RULE-BASED SENTIMENT ANALYSIS OF UKRAINIAN REVIEWS

Mariana Romanyshyn

Department of Computer-Aided Design, Lviv Polytechnic National University, Lviv, Ukraine
mariana.scorp@gmail.com

ABSTRACT

Last decade witnessed a lot of research in the field of sentiment analysis. Understanding the attitude and the emotions that people express in written text proved to be really important and helpful in sociology, political science, psychology, market research, and, of course, artificial intelligence. This paper demonstrates a rule-based approach to clause-level sentiment analysis of reviews in Ukrainian. The general architecture of the implemented sentiment analysis system is presented, the current stage of research is described and further work is explained. The main emphasis is made on the design of rules for computing sentiments.

KEYWORDS

Sentiment analysis, rule-based approach, sentiment dictionary, sentiment grammar

1. INTRODUCTION

Sentiment analysis, or opinion mining, is the field of natural language processing, the main task of which is to identify subjective or emotional indicators in the text, i.e. to find the attitude of the author towards a certain object or event. It 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, tracking the change of attitude towards certain governmental issues), marketing (e.g. creating ratings of products/companies/people), psychology (e.g. detecting signs of psychological illnesses or signs of depression in users' messages), and, of course, artificial intelligence, for which the understanding of human feelings and emotions is really important.

There are different kinds of sentiment analysis tools. First of all, there are sentiment maps, like [1] or [2]. These maps represent the general mood of some social group at the certain point of time. The maps are usually composed, relying on positive/negative posts in microblogs.

Another popular kind of sentiment analysis is entity-centric analysis, like [3] or [4]. This kind of analysis represents the rating of a certain organization, person, product or event at a certain point of time. The information is usually collected from blogs, microblogs and websites which contain reviews.

Review-based sentiment analysis has become a valuable substitution of or a supplement to the star rating at review websites. The examples of review-based sentiment analysis can be found in [5] and [6].

DOI : 10.5121/ijaia.2013.4410

Deep sentiment analysis involves the identification of sentiment or emotional information at the level of a sentence or a clause. This kind of sentiment analysis is especially difficult to implement, but it can be the most valuable for the task of understanding human emotions and attitudes. A review can include both positive and negative comments about a certain entity, and deep sentiment analysis gives a possibility to get not just the general sentiment evaluation of the review, but positive or negative clauses, which is more valuable, as it gives the ability to analyse the main positive or negative points about the entity. This paper deals with deep sentiment analysis for Ukrainian reviews.

2. RELATED WORK

The majority of works on sentiment analysis is dedicated to review-level or entity-centric sentiment analysis, described in the introduction. Sentiment analysis on the level of sentence or clause, or deep sentiment analysis, is more poorly researched, but has been implemented for English language with the mix of rule-based approach and machine learning techniques by Jason Kessler [7, 8] and Theresa Wilson [9, 10]. One of the reasons, why sentiment analysis on the level of clause is less researched, is the necessity of additional NLP tools to compute and aggregate sentiment information. Machine learning proved to be indeed effective for other types of sentiment analysis, which can be seen from the works by Eric Breck, Yejin Choi & Claire Cardie [11], Tony Mullen and Nigel Collier [12], Bo Pang, Lillian Lee & Shivakumar Vaithyanathan [13] and Sarah Schrauwen [14]. Still, we believe that, however time-consuming, the rule-based approach is capable of giving a better understanding of the author's attitude, is more suitable for deep sentiment analysis and is more efficient when big number of annotated corpora is unavailable. This work has been inspired, first of all, by rule-based sentiment analysis systems described in [15], implemented for English, and [16], implemented for Russian. A more detailed comparison of approaches to sentence-level sentiment analysis can also be found in [17].

Implementing sentiment analysis for Ukrainian language seems to be a challenging task. There are no sentiment-annotated corpora for Ukrainian language yet, and the syntactic parser is still under development. In this paper we propose our attempt to create a clause-level rule-based sentiment analysis system for reviews in Ukrainian. Section 3 describes the general architecture of our system. Sections 4 and 5 are dedicated to the description of the rules of sentiment grammar.

