methods and phys. mech. field. – 2003. – Vol. 46, № 3. – Pp. 149–150 (in Russian). 23. Surdutovich G.I., Vitlina R. Z. and Baranauskas V. Simple reflectometric method for measement of weakly absorbing films// Thin Solid Films. – 1999. – Vol. 355–356.– Pp. 446 –450. 24. Rgevkin S.N. Lectures course on the theory of sound. – M: Moscow State University, 1960. – 335 p. 25. Podilchuk U. N., Rubczov U.K. Ray methods in the theory of propagation and scattering of waves.. – Kyiv: Naykova dumka, 1988. – 220 p. 26.Kosoboutskyy P.S., Karkulovska M.S., Kosoboutskyy J.P. On phase-amplitude correlation in reflection spectra of Fabry-Perrot interferometers // Optica and Spectroscopy. – 2003. – Vol. 94, № 3. – Pp. 434–436. 27. Dwigt H.B. Tables of Integrals and other Matematical Data. 4nd ed. New York, The Mac Millan Company.1961.

UDK 811.161.2

M. Lobur, A. Romaniuk, M. Romanyshyn
Lviv Polytechnic National University
Computer-Aided Design Department

DEFINING AN APPROACH FOR DEEP SENTIMENT ANALYSIS OF REVIEWS IN UKRAINIAN

This paper studies the approaches commonly used for sentiment analysis and defines an optimal approach for Ukrainian language analysis.

Key words: rule-based sentiment analysis, deep sentiment analysis, sentiment dictionaries.

Описано підходи до емоційно-смислового аналізу, а також визначення найкращого підходу для української мови.

Ключові слова: емоційно-смисловий аналіз на основі правил, глибокий емоційно-смисловий аналіз, тональні словники.

1. Problem

Sentiment analysis is the task of natural language processing, which is widely used nowadays in such areas as sociology (e.g. collecting data from social networks about people's likes and dislikes), political science (e.g. collecting data about political views of certain social groups), marketing (e.g. creating ratings of products/companies/people), medicine and psychology (e.g. detecting signs of psychological illnesses or signs of depression in users' messages, detecting bullies with the help of messages in microblogs, like Twitter), etc. [1].

Unfortunately, no matter how useful such a tool would be, there is no available sentiment analysis system for Ukrainian language yet. The aim of this paper is to study the most effective approaches to sentiment analysis and thus find the optimal approach for implementing such an analyser for Ukrainian. The approaches researched in this paper include a rule-based approach, statistical analysis based on sentiment dictionaries and approaches based on machine learning algorithms.

2. Recent Research Analysis

The previous decade showed a rising interest in the area of sentiment analysis. This has been proven with a large number of projects, which appear every day: sentiment analysis of hotel reviews [9], bank reviews [5], restaurant reviews, comments on movies [21], products, messages about political events in blogs and social networks, etc. A big number of studies are dedicated to sentiment analysis of messages in microblogs (e.g. Twitter, Google Buzz).

This paper observes recent studies in the field of sentiment analysis, which use different approaches. These works are used for comparison here. A very interesting research was conducted by W. Kasper and M. Vela from DFKI GmbH [8, 9]. This work has been presented this year. The developers managed to combine statistical and rule-based approaches to implement sentiment analysis for hotel reviews in German. Another important research in this area was conducted by K. Moilanen and S. Pulman from the University of Oxford [12], where compositional semantics was used. The results of their work were presented in 2009. The third study that drew our attention is sentiment analysis of Russian messages, implemented at ЗАО «Ай-Теко», Moscow in 2011. This study used a detailed sentiment dictionary for analysis of reviews in Russian [2]. Another research for Russian, implemented by D. Kan in 2011, used a rule-based approach for analyzing emotions in messages [7].These and some other studies will be mentioned further in the article.

3. Research Aims

The aim of this paper is to define an approach to sentiment analysis of reviews in Ukrainian.

The objectives of the article are the following:

- analysis of text preprocessing needs and tools;
- analysis and comparison of existing approaches to sentiment analysis;
- defining an optimal approach suitable for analysing reviews in Ukrainian.

