Sentiment Analysis of Twitter for the Serbian Language

Adela Ljajić*, Ulfeta Marovac*, Aldina Avdić*

* State University of Novi Pazar, Novi Pazar, Serbia

acrnisanin@np.ac.rs

umarovac@np.ac.rs

apljaskovic@np.ac.rs

Abstract— **Although short texts, tweets and other types of comments that can be found on social networks, carry significant information about the person who wrote them and the object to which they refer when they are cumulatively observed. The interest for machine analysis of the feelings that are expressed by some text grows, but algorithms that perfectly determine emotion that text carries are still not found. Problems that occur in the machine analysis of the text are different, starting with the complexity of the language in which the text is written to the complexity of the feelings expressed by it. This paper provides an overview of some of the existing solutions for sentiment analysis and gives an overview of our solution. Sentiment analysis will be shown in the case of tweets written in the Serbian language. Although the method is still in the development phase, the resulting accuracy of 82% is encouraging at this stage of the method development.**

I. Introduction

The increasing use of social networks offers the opportunity to explore a variety of social issues by analyzing the comments that people are posting. The problem of sentiment analysis and determination of comments sentiment is not new, but with the appearance of machine processing of natural language and machine learning algorithms for classification, greater progress in this area happened. Most developed tools for sentiment analysis are for texts in English [1-4]. To determine the sentiment of a text is not easy, there are problems which are difficult to resolve, such as, for example, irony and sarcasm. The complexity of the grammar of the Serbian language makes the analysis of texts in the Serbian language even more difficult. It is necessary to create an algorithm that manages to classify text written in the Serbian language as positive or negative, without the use of huge and abundant lexical resources.

Sentiment analysis belongs to the field of natural language processing which is on the rise, unlike areas such as spam detection, detection and labeling type, gender and number of words in a sentence, which are mostly areas where they obtained pretty good results.

Sentiment analysis is one of the sub-areas of natural language processing, i.e. NLP (Natural Language Processing). Most generally speaking, sentiment analysis is a classification of text documents into two or more classes. The simpler task of sentiment analysis may require the determination of whether the text has a positive or negative sentiment. The more complex task of sentiment analysis may require numerical quantification of sentiment text using scales, for example, from 1 to 10. Text classification and sentiment analysis as a special case of text classification are problems to be solved with the help of artificial intelligence or with the help of supervised machine learning algorithm.

Since it is a language processing, it is necessary to do some preprocessing to help in getting better results even on small training data sets. One such process is to prepare the text in a single format (for the Serbian language select one alphabet, for example, Latin alphabet), to process special characters, to remove redundant words (stop words), to tokenize text, etc.

For the Serbian language, only a few solutions for sentiment analysis were developed [5-9]. Some of problems of sentiment analysis of texts in the Serbian language are:

- The lack of labeled data;
- Lack of vocabulary of terms with positive and negative sentiment on the Serbian language;
- Non-uniform normalization (stemming, lemmatization) because of complexity of the Serbian language;
- Various pronunciation (*ekavica*, *ijekavica*) and dialects;
- Characters with diacritical marks (č,ć,š,...)
- Two alphabets (need of translation into one, Latin or Cyrillic);
- Informal writing on social networks.

The process of text classification based on sentiments and altogether with the preprocessing process can still be upgraded by seeking solutions to problems such as normalization, keywords extraction, symbols, negation, sarcasm, etc. Previous solutions of sentiment analysis for the Serbian language are described in the sequel. Section 3 describes the methodology of work. The normalization and feature extraction are presented in Section 4 and

Section 5. Applied machine learning methods and results are discussed in Section 6, while Section 7 contains our conclusions and some directions of future work.

II. Previous solutions for the Serbian language

As previously stated, there are several solutions which attempt to classify the text in the Serbian language, based on the sentiment that text carries. They differ in the domain to which had applied to, the method of text normalization used, and the methods of the classifications used.

Milošević in [5] described implementation of sentiment analyzer for Serbian language using Naive Bayes algorithm of machine learning. As a part of sentiment analyzer is described and developed hybrid stemmer for Serbian language that works on principles of suffix stripping and dictionary. Analyzer of sentiment does binary classification of text into two classes - positive and negative. He processed negation by adding prefix to words near negation.

Batanović et al. presented in [8] a dataset balancing algorithm that minimizes the sample selection bias by eliminating irrelevant systematic differences between the sentiment classes. They use it to create the Serbian movie review dataset – SerbMR – the first balanced and topically uniform sentiment analysis dataset in Serbian. They proposed an incremental way of finding the optimal combination of simple text processing options and machine learning features for sentiment classification.

Mladenović's doctoral thesis [7], whose main task is the analysis of emotions in text, presents research related to the sentiment classification of texts in the Serbian language, using a probabilistic method of machine learning of multinomial logistic regression i.e. maximum entropy method. She developed a system for sentiment analysis of Serbian language texts, with the help of digital resources such as: semantic networks, specialized lexicons and domain ontologies.

Doctoral dissertation of Grljević [9] provides a comprehensive approach to modeling and automation of analysis of sentiments contained in student reviews of teaching staff available on social media and social networking sites. Reviews of professors from various institutions of higher education in Serbia, have been collected from the site "*Oceni profesora*". The prepared and annotated corpus was used in sentiment analysis for training the supervised algorithms (Naive Bayas, SVM, K-Nearest Neighbor). Five dictionaries were developed based on the annotation of sentiment words, negation keywords, and observations derived during the annotation process, which were consequently used in sentiment analysis based on lexicons.

III. Data Set and Methodology

The first step in the sentiment analysis is the collection of resources (texts) to be analyzed. In our case, these are the tweets on the Serbian language related to one area (for example, for different media in Serbia).

Collecting tweets 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 tweets 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 tweets 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 tweets. Tweets in the field of printed media, for which we have been doing the collection of data, contains tweets that carry more negative than positive sentiment. Therefore, it is necessary to extend the model with positive tweets in order to design a relevant training set.

Collecting data for sentiment analysis in resource-limited languages carries a significant risk of sample selection bias, since the small quantities of available data are most likely not representative of the whole population.[8]

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.

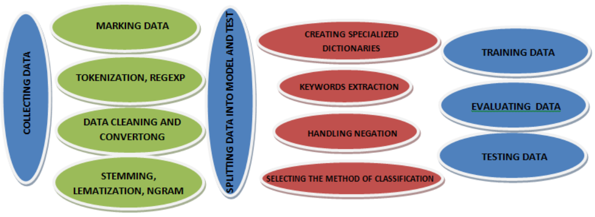
COLLECTING DATA
MARKING DATA
TOKENIZATION, REGEXP
DATA CLEANING AND CONVERTONG
STEMMING, LEMATIZATION, NGRAM
SPLITTING DATA INTO MODEL AND TEST
CREATING SPECIALIZED DICTIONARIES
KEYWORDS EXTRACTION
HANDLING NEGATION
SELECTING THE METHOD OF CLASSIFICATION
TRAINING DATA
EVALUATING DATA
TESTING DATA

Figure 1. The algorithm for sentimental analysis which is applied to the set of tweets.

Our work was organized in next six steps (Fig. 1):

1. Collecting data using Twitter API;
2. Labeling data manually;
3. Normalization;
4. Tokenization;
5. Stemming;
6. Splitting data for training and test sets;
7. Feature extraction;
8. Applying method for training and test.

Our initial data set from the Twitter have no assigned sentiment so that one of the first steps must be their labeling. The dataset is divided in 2:1 ratio, the data for training and testing, respectively.

IV. 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. In tokenization process, before applying sentiment analysis, special attention should be paid to emoticons and their way of writing. Special regular expressions that include known emoticons have been used. Correct processing of emoticons could be important for labeling the starting data.

Another aggravating circumstance for the text written in the Serbian language is the use of two alphabets. Therefore, all data is transferred in one (Latin). A collection of stop words that do not contain a sentiment and have high occurrence are removed from the data.

Normalization of documents in the Serbian language can be done by lemmatization or stemming. It's hard to make a perfect lemmatizer so that we have to be satisfied by using stemmers [10]. Disadvantages of stemming, caused by removing only the suffix (not the prefix) can be solved using n-gram analysis [11]. Normalization and tokenization in our work have been done in the following steps:

- Removing retweets
- Tokenization
    - Dates handling
    - Currency expression handling
    - Numbers in specific format handling
    - URLs handling
    - Hash tags handling
    - Emoticon handling
- Translate to Latin letter
- Removing stop words
- Stemming.

V. FEATURE EXTRACTION

Each document or sentences which are the subject of the sentiment analysis can be represented as a vector consisting of all the words which appear in them. The problem that arises is the high dimensionality of the vectors that describe the text item. When applying a machine learning method, high vectors dimensionality can lead to excessive training (overfitting). Therefore, we seek to reduce the dimension of a vector, without losing the information that is essential for the classification. Attributes that stand out as carriers of information may be different [12]: the presence and frequency of key words, parts of speech, sentiment words and phrases, negation. Feature extraction was based on the lexical resources and allocation of key terms.

Sentiment words 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 are specific to the domain that is the subject of analysis. As these general lexicons don’t cover the entire set of words, lexicon which represent the adjustment of the general sentiment lexicon to the corpus, have also been extracted. So we obtained two types of sentiment lexicons:

- General sentiment lexicon of positive and negative terms, manually created
- Model generated sentiment lexicon.

Every lexicon has its own advantages and disadvantages:

- General lexicon advantage: large word set.
- General lexicon disadvantage: ambiguity of word meaning.
- Model generated sentiment lexicon advantages: closely related to the specific field.
- Model generated sentiment lexicon disadvantages: limited with size and quality of the model.

Negation was treated by isolating a set of words which form negation in the Serbian language and using this set for creating special rules for handling negation.

VI. MACHINE LEARNING METHOD

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 Simple method, Naïve Bayes and SVM, as shown in Table I.

TABLE I. RESULTS OF APPLYING DIFFERENT MACHINE LEARNING METHODS

	Precision	**Recall**	**F-Measure**	**Correctly**	**Incorrectly**

Simple method	0-0.6567;
	2-0.8691;
	4-0.7319
		0-0.9882;
		2-0.2087;
		4-0.9070
			0-0.7889;
			2-0.3366;
			4-0.8096
				68.4	31.6

Simple method (test set)	0-0.5690;
	2-0.4835;
	4-0.3736
		0-0.8962;
		2-0.1237;
		4-0.4096
			0-0.6960;
			2-0.1971;
			4-0.3907
				54.1	45.9

Naïve Bayes (cross valid. 10-fold)	0.743	0.737	0.739	73.7457	26.2543
Naïve Bayes (test set)	*0.531*	*0.527*	*0.515*	*52.6651*	47.3349
SVM with normalized poly kernel (cross valid. 10-fold)	**0.822**	**0.822**	**0.822**	**82.1836**	17.1836
SVM with normalized poly kernel (test set)	0.506	0.478	0.466	47.7984	52.2016

Data set consists of 3508 tweets for training and 1726 tweets for test.
Training feature we used are the following:

- number of positive terms from lexicon
- number of negative terms from lexicon
- number of positive terms appearing only in positive tweets and not in negative tweets
- number of negative terms appearing only in negative tweets and not in positive tweets
- number of terms in tweet.

Simple method is simply counting features: terms from positive/negative lexicon and terms appearing only in positive/negative tweets. Naive Bayes and SVM are using the same features as Simple method and additionally the number of terms in tweet.

The results are still not satisfactory and mostly because of the small set of data. Cross-validation indicates that this method can get a good result but we still have to work on attributes for new features. The problem is irony, incomplete tweets, informal communication on Twitter, few tweets with positive sentiment.

VII. Conclusion

Presented algorithm is adapted to the structure of the text as a tweet, and to the topic on which tweets was referring. The accuracy of classification was increased with the deeper observation of the individual areas and aspects of the text which is analyzed, as well as the creating of specific vocabulary and rules for each area. The method should be improved with:

- Dataset with better positive-negative ratio;
- Handling negation and irony;
- Normalization with n-grams.

Our method is at the beginning of development. A small number of features used for training. Model still gives results that are encouraging for this phase of research. The contribution of this work is mostly in constructed set of lexical resources which are general purpose and don’t depend on the data source.

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., 1988, 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] N. Milošević, “Mašinska analiza sentimenta rečenica na srpskom jeziku”, *Master's Degree Thesis*, University of Belgrade, Belgrade, Serbia, 2012.

[6] 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.

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

[8] 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.

[9] 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.

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

[11] U. Marovac, A. Pljasković, A. Crnišanin, E. Kajan, „N-