3. GENERAL DESCRIPTION OF THE PROPOSED APPROACH

Rule-based approach to sentiment analysis allows deep analysis of the opinion content of the review, meaning you can not only find the general sentiment of the review, but also a separate sentiment of each clause. This provides enough information to single out separate positive and negative points about an entity or event. Thus, such analysis gives more information than the general sentiment of the review or the rate of certain organization or person in social networks. Also the rule-based approach makes it possible to later compute the sentiment of the sentences with relative clauses and get more precise sentiment information.

Fig. 1 presents the general scheme of sentiment analysis system that we implemented for Ukrainian reviews.

As you can see from Figure 1, the first stage is text pre-processing. In every review the date, the author, the citation of the previous message (if present) and the text of the review itself are defined. Then the extracted text of the review is part-of-speech tagged and split into clauses. POS tagging is conducted with the help of UGTag 2.0 – a morphological tagger for Ukrainian language [18], which proved to be a vital tool for our system. Clauses are defined with the help of punctuation, which is really strict in Ukrainian, count of potential predicates, coordinate conjunctions and relative pronouns. Here we rely on the assumption that the rules of punctuation have been followed by the author of the review. Named-entity recognition has been implemented at two stages: before and after POS tagging. Entities that have distinct features (capital letters, quotation marks, non-Ukrainian characters, etc.) are defined before POS tagging with the help of regular expressions. The ones that are harder to define are recognized after POS tagging with the help of obtained morphological information and right and left context of the word or phrase.

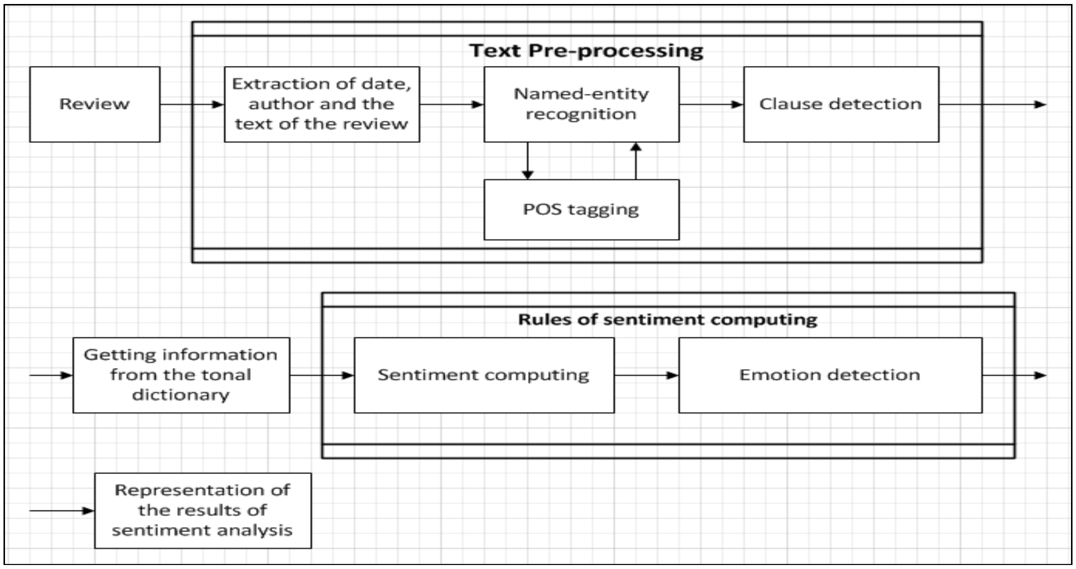
Text Pre-processing
Review
Extraction of date, author and the text of the review
Named-entity recognition
Clause detection
POS tagging
Rules of sentiment computing
Getting information from the tonal dictionary
Sentiment computing
Emotion detection
Representation of the results of sentiment analysis