4. Main Part

4.1. The Task of Sentiment Analysis

Sentiment analysis, or opinion mining, is a kind of text analysis, which aims to identify emotional attitudes or subjective judgments of the author concerning a particular object in the text message. The main objective of sentiment analysis is the automatic evaluation of a particular object (a person, a message in media, an event, an organization, etc.) in a text message in order to get a numerical or categorical indicator of general subjective attitude to the object. The aim of sentiment analysis is to understand the opinions and preferences of users, customers or clients. It is used by managers of various organizations for learning more about the advantages and disadvantages of their products, services or about the organization itself.

This paper discusses the existing approaches to the implementation of sentiment analysis. The approach defines preprocessing needs and supplementary tools for sentiment analysis, such as dictionaries, corpora, ontologies. The accuracy of the outcome of the sentiment analysis system depends largely on the choice of the approach, too.

4.2. Text Preprocessing

Before getting directly to analyzing emotions in a text fragment, some studies suggest that the fragment itself should be preprocessed. Usually words go through stemming and lemmatization. Stemming means removing word endings, and lemmatization involves transforming each word to its initiate form. The aim of these procedures is to unify all word forms of one word, which will reduce the number of words needed to process. However, not all of the researchers think it is necessary to stem or lemmatize words, as in that way we lose important morphological information that could be necessary for sentiment analysis, e.g. words ‘love’ and ‘loved’ may convey different emotions. While the first word would be positive, the second one could convey more negative attitude, like regret.

The approach that is going to be used also defines whether or not we need part of speech tagging or parsing to be done. These do not necessarily have to be deep; they can be shallow, depending on the depth of the needed analysis. For example, the research on sentiment analysis of Russian reviews [2] involved tagging nouns, verbs, adjectives and adverbs only. The same research used shallow parsing for defining an object, a predicate and a subject of each clause. The text preprocessing stage in the research on sentiment analysis of hotel reviews [9] included stemming, part of speech tagging and an n-gram-based spellchecker. Part of speech tagging and parsing were also used in [12] for English and in [16] for sentiment analysis for seven languages.

Most studies also define the author and the object of attitude. The author is usually easy to define as it is usually stated in the annotation to the review. The object of attitude is, though, more difficult to find. For that purpose researchers use named-entity recognition, study the role of the object in the sentence, punctuation, etc. [2]. A similar approach was used in the research on German reviews, except they used an entity extraction system SProUT [9]. Some studies just use noun clusters as objects [16, 17].

4.3. Approaches to Sentiment Analysis

Having generalised recent research in the field of sentiment analysis, we got an ability to define the following approaches:

1. statistical approach based on sentiment dictionaries;
2. rule-based approach;
3. supervised machine learning;
4. unsupervised machine learning.

4.3.1. Statistical approach based on sentiment dictionaries

The first approach uses so-called sentiment dictionaries. A sentiment dictionary is a list of words with their sentiment values. A sentiment value can be a number (e.g., 1-10, where 1 is a negative word, and 10 is a positive word) or a certain category (e.g., positive or negative).

According to this approach every word in a review is assigned a sentiment value stated in a dictionary, and after that the sentiment of the whole review is computed. The general sentiment is computed statistically. This does not usually give high accuracy, which is the first disadvantage of this approach. The next disadvantage is no space for conducting deep analysis of the text message. However, this kind of statistical approach needs neither part of speech tagging nor parsing to be conducted, which is of high importance for languages, which lack such tools of text processing.

Very often only nouns, verbs, adjectives and adverbs are listed in a sentiment dictionary. For example, the research of sentiment analysis of Russian texts [2] used a sentiment dictionary, which contained only the most frequent nouns, verbs, adjectives and adverbs, collected from mass media articles. Every word was assigned its part of speech and strength of the sentiment (from 1 to 3). Adjectives and adverbs are divided into positive, negative and amplifying. Nouns are divided into positive, negative, potentially positive and potentially negative (these are the words, whose sentiment relies on the surrounding words; they are positive in positive surrounding, and negative in negative surrounding). Verbs fall into eight categories, depending on their surrounding and the role they play in a sentence. Linking verbs comprise a separate category.

