Processing of Negation in Sentiment Analysis for the Serbian Language

Adela Ljaić, Ulfeta Marovac, Aldina Avdić

Abstract— The different approaches to the processing of negation in sentences written in Serbian and their influence on the sentiment analysis are presented in this paper. Sentiment analysis is still not sufficiently elaborated, especially for resource-limited languages. Processing of negation is important for sentiment analysis and significantly increases its quality. The existing solutions for the Serbian language which describe the sentiment analysis have not quite provided an appropriate and efficient way of processing the negation in sentiment analysis.

Index Terms—sentiment analysis, negation, lexicon, Serbian language.

I. Introduction

The problem of determination of texts sentiment and sentiment analysis is not new. With the appearance of machine processing of natural language and machine learning algorithms for classification, greater progress in this area happened. The largest number of developed tools and solutions for sentiment analysis are intended for texts in English [1-4]. The complexity of the Serbian grammar makes the analysis of texts in the Serbian language even more difficult. It is necessary to create an algorithm that classifies text written in the Serbian language as positive or negative, without the use of huge and abundant lexical resources.

To determine the sentiment of a text is not easy, there are problems which are difficult to resolve, such as, for example, negation, irony, and sarcasm.

Negation in a sentence significantly changes the meaning of a sentence and its sentiment. The processing of a negation is not an easy task. Most authors believed that a negation in sentence reverses the meanings of the sentence. They simply change the sentiment polarity. However, it is shown that such an approach is not enough. There are cases when the negation does not change sentence sentiment. Thus, we must be careful in negation processing. For the Serbian language, a negation must be processed according to specific rules and established usage.

The following sections describe the processing of negation in sentences written in the Serbian language. Section 2 describes algorithm of sentiment analysis. Section 3 presents an analysis of related work in the area of negation processing in the sentiment analysis. In Section 4 are proposed different solutions of the negation processing.

Applied machine learning methods and results of negation processing are discussed in Section 5, while Section 6 contains our conclusions and some directions of future work.

Adela Ljaić is with the State University of Novi Pazar, Vuk Karadžić bb, 36300 Novi Pazar, Serbia (e-mail: acrnisanin@ np.ac.rs).

Ulfeta Marovac is with the State University of Novi Pazar, Vuk Karadžić bb, 36300 Novi Pazar, Serbia (e-mail: umarovac@ np.ac.rs).

Aldina Avdić is with the State University of Novi Pazar, Vuk Karadžić bb, 36300 Novi Pazar, Serbia (e-mail: apljasković @ np.ac.rs).

II. Sentiment analysis

The aim of sentiment analysis is to find out the feelings, emotions, and opinions written in a text. A simpler approach to the sentiment analysis may require positive or negative attitude toward the text. A more complex approach to the sentiment analysis may be a required attitude to a text expressed numerically using a scale of 1 to 10.

The first step in the sentiment analysis is collecting and labeling data. Collecting data on the Serbian language is not an easy task. All available services do not offer enough options that would in any way limit the collection of texts written only in the Serbian language. Although there is the possibility of specifying the language for Serbian, Cyrillic is the default, so specifying the language leaves out the collection of texts that are written in Latin letters (larger than that of the Cyrillic alphabet).

Another problem with the data collection is the unequal ratio between positive and negative texts. For example, data from social networks in the field of sports contain the significantly higher percent of positive than negative feedback. In contrast, the data in the area of media contain the significantly higher percent of negative relative to the positive feedback (posts, tweets).

The collected text data are input in the process of sentiment analysis that takes place in several phases, depending on a specific algorithm.

Our approach applies machine learning algorithms in order to train a polarity classifier using a labeled corpus. One of famous approach is Pang and Lee's algorithm for sentiment analysis [1] which contains the following steps:

1. Pre-processing;
2. Separation of characteristics (attributes) that can be numeric or text;
3. Classification using the appropriate algorithm for the classification (Naive Bayes, Maxent, SVM, ...), using previously allocated attributes.

Our algorithm was organized in next six steps (II.A – II.F) which are closely described in [15].