Figure 1. The general scheme of sentiment analysis system

The next stage of our algorithm is assigning sentiment and emotion to each word of the clause. This is done with the help of a sentiment dictionary, generated from a manually annotated sentiment corpus. The corpus has been collected from the reviews on the popular Ukrainian review websites [19] and [20] and contains almost 800 reviews now. Each dictionary entry contains the lemma of the word, its morphological tag, sentiment and emotion. The sentiment of a word can be positive, negative, invertor or intensifier. We decided to start from this small collection of sentiment categories, but we feel that it should be extended in the future. The emotion type for each word can be anger, disgust, fear, joy, sadness, surprise, as basic human emotions, according to the psychologist Paul Ekman [21]. If a word cannot be referenced to a definite sentiment or emotion category, it is assigned a ‘neutral’ value for sentiment and ‘none’ value for emotion.

The identification of emotions as a supplement to sentiments is something quite novel in the field of sentiment analysis. Still, emotions convey a different kind of information than sentiments. The most prominent example would be the emotion of surprise, which can be either positive or negative:

I was very much surprised with their fantastic service.
vs
Their impoliteness was a surprise.

On the other side, there can be positive or negative clauses that do not convey any specific emotion from P. Ekman’s range:

The food was quite good.

As a result, we get a collection of clauses, for which a sentiment will be defined. Each clause represents a collection of words with their attributes, namely each word's lemma, morphological tag, sentiment and emotion.

The next stage is the computing of sentiments for each clause.

4. LEVELS OF SENTIMENT GRAMMAR

In order to define the sentiment of a clause properly, the sentiments of separate words have to be composed in a certain order. The best choice would be to use the syntactic parser, but, unfortunately, Ukrainian language lacks such kind of a tool yet. That is why we decided to define the levels of sentiment composition, which will somehow reflect the syntactic tree. We defined the rules of sentiment composition at the phrasal level and at the level of a sentence. Each rule unites the given words/phrases, computes their composite sentiment and saves the morphological information of the head word.

The order of rules for word/phrase unification and their sentiments computing was defined by the peculiarities of Ukrainian language. At the phrase level the general order of sentiment computing rules is the following:

- Adverbs unification.
  First we unify adverbs with adverbs, with or without a coordinate conjunction between them.
- Verbs unification.
  Ukrainian language possesses few composite verb forms, but they have to be unified after adverbs, as an adverbial modifier can be placed inside of the composite verb form.
- Adverbs and verbs unification.
  As we know, adverbs can modify different parts of speech. Still, on the first place we process the verb and adverb collocations. At this stage we also unify verbs with 'не' (Eng.: not), which is a particle in Ukrainian, and with 'б' or 'би', which are particles that define the conditional mood of a Ukrainian verb. The inverting particle 'не' changes the sentiment of the verb to the opposite and the conditional mood of the verb is a strong neutralizing indicator.
- Adjectives unification.
  The next step is to count the sentiment of a sequence of adjectives, with or without a coordinate conjunction between them.
- Adverbs and adjectives unification.
  Here the adverbs, which modify adjectives, are unified with adjectives. At this stage we also unify adjectives with 'не' (Eng.: not), which is a particle in Ukrainian.
- Nouns unification.
  First we count the common sentiment of nouns that do not have any commas or conjunctions between them. These noun clusters represent the relation of possessive case in Ukrainian language for both animate and inanimate nouns. The main noun goes first, as opposed to English, for example.
- Adjectives/numerals/certain pronouns and nouns unifications.
  By this time adjectives have already been unified with adverbs and nouns have been unified into noun clusters. This gives the possibility to count the sentiment of the whole noun phrase. At this stage we also unify nouns with 'не' (Eng.: not).
- Prepositions and nouns unification.
  Prepositions play an important role, as they can influence the sentiment. This is also taken into account when computing the composite sentiment.
- Nouns and nouns with prepositions unification.
- Noun enumerations unification

At this stage, separate noun phrases, divided by commas and coordinate conjunctions, are unified and their composite sentiment is counted.