All the words in a sentiment dictionary usually refer to a specific domain (e.g. banks, restaurants, movies, etc.) as it is much more difficult to implement sentiment analysis for general domain. As manual creation of such a dictionary is an extremely time-consuming task, there have been developed methods of automatic generation of sentiment dictionaries on the basis of sentiment-annotated corpora and ontologies that refer to a specific domain. If the domain of the dictionary is not defined, general-domain semantic networks are used, too. For example, the work [12] used a sentiment dictionary, created on the basis of WordNet 2.1. Among the available sentiment dictionaries, the most commonly used ones will have to be SentiWordNet and General Inquirer Lexicon. SentiWordNet embraces seven languages and is frequently used in the field of sentiment analysis [6]. Unfortunately, it does not support Ukrainian language yet. General Inquirer Lexicon can be used only for English and appears to be a useful sentiment database, too [3].

To conclude with, the statistical approach to sentiment analysis that is based on sentiment dictionaries is rather easy to implement, as it needs only a sentiment dictionary and an algorithm of computing the average sentiment value of a text fragment. Also this approach does not require part of speech tagging or parsing. The downside of this approach is the low quality of the results and no possibility to conduct deeper text analysis (e.g. define the reason of attitude, define the range of emotions the author wanted to voice, distinguish text fragments with a particular sentiment or emotion, etc.).

4.3.2. A rule-based approach to sentiment analysis

This type of systems is based on a set of rules, which the system uses in order to define the sentiment of a text fragment.

The majority of commercial systems use this approach, despite the fact that it is extremely time-consuming, since the system requires a large number of manually-written rules for effective operation. As well as the dictionaries, almost always the rules refer to a particular domain (such as hotels, restaurants, banks, movies, music, etc.), which complicates their being used for analysis of the texts of other topics. Nevertheless, a good base of rules makes this approach work quite accurately within a certain domain. An obligatory tool for implementing this approach is a sentiment dictionary. It is used to assign a sentiment value to every word in a sentence (if such a value exists), and this information will be used later to actually define the sentiment of the sentence. One of the simplest rule-based algorithms for calculating the sentiment of the sentence was presented in work [7], which describes the processing of Russian text messages about books, movies, and digital cameras. In this study, sentiment analysis is comprised of the following steps:

- first of all, the words-invertors are searched for in every sentence (if such a word is found, the sentiment of the following three words is changed to the opposite);
- the number of positive and negative words is counted separately;
- an opposite clause is searched for, and, if found, the number of subjective words in the sentence is divided by two;
- finally, the general sentiment is calculated: the number of negative words and the result of the previous stage are subtracted from the number of positive words. If the general sentiment is greater than zero, the text fragment is positive; if it is less than zero, the text fragment is considered to be negative; in case of a zero outcome, the text is neutral.

As you can see, this example also does not require any specific text preprocessing tools, but the results of such an analyzer will be better than that of the previous approach, since in this case the words-invertors and opposite clauses are taken into account.

In study [3] a similar approach was used, but here the sentiments of five words after the word-invertor are changed to the opposite. Furthermore, the hidden English syntactic inverting structures are taken into account. This study also emphasizes the importance of subjective adjectives, which are pre-defined with the help of part of speech tagging. The sentiment of adjectives is determined on the basis of sentiment dictionaries, but if a particular adjective is not in the dictionary, the adjective's relations with other words are searched for in WordNet. Using such a simple set of rules, the researchers managed to achieve 81 % accuracy when applying this approach to a certain domain, but only 68 % for texts on general topics.

An example of a deeper and therefore more precise study of the text subjectivity using compositional semantics was presented in [12], a work on English text messages. Every sentence here is run through part of speech tagging and parsing, and each word is assigned a sentiment value according to a sentiment dictionary. Starting from the main verb in the predicate, the words in the sentence are combined. The words are added one by one, moving along the syntactic tree, and the sentiment value of the text fragment is immediately computed, depending on the dominant word in that part of a sentence. This algorithm allows conducting deep analysis that involves distinguishing subjective text fragments that reflect the author's opinion. Also the important thing is the ability to determine the reason of the author's subjective attitude towards a particular object.