A. Collecting data

We collected tweets on the Serbian language related to one area (for different media in Serbia). The collected text data are input in the process of sentiment analysis.

B. Labeling data

Our initial data set from the Twitter have no assigned sentiment so it must be manually labeled. The dataset is

divided in 2:1 ratio, the data for training and testing, respectively.

C. Normalization

As a part of the pre-processing step in our algorithm, all signs of excessive punctuation, links, tags that do not affect the content of the message have been removed. Tokenization has been performed based on the rules that apply to writing on twitter. For example, tagging of posts is followed by a hashtag (#), the @ sign have used to present usernames in tweets. All data is transferred in one alphabet (Latin). A collection of stop words that do not contain a sentiment and have high occurrence are removed from the data.

D. Stemming

We used stemmer described in [5] in order to transform the entire word occurrence in a unique form. Application stemmers showed an advantage over other approaches due to the inclusion of a wider set of words.

E. Feature extraction

Any information that does not affect the sentiment analysis of Tweets is removed by the previous steps. Each tweet can be represented by a set of words that are appearing in it. In order to reduce the dimension of a vector of tweet presentation, without losing the information that is essential for the sentimental analysis, we create lexicons of sentiment terms. Sentiment terms express the desired and undesired situations and can be divided into positive and negative sentiment terms. We constructed lexicons of positive and negative sentiment term, and they are general purpose.

Based on the existing corpus, we have also created general sentiment term lexicon that is specific to the domain that is the subject of analysis.

F. Machine learning

Although there are a number of very complex algorithms of supervised machine learning, the best results in the sentiment analysis and the general classification of human language and the speech were obtained with one of the simplest algorithm – Naive Bayes. Our data have been trained and tested using Naïve Bayes, Simple Logistic, KNN with one, and KNN with two nearest neighbors.

III. Literature review

The occurrence of negation in a sentence essentially affects the change of its sentiment so sentiment analysis must include the negation processing. Many authors is watching a sentiment for each term individually in the process of sentiment analysis, whereby the occurrence of negative keywords solved by changing the sentiment polarity of the term that follows the negation. According to [1], negated context is a segment of a tweet that starts with a negative term and ends with punctuation as a comma, period, colon, semicolon, exclamation mark, or question mark. One of the advanced approaches presented Kiritchenko in [6], where the term immediately after the negation is subject to the greater influence of negation than terms that follow later.

The negation changes sentiment polarity of a term or in some cases the intensity of sentiment. Most previous solutions are based on the change of a sentiment polarity. However, it could happen that the negation does not change sentiment polarity, but only the intensity of sentiment. Several studies have pointed to this problem ([8], [9]). In [6] is proposed a corpus-based statistical approach to estimate sentiment scores of individual terms in the presence of negation. They build two lexicons: one for terms in negated contexts and one for terms in affirmative (non-negated) contexts. They showed that negation of positive terms tends to imply negative sentiment, whereas in the case of negative terms the sentiment remains negative in the negated context.

All these solutions are related to the English language and for another language can have completely different results. There are several works that are processed sentiment analysis for the Serbian language [9-13]. By Milošević [9] a negation has an impact on all terms after appearing negative term ("ne"(not) and negation of the verb "jesam"(I'm)-"nisam"(I'm not) and "hteti"(want)-"neću"(don't want) to the first punctuation character. Mladenovic in [10] and [11] used ready-made lexical resources for the Serbian language. These lexical resources include the negation of each word if it exists. Batanović [12] treats the negation on the basis of work [1], but he use that scope of the negation applies only to the first two words after the negation, as opposed to the basic method where the scope of negation is valid for all the words to the first punctuation mark. He showed that his solution gives better results for his data set. Grljević [13] made the lexicon of negative terms and the lexicon of modifiers that change the intensity of sentiment. She processes all the terms after the negative term to the first punctuation character. The number of occurrences negation is taken as an attribute in the training data. It was pointed to the problem of double negation phenomenon that couldn't be processed using unique rule.