When all chunks are ready, sentiment has to be computed at the clause level. Ukrainian language does not have strict rules of subject-predicate-object order. In order to decide, which noun phrase plays the role of a subject, we look at the case of a noun. In Ukrainian the noun has seven cases and only nouns in nominative case can represent a subject in the clause. Morphological information is also vital, when the subject is represented by a numeral or a pronoun.

We also take into account sentences, in which the linking verb 'be' is omitted, which is very common for Ukrainian language. These are usually predicates, which present copulative relations. Such cases can be detected, as either the verb is replaced with a dash, or we find a noun followed by an adjective.

After the main parts of the clause are identified, we check for adverbial modifiers that modify the whole clause and can influence the sentiment of the clause (e.g., luckily, unfortunately, etc.).
The examples, provided in the end of Section 5, show the above described order of word/phrase unification for computing sentiments.

5. TYPES OF RULES

Sentiment analysis is usually viewed as a domain-oriented task. It is really complicated to create a sentiment analysis system, which would perform well for any topic, as certain words and phrases can be positive in one domain, but negative in another. Take, for one, the word 'small'. If you're talking about mobile phones, it most possibly will be positive, but if you are talking about the size of portions in a restaurant, it will almost certainly be negative.

Sentiment analysis tools have been developed for a variety of domains by now. These include news [22], product reviews [23], comments on political issues [24], legal blogs [25], financial news [26], etc. We decided to take restaurant reviews as a domain for research due to the popularity of the topic at all kinds of blogs and forums. Restaurant reviews from the already mentioned review websites [19] and [20] were taken as a material for our sentiment-annotated corpus and became the basis for the sentiment dictionary, generated from these reviews.

Furthermore, even when the domain of sentiment analysis is chosen, there are still some words and phrases that cannot be referred to a certain sentiment and/or emotion category on their own, but in a collocation with other words they convey certain definite sentiment and/or emotion. Thus, our system includes two kinds of rules at each level of sentiment grammar: context-dependent and context-independent.

As we have mentioned above, the categories of separate words are positive, negative, intensifier, invertor and neutral. The final sentiment of the clause, though, can be positive, negative, very positive, very negative or neutral. Some researches use numerical values to have a range of sentiments, but we feel that categories are in no way worse.

5.1. Context-Independent Rules

At each level of our sentiment grammar we realized 10-25 context-independent rules of sentiment composition. Context-independent rules take into account only the sentiment of the words/phrases that are unified. For example, at the level of adverbs and adjectives unification, the following set of rules has been realized:

INTENSIFIER + POSITIVE -> VERY POSITIVE
INTENSIFIER + NEGATIVE -> VERY NEGATIVE
INVERTOR + POSITIVE -> NEGATIVE
INVERTOR + NEGATIVE -> POSITIVE
INTENSIFIER + NEUTRAL -> POSITIVE
INVERTOR + NEUTRAL -> NEGATIVE
POSITIVE + NEUTRAL -> POSITIVE
NEGATIVE + NEUTRAL -> NEGATIVE
POSITIVE + POSITIVE -> POSITIVE
NEGATIVE + NEGATIVE -> NEGATIVE
NEUTRAL + POSITIVE -> POSITIVE
NEUTRAL + NEUTRAL -> NEUTRAL
NEUTRAL + NEGATIVE -> NEGATIVE
POSITIVE + NEGATIVE -> NEGATIVE
NEGATIVE + POSITIVE -> NEGATIVE

The rules at each level are processed one by one, but they are sorted from the most common ones to the least common, in order to optimize the time spent on the search of a matching rule at a certain level of words/phrases unification.

The least common rules here turned out to be POSITIVE + NEGATIVE and NEGATIVE + POSITIVE. Having reviewed the examples of these unifications, we have found examples of irony, so the result of their unification was defined by the test cases.

5.2. Context-Dependent Rules

Context-dependent rules take into account both the lemmas of the words/head words and the sentiments. We would like to provide a few examples of such rules.