The work on sentiment analysis of German hotel reviews [9] the rule-based approach was used together with computing the sentiment statistically. Thus, the system calculates and overall sentiment of the message and some emotional peculiarities of each sentence's sentiment. The same approach is seen in another commercial sentiment analysis system, presented in [15].

To sum up, the rule-based approach to sentiment analysis is really time-consuming, as it requires a substantial number of manually written rules for the effective work of the analyzer, but when being applied within a particular domain, such an approach can provide more than 90 % accuracy. Text preprocessing proved to be very helpful for this approach, and it also gives the developers more opportunities for research.

4.3.3. A supervised machine learning approach to sentiment analysis

This approach has become popular during the past few years. The implementation of sentiment analysis using machine learning algorithms involves training a machine learning classifier on a sentiment-annotated corpus, and then using the resulting model to analyze new texts.

Although a sentiment-annotated corpus was not obligatory for the previous approaches in case there was an available sentiment dictionary, the approach based on supervised machine learning requires such a corpus. Moreover, the accuracy of the analyzer depends greatly on the quality and magnitude of the sentiment-annotated corpus.

Supervised machine learning allows us to apply classification and linear regression algorithms to define the sentiment of a text fragment.

The classification is used when there is a finite set of subjective classes. Classification can be flat or hierarchical. Flat classification means training one classifier for differentiating all classes. Hierarchical classification means that classes are divided into groups, and classifiers are trained to differentiate between the groups. For example, if we use five classes: very positive, positive, neutral, negative and very negative messages, we could consistently apply three classifiers. Firstly, we could teach a binary classifier, which will differentiate between neutral and subjective texts. The second classifier will define positive and negative texts, and, finally, the third classifier's work will be to differentiate between positive and very positive message, as well as between negative and very negative.

Linear regression is used to obtain numerical values of subjectivity, for example, from 1 to 10, where, for example, 1 is a negative message, and 10 is a positive one. In some studies, these two algorithms are used in parallel for comparison.

The process of implementation of sentiment analysis with the use of supervised machine learning can be divided into the following stages:

1. annotating a corpus for the classifier;
2. representation of each annotated review in the form of a feature vector (features can be presented by words, n-grams, surrounding words, or even punctuation; the words may or may not be transformed to their initial form);
3. choose a classification algorithm (Bayesian classifier, SVM, MaxEnt classifier, etc.) and train a classifier.

Linear regression was used in project [11], where also ToneADay.com website was taken advantage of in order to attract new languages. An example of a machine learning classifier can be found in the project [14], where the classifier is trained on a sentiment-annotated corpus of movie reviews from NLTK. The classification algorithm was used to distinguish neutral and subjective text messages. If the text is neutral, the sentiment is not defined.

Thus, supervised machine learning algorithms are an effective tool for sentiment analysis, but only if they are trained on a sentiment-annotated corpus of a substantial size (500 thousand words and more). Nevertheless, it does not allow for detailed analysis of the parts of the text. For example, distinguishing separate text fragments that express certain emotions would involve using a combined approach (e.g., machine learning together with rule-based approach, as it was done in [10; 13]).

4.3.4. An unsupervised machine learning approach to sentiment analysis

Unsupervised machine learning hasn't proven itself to be very effective for sentiment analysis yet.

To implement this approach a corpus is also needed, but it doesn't need to be sentiment-annotated. The task of the system is to find phrases, which seem to be subjective. Having a part of speech tagging done, the system looks for adjectival and adverbial phrases and tries to put them into categories. Using sentiment dictionaries or rules, these phrases are assigned to certain sentiments. These data allow the system to define the general sentiment direction of the text to be positive, negative or neutral.

A detailed example of using unsupervised machine learning for sentiment analysis was described by Peter Turney in [18]. P. Turney used a specific method of unsupervised machine learning, which was based on the information retrieval of the most frequently used words and phrases in a text.

4.4. Deep Sentiment Analysis of Ukrainian Reviews

The aim of our study is the implementation of deep sentiment analysis of reviews in Ukrainian. Having conducted a detailed research of existing approaches to sentiment analysis, we found that the most suitable approach for deep sentiment analysis will be a rule-based approach. This approach makes it possible to define subjective text fragments that reflect the opinion of the author.