IV. Analysis of negation for the Serbian language

In the Serbian language, it is distinguished lexical and syntactic negation. Lexical negation relates to nouns, adjectives, adverbs and pronouns and it is realized by the addition of a prefix "ne", "ni" and "bez". Such lexemes are called negative lexemes or negative words. Negation of the syntactic level, the level of the sentence is carried out using "ne" and "ni", it stands alone or attached to the verb[15]. Verbs are negated by adding the negative term "NOT" before the verb.

There are exceptions when the negation of verb creates a new negative term: "ne biti"(not to be) in a present: "nisam"(I'm not), "nisi"(you are not), "nije"(it's not), "nismo"(we are not), "niste" (you are not), "nisu" (they are not); "ne biti" (not to be) in an imperative: "nemoj"(do not), "nemojmo"(let's not), "nemojte"(do not); "ne hteti"(do not want) in a present: "neću"(I do not want), "nećeš" (you do not want), "neće" (he does not want), "nećemo" (we do not want), "nećete" (yoi do not want), "neće" (they do not want); "ne imati"(do not have) in a present: "nemam"(I do not have), "nemaš"(you do not have), "nema"(he does not have), "nemamo"(we do not have), "nemate"(you do not

have), “nemaju”(they do not have).

Lexical negation is covered by the sentiment lexicons in sentiment analysis. Negation of the syntax level refers to the sentence and it should be treated. In order to process the negation, we have made a collection of keywords which negate some of the content of the sentence. This collection, lexicon of negation keywords contains 60 terms. Beside negative terms in this lexicon are included terms like: “nijedan”(no one), “nikakav”(no, neither), etc.

On the basis of this lexicon, it is determined the beginning of the tweet negation. The range to which negation has influence in the sentence is divided into two cases:

1. The first term after negation keywords.
2. All terms after negation keywords to the first punctuation character.

In the first case, the first term that appears after the negative keyword will change the sentiment polarity without changing the intensity of sentiment.

For the second case, negation will change the sentiment of all the terms until first punctuation character on the following way:

a) By changing only the sentiment polarity of terms.
b) By changing the polarity of all of the terms but reducing the sentiment intensity of each term after the first term after negation by 20%.

The intensity of sentiment is reduced by 20% for every word (case b) because we have experimentally found that this percentage of reduction of intensity gives the best improvement.

V. RESULTS

First, we examined the effects of the negation of the term immediately after the negative key term, for terms belonging to the positive and negative lexicons.

Lexicon of negative terms contains 3089 terms while lexicon of positive terms contains 1249 terms.

The negation of the positive term appears more in tweets with a negative sentiment than in the tweets with a positive sentiment; only 4 positive terms with negation appeared more in tweets with positive sentiment than in tweets with negative sentiment. Negation for 4 positive terms is presented in Table 1. The number of terms occurrences in tweets is normalized by the ratio of the number of positive and negative tweets.

TABLE 1. THE NEGATION OF THE POSITIVE TERMS EXAMPLE

stemmed term	regular in positive tweets	negation in positive tweets	regular in negative tweets	negation in negative tweets	regular in neutral tweets	negation in neutral tweets
napred	0	11.65449	1.87594	0	2.62378	0
pl	0	11.65449	16.88342	0	10.49514	0
sjaj	23.30897	11.65449	0	0	0	0
svid	11.65449	11.65449	0	1.87594	0	0

These results indicate justified changing the polarity of terms with negation. There are cases when negation changes the intensity of sentiment - these cases will be handled testing vocabulary on a larger corpus.

We processed negation in three ways:

1. Method1 – changing polarity only of the first term after negation keywords.
2. Method2 – changing polarity of all terms after negation keyword to the first punctuation character
3. Method3 – changing polarity of all terms after negation keywords term to the first punctuation character, but reducing the sentiment intensity of each term after the first term after negation by 20%.