At the stage of adjectives/numerals and nouns unification:

(“high”, NEUTRAL) + (“price”, NEUTRAL) -> NEGATIVE
(“high”, NEUTRAL) + (“quality”, NEUTRAL) -> POSITIVE
(“small”, NEUTRAL) + (“portion”, NEUTRAL) -> NEGATIVE
(“small”, NEUTRAL) + (“range”, NEUTRAL) -> NEGATIVE
(“big”, NEUTRAL) + (“portion”, NEUTRAL) -> POSITIVE
(“big”, NEUTRAL) + (“range”, NEUTRAL) -> POSITIVE
(“few”, NEUTRAL) + (“people”, NEUTRAL) -> POSITIVE
(“a lot of”, NEUTRAL) + (“people”, NEUTRAL) -> NEGATIVE
The sentiment information is used in these rules. For example:
(“few”, NEUTRAL) + (“*”, NEGATIVE) -> POSITIVE
(“a lot of”, NEUTRAL) + (“*”, NEGATIVE) -> NEGATIVE
(“few”, NEUTRAL) + (“*”, POSITIVE) -> NEGATIVE
(“a lot of”, NEUTRAL) + (“*”, POSITIVE) -> POSITIVE

Context-dependent rules are strongly bound to the domain, chosen for sentiment analysis. In case the domain is changed, this is the part that has to be adapted to the new topic in the highest measure.

Consider a few simple examples of sentiment computing at a clause level.

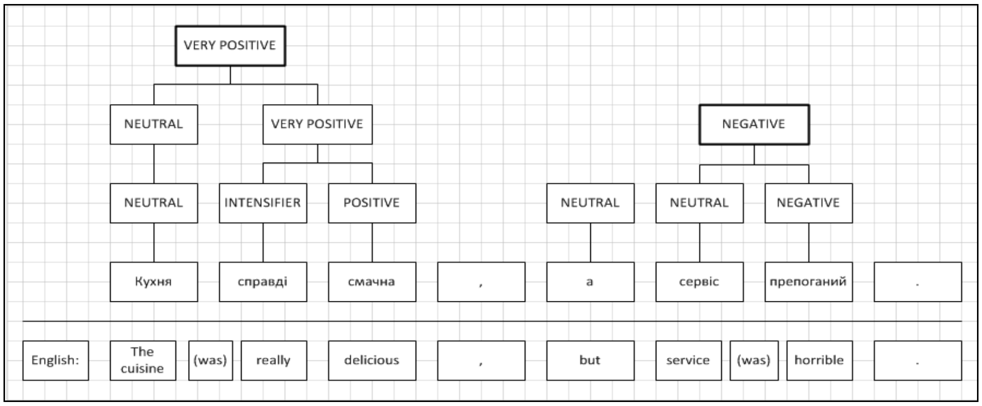

Figure 2. An example of sentiment computing (a)

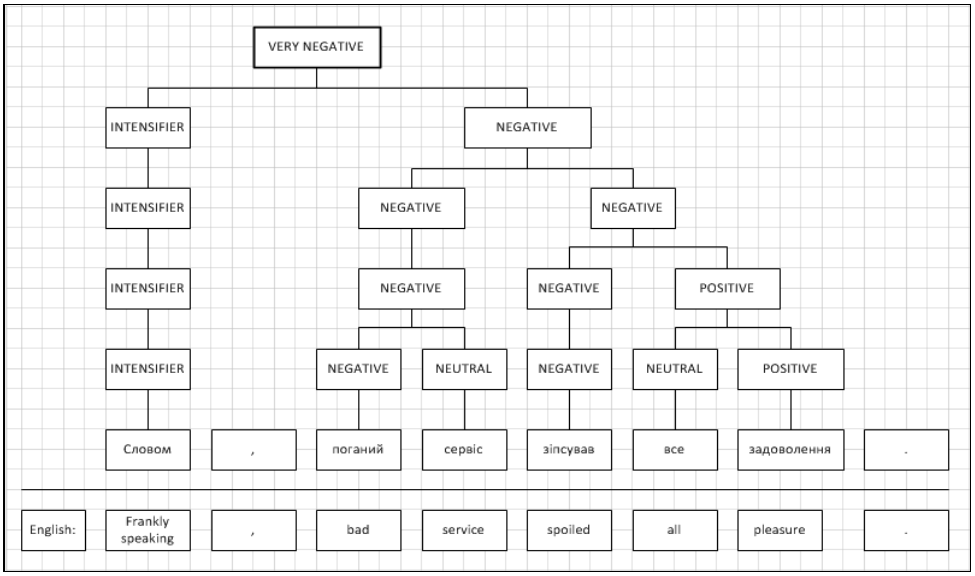

Figure 3. An example of sentiment computing (b)