The implementation of deep sentiment analysis for Ukrainian language can be divided into the following steps:

1. defining text preprocessing tools;
2. creating a sentiment dictionary;
3. constructing rules;
4. creating a result representation tool.

The text preprocessing stage for the chosen approach involves shallow part of speech tagging and parsing. A part of speech tagger for Ukrainian language was implemented in the project UGTag [19], and the tagger itself is free to download and use. Although this analyzer lacks morphological disambiguation, we still can use it, as we need only shallow tagging for our research, and the information about the part of speech will be enough. Luckily, such information is usually unambiguous. Unfortunately, there is no available parser for Ukrainian language, and the implementation of such a tool goes beyond this study. Thus, it was decided to write our own part-of-speech patterns for distinguishing a subject, a predicate and an object of each sentence, which would play the role of shallow parsing in our research.

Stemming and lemmatization will not be needed as the abovementioned part of speech tagger already provides information about the initiative form of the word. This initiative form will be useful, when the desired word is not in the dictionary. Then the word will be assigned the sentiment value of its initiative form. Also, according to earlier research, different forms of one word can have different sentiment values. Thus, using the sentiment value of the initiative form of the word may lower the accuracy of the analyser, however, it will, at the same time, reduce the computational time.

At the stage of preprocessing the object of subjectivity will also be defined. Since we have chosen restaurant reviews as the domain for our study, the objects of subjectivity will be the named entities – the names of the restaurants. The main methods to find these objects are the form of words (capital letters, numbers usage, unusual combination of nouns, etc.), punctuation (such as quotes), and surrounding words (using collected examples from the corpus).

It is also worth mentioning that the sentiment of each clause will be defined separately, as one complex sentence may contain text fragments with opposite sentiments. Complex sentences are going to be divided into clauses with the help of punctuation and conjunctions.

Getting to the next step, we need a sentiment dictionary. Unfortunately, there is no available sentiment dictionary for Ukrainian language neither for the restaurant domain, nor for the general domain, which defines the problem of creating such a sentiment dictionary. Sentiment dictionaries are usually generated on the basis of ontologies or sentiment-annotated corpora. In order to create a sentiment dictionary, we created a sentiment-annotated corpus of restaurant reviews in Ukrainian (600 annotated reviews). On the basis of this corpus we managed to generate the main part of the sentiment dictionary for restaurant domain. It's worth mentioning that the dictionary contains only nouns, verbs, adjectives and adverbs. Each word is assigned a sentiment value (positive or negative) and emotion, if present (anger, disgust, fear, joy, sadness, surprise, which are the basic emotions by P. Ekman). Words that play the role of invertors («не», «нема», «немає», «неможливо», etc.), and words that play the role of amplifiers («дуже», «надзвичайно», «безмежно», etc.) are added to the dictionary, too.

The next step is writing the rules. The first part of the rules has to do with the words-invertors. When such a word is found in the sentence, the sentiment of the next word or set of words (up to five homogeneous words) is changed to the opposite. Then the sentiment of the clause is defined. This is done with the help of sequential composing of sentiments, which was used in the work [12] for sentiment analysis of English. The words are added one by one, starting from the predicate, and the sentiment is defined, depending on the main word in the text fragment. This approach gives more accurate results than counting positive and negative words. In the end, the amplifying words are processed: if such a word is

found, the sentiment of a positive clause is changed to very positive and the sentiment of a negative clause is changed to very negative.

The results are going to be presented in a table with positive, negative, very positive and very negative clauses from every review. If the clause possesses any specific emotion of the six basic emotions by Ekman, it is stated too.

Conclusion

This paper reviewed the most commonly used approaches to the implementation of sentiment analysis, which use sentiment dictionaries, rules and supervised and unsupervised machine learning algorithms.

Having conducted a detailed analysis of recent research in the field of sentiment analysis, we have designed an algorithm for implementing deep sentiment analysis of reviews in Ukrainian for the restaurant domain, which involved a rule-based approach. A sentiment-annotated corpus and a sentiment dictionary for the restaurant domain have been created. The rules of sequential sentiment compounding are being developed.