TABLE 2. THE NEGATION OF THE NEGATIVE TERMS EXAMPLE

stemmed term	regular in positive tweets	negation in positive tweets	regular in negative tweets	negation in negative tweets	regular in neutral tweets	negation in neutral tweets
poj	0	0	0	3.75187	0	0
bezobrazl	0	0	0	1.87594	0	0
skup	0	0	1.87594	1.87594	2.62378	0
st	0	0	16.88342	1.87594	0	0
student	0	0	3.75187	1.87594	0	0
sukob	0	0	7.50374	1.87594	2.62378	0
sumnj	0	0	1.87594	1.87594	0	0
ter	11.65449	0	1.87594	1.87594	0	0
trag	0	0	7.50374	1.87594	2.62378	0
ubi	0	0	7.50374	1.87594	2.62378	0
uc	0	0	0	1.87594	0	0

Sentiment analysis has been done with each of three methods. The methods are tested on the set of 3508 tweets for training and 1726 tweets for test. Processing was carried out in Weka software using:

a) Naïve Bayes multinomial (NBM),
b) Simple Logistic (Slog),
c) k-nearest neighbors with 2 nearest neighbor (KNN-2) and
d) k-nearest neighbors with 3 nearest neighbor (KNN-3) machine learning methods.

Method1 shows an increased accuracy in the case of use of KNN-2 and KNN-3. The precision is improved in the case of a SLog, KNN-2 and KNN-3. The results for Method1 are presented in Table 3.

TABLE 3. METHOD1-CHANGING POLARITY ONLY TO FIRST TERM AFTER NEGATION

		No negation	First term after negation
Accuracy	NBM	53.3637	53.2782
	Slog	57.1266	56.8985
	KNN- 2	56.0433	56.2999
	KNN-3	56.1859	56.2429
Precision	NBM	0.504	0.503
	Slog	0.565	0.559
	KNN- 2	0.541	0.547
	KNN-3	0.543	0.547
Recall	NBM	0.534	0.533
	SLog	0.571	0.569
	KNN- 2	0.56	0.563
	KNN-3	0.562	0.562
F-measure	NBM	0.504	0.503
	SLog	0.556	0.552
	KNN- 2	0.544	0.548
	KNN-3	0.546	0.548

Method2 shows increased accuracy in all four methods of machine learning. Precision, Recall, and F-measure also gives much better results. The results for Method2 are presented in Table 4.

TABLE 4. METHOD2-CHANGING POLARITY TO ALL TERMS AFTER NEGATION KEYWORDS TO THE FIRST PUNCTUATION CHARACTER

		No negation	All terms after negation
Accuracy	NBM	53.3637	53.5063
	Slog	57.1266	57.5257
	KNN- 2	56.0433	56.4139
	KNN-3	56.1859	56.3284
Precision	NBM	0.504	0.512
	Slog	0.565	0.575
	KNN- 2	0.541	0.546
	KNN-3	0.543	0.546
Recall	NBM	0.534	0.535
	SLog	0.571	0.575
	KNN- 2	0.56	0.564
	KNN-3	0.562	0.563
F-measure	NBM	0.504	0.51
	SLog	0.556	0.563
	KNN- 2	0.544	0.549
	KNN-3	0.546	0.549

Method3 shows the enhancement of Accuracy, Precision, Recall and F-measure by using all four methods of machine learning. The results for Method3 are presented in Table 5.

It should be noted that the sentiment analysis was performed with a small number of features, in order to see how the negation processing influence on the result. Hence, the results of the analysis of the sentiment are generally less good than that is shown in the earlier paper [14].

TABLE 5. METHOD3-CHANGING POLARITY TO ALL TERMS AFTER NEGATION KEYWORDS TO THE FIRST PUNCTUATION CHARACTER BUT REDUCING THE SENTIMENT INTENSITY FOR FOLLOWING TERMS

		No negation	All terms after negation- with scale
Accuracy	NBM	53.3637	53.7058
	Slog	57.1266	57.6682
	KNN- 2	56.0433	56.3854
	KNN-3	56.1859	56.2429
Precision	NBM	0.504	0.508
	Slog	0.565	0.576
	KNN- 2	0.541	0.548
	KNN-3	0.543	0.547
Recall	NBM	0.534	0.537
	Slog	0.571	0.577
	KNN- 2	0.56	0.564
	KNN-3	0.562	0.562
F-measure	NBM	0.504	0.506
	Slog	0.556	0.564
	KNN- 2	0.544	0.549
	KNN-3	0.546	0.547

VI. CONCLUSION

Testing how negation changing sentiment positive and negative terms in Serbian language, we concluded that in most cases the negation changes the polarity of terms. The scope of influence of negation is greater than the first following terms but negation still has the greatest impact on the first term. The impact on following terms is reduced. Conducted an experiment aimed to show how the negation behaves in Serbian language and provides a good basis for the improvement of sentiment analysis algorithms. Using a larger corpus, we plan to extract coefficients which show the impact of negation to the intensity of the polarity of a term. Negation can be treated better with additional rules but expected improvement in predicting sentiment is little. The question is how much it's worth to improve the impact of negation on sentences' sentiment, because of low accuracy magnification expectance.

ACKNOWLEDGMENT

This work was partially funded by the Ministry of Education and Science of the Republic of Serbia after the project III-44 007.

REFERENCES

[1] B. Pang, L. Lee, H. Rd, and S. Jose, "Thumbs up ? Sentiment Classification using Machine Learning Techniques", *EMNLP '02 Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, vol. 10., 2002, pp. 79–86.

[2] C. D. Manning, J. Bauer, J. Finkel, and S. J. Bethard, "The Stanford CoreNLP Natural Language Processing Toolkit", *Association for Computational Linguistics*, 2014, pp. 55–60.

[3] S. Paumier, *Unitex*, e Institut Gaspard-Monge (IGM), University Paris-Est Marne-la-Vallée, Paris, 2002.

[4] M. Silberztein, *Nooj*, http://www.nooj4nlp.net, 2002.

[5] Nikola Milošević, *Stemmer for Serbian language*, http://arxiv.org/abs/1209.4471, arXiv preprint arXiv:1209.4471, 2012.

[6] S. Kiritchenko, X. Zhu, and S. M. Mohammad,"Sentiment Analysis of Short Informal Texts", Volume 50, pages 723-762 Journal of Artificial Intelligence Research

[7] A Kennedy, D Inkpen , Sentiment classification of movie reviews using contextual valence shifters, Computational intelligence 22 (2), 110-125, 2006.

[8] M Taboada, J Brooke, M Tofiloski, K Voll, M Stede, Lexicon-based methods for sentiment analysis, Computational linguistics 37 (2), 267-307, 2011.

[9] N. Milošević, "Mašinska analiza sentimenta rečenica na srpskom jeziku", *Master's Degree Thesis*, University of Belgrade, Belgrade, Serbia, 2012.

[10] M. Mladenović, J. Mitrović, C. Krstev, D. Vitas, " Hybrid SentimentAnalysis Framework For A Morphologically Rich Language", *Journal of Intelligent Information Systems*, vol. 46:3, 2016, pp 599–620.

[11] M. Mladenović, *Information Models in Sentiment Analysis Based on Linguistic Resources*, Doctoral Dissertation, University of Belgrade, Belgrade, Serbia, 2016.

[12] V. Batanović, B. Nikolić, M. Milosavljević, "Reliable Baselines for Sentiment Analysis in Resource-Limited Languages: The Serbian Movie Review Dataset", *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 2688-2696, Portorož, Slovenia.

[13] O. Grljević, *Sentiment in Social Networks as Means of Business Improvement of Higher Education Institutions*, Doctoral Dissertation, University of Novi Sad, Novi Sad, Serbia, 2016.

[14] A. Ljaić, U. Marovac, A. Avdić, Sentiment Analysis of Twitter for the Serbian Language, ICIST, Kopaonik, Proceedings of Papers, 2017

[15] Branka L. Mladenović ,Equivalence of Syntactical and Lexical Negation, Univerzitet of Belgrade, DOI10.7251/FIN1301391M ,2013