Figures 2 and 3 show the above described approach to the sentiment computing. The sentence in Figure 2 includes two clauses, which present both a classic case of linking verb omission in Ukrainian language and different sentiments inside of the same sentence. Thus, the positive and negative points about the restaurant can be singled out. The case in Figure 3 shows how the adverbial modifier can influence the sentiment of the clause.

6. EVALUATION AND FUTURE WORK

Although the proposed sentiment analysis system is still under development and needs a lot of improvements, it has already given an average precision of 62% on our test set of 200 reviews. There are currently no available sentiment analysis systems for Ukrainian language, so we cannot conduct any proper comparison. Still, the system is compatible to sentiment analysis applications, implemented for English [17]. It performs better than some simple bag-of-words approaches, but worse than systems implemented with the help of machine learning techniques.

The future work will first of all include the further development of sentiment grammar and defining the best way to detect the emotional content of the clauses, relying on the emotions, conveyed by separate words and phrases. The system is also planned to be enriched with the detection of the sentiment target, meaning the entity, towards which the sentiment/emotion is expressed. This will be necessary when we start working on the discussions on the forums, and not just the reviews, written for a specific restaurant.

The most important tool that we are lacking is a syntactic parser for Ukrainian language, which is currently under development. We believe that the chosen approach makes it easy to include a syntactic parser into the implemented system as soon as such a tool is ready and available. The next step after that will be the definition of anaphoric relations and sentiment analysis at the level of a sentence with relative clauses.

Another important issue for future work is searching for the ways to eliminate errors, related to POS tagger mistakes and spelling and punctuation mistakes made by the author of the review, which again lead to POS tagger mistakes.

7. CONCLUSIONS

The identification of sentiment and emotional content in textual information is a very important issue, and we believe that it has the future in the field of artificial intelligence. Sentiment information is truly significant for natural language understanding, and deep sentiment analysis provides detailed information on positive and negative attitude towards a certain entity.

This paper presented a rule-based approach to clause-level sentiment analysis of Ukrainian restaurant reviews, collected from the popular forums. The steps of the general pipeline have been described, and the manner of sentiment computation has been explained. The levels of sentiment grammar have been enumerated and explained. Two types of rules of sentiment grammar have been developed: context-independent, which take into account only the part of speech and the sentiment, and context-independent rules, which also consider the lemmas of the words. The rules for computing emotional content are under development.

The developed system still has a lot of room for improvement, so the future work has been also discussed. The main improvement for the system will be the syntactic parser.

ACKNOWLEDGEMENTS

The author would like to thank Natalia Kotsyba, Igor Shevchenko and Andriy Mykuliak, who developed UGTag 2.0, which is a vital tool not only for this research, but also for the development of NLP in Ukraine.

REFERENCES

[1] Priyesh Patel, and Daniel Saul, “MoodMap - Correlating Sentiment Data from Tweets with Deprivation Data from the Government,” (MoodMap), [online], http://themoodmap.co.uk/ (Accessed: 15 July 2013).

[2] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist, “Pulse of the Nation: U.S. Mood Throughout the Day inferred from Twitter,” (Northeastern University), [online] 2010, http://www.ccs.neu.edu/home/amislove/twittermood/ (Accessed: 15 July 2013).

[3] Semantic Engines LLC, “Opinion Crawl - sentiment analysis tool for the Web and social media,” (Opinion Crawl), [online] 2010, http://opinioncrawl.com/ (Accessed: 15 July 2013).

[4] GROUBAL, “Business Social Media Scores | Groubal Community Sentiment Index,” (Groubal) [online], http://www.groubalcsi.com/ (Accessed: 15 July 2013).

[5] Lybmix Inc., “Lymbix. Sentiment Analysis Reinvented,” (Lymbix), [online] 2009, http://www.lymbix.com (Accessed: 15 July 2013).

[6] Market Sentinel Ltd, “Skyttle API,” (Skyttle), [online] 2010, http://nlp.skyttle.com (Accessed: 15 July 2013).

[7] Jason S. Kessler, “Polling the Blogosphere a Rule-Based Approach to Belief Classification,” in *Proceedings of the Second International Conference on Weblogs and Social Media*, *ICWSM*, 2008.

[8] Jason S. Kessler, and Nicolas Nicolov, “Targeting Sentiment Expressions through Supervised Ranking of Linguistic Configurations,” *3rd International AAAI Conference on Weblogs and Social Media*, 2009.

[9] Theresa Wilson, Janyce Wiebe, and Rebecca Hwa, “Just how mad are you. Finding strong and weak opinion clauses,” *Proceedings of AAAI*, 2004.

[10] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann, “Recognizing contextual polarity in phrase-level sentiment analysis,” *Proceedings of HLT-EMNLP*, 2005.

[11] Eric Breck, Yejin Choi, and Claire Cardie, “Identifying expressions of opinion in context,” *Proceedings of the 20th international joint conference on Artificial intelligence, IJCAI'07*, 2007, pp2683-2688.

[12] Tony Mullen, and Nigel Collier, “Sentiment analysis using support vector machines with diverse information sources,” *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2004.

[13] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, “Thumbs up? Sentiment classification using Machine Learning Techniques,” *Proceedings of EMNLP*, 2002, pp79-86.

[14] Sarah Schrauwen, “Machine Learning Approaches to Sentiment Analysis Using the Dutch Netlog Corpus,” *Computational Linguistics and Psycholinguistics Technical Report Series, CTRS-001*, 2009.

[15] Karo Moilanen, and Stephen Pulman, “Multi-entity Sentiment Scoring,” *Proceedings of Recent Advances in Natural Language Processing (RANLP 2009)*, 2009, pp258-263.

[16] А. Г. Пазельская and А. Н. Соловьев “Метод определения эмоций в русском языке,” *Компьютерная лингвистика и интеллектуальные технологии. Сб. научных статей*, Vol. 10 (17), 2011, pp510-522.

[17] V. S. Jagtap, and Karishma Pawar, “Analysis of different approaches to Sentence-Level Sentiment Classification,” *International Journal of Scientific Engineering and Technology*. Vol. 2, Issue 3, 2013, pp164-170.

[18] Andriy Mykulyak, “*UGTag*,” [online] 2009, http://www.domeczek.pl/~polukr/parcor/ (Accessed: 15 July 2013).

[19] Дівочі посиденьки, [online], http://posydenky.lvivport.com/ (Accessed: 10 July 2013).

[20] Львів. Відпочинок у Львові, [online], http://v.lviv.ua/ (Accessed: 10 July 2013).

[21] Paul Ekman “Basic Emotions”, *Handbook of Cognition and Emotion*, 1999, pp45–60.

[22] Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella et al., “Sentiment Analysis in the News,” *LREC 2010*, 2010, pp2216-2220.

[23] Kushal Dave, Steve Lawrence, and David M. Pennock, “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,” *WWW2003*, 2003.

[24] B. Yu, “Exploring the characteristics of opinion expressions for political opinion classification,” *Digital Government Conference’08*, 2008, pp82-91.

[25] Jack G. Conrad, and Frank Schilder, “Opinion Mining in Legal Blogs,” *ICAIL ’07*, 2007, pp231-236.

[26] Ann Devitt, and Khurshid Ahmad “Sentiment Polarity Identification in Financial News,” *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007, pp984–991.

AUTHOR

Mariana ROMANYSHYN is a Master of Applied Linguistics and is currently a PhD candidate of the Department of Computer-Aided Design, Lviv Polytechnic National University in Lviv, Ukraine. Her research area is Natural Language Processing.