1. Давыдов А. А. Системная социология: Opinion Mining / А. А. Давыдов. – М.: ИС РАН. – 2009. – Режим доступу: http://www.isras.ru/index.php?page_id=1024/ 2. Пазельская А. Метод определения эмоций в текстах на русском языке / А. Г. Пазельская, А. Н. Соловьев // Компьютерная лингвистика и интеллектуальные технологии: сб. научных статей / Вып. 10 (17). – М.: Изд-во РГГУ, 2011. – С. 510–522. 3.Agrin N. Developing a Flexible Sentiment Analysis Technique for Multiple Domains / Nate Agrin. – 2006. – Режим доступу: http://courses.ischool.berkeley.edu/i256/f06/projects/agrin.pdf 3) Cheng T. Corpus and Sentiment Analysis / Tai Wai David Cheng. – Guildford, 2007. – 144 p. 4.Deep sentiment analysis with attensity analyze optimises Lloyds' customer service. – Режим доступу: http://www.attensity.com/wp-content/uploads/2010/09/LloydsSuccessStory.pdf 5) Denecke K. Using SentiWordNet for Multilingual Sentiment Analysis / Kerstin Denecke. – ICDE Workshops. – 2008. – pp. 507-512. 6. Kan D. Rule-based approach to sentiment analysis at ROMIP 2011 / Dmitry Kan. – Режим доступу: http://www.slideshare.net/dmitrykan/rule-based-approach-to-sentiment-analysis-at-romip-2011 7. Kasper W. Monitoring and Summarization of Hotel Reviews / Walter Kasper, Mihaela Vela. – Information and Communication Technologies in Tourism 2012 (ENTER-2012). – Helsingborg, Wien: Springer, 01/2012. – pp. 471-482. 8. Kasper W. Sentiment Analysis for Hotel Reviews / Walter Kasper, Mihaela Vela. – Proceedings of the Computational Linguistics-Applications Conference. – Jachranka, Poland: Polskie Towarzystwo Informatyczne, Katowice, 10/2011. – pp. 45–52. 9. Kawathekar S. A. Sentiments analysis using Hybrid Approach involving Rule-Based & Support Vector Machines methods / Swati A. Kawathekar, Dr. Manali M. Kshirsagar. – IOSR Journal of Engineering (IOSRJEN). – Vol. 2 Issue 1. – Jan., 2012. – pp. 55-58. 10. Lybmix. – Режим доступу: http://www.lybmix.com/live-demo 11) Moilanen K. Multi-entity Sentiment Scoring / Karo Moilanen, Stephen Pulman. – Proceedings of Recent Advances in Natural Language Processing (RANLP 2009). – Borovets, Bulgaria, September 14–16 2009. – pp. 258–263. 12. Prabovo R. Sentiment Analysis: A Combined Approach / Rudy Prabowo, Mike Thelwall. – Journal of Informetrics. – Vol. 3, No. 2. – April 2009. – pp. 143–157. 13. Python NLTK Demos for Natural Language Text Processing. – Режим доступу: http://text-processing.com/demo/sentiment 14. SAS® Sentiment Analysis. – Режим доступу: https://www.sas.com/text-analytics/sentiment-analysis/#section=1 15. Skyttle API. – Режим доступу: http://nlp.skyttle.com/api/nlp/ 16. Trustyou. – Режим доступу: http://www.semantic-api.com/demo-statistical-sentiment-analysis.html 17. Turney P. D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews / P. D. Turney // Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (Stroudsburg, PA, USA 2002): ACL, 2002. – pp. 417–424. 18. UGTag – a morphological tagger for Ukrainian language. – Режим доступу: http://www.domeczek.pl/~polukr/parcor/ 19. Yang C. A Rule-Based Approach for Effective Sentiment Analysis / Chin-Sheng Yang, Hsiao-Ping Shih. – Режим доступу: http://pacis2012.org/files/papers/pacis2012_T25_Yang_288.pdf 20. Yessenov K. Sentiment Analysis of Movie Review Comments / Kuat Yessenov, Sasa Misailovic. – Massachusetts Institute of Technology, Spring 2009. – Режим доступу: http://people.csail.mit.edu/kuat/courses/6.863/report.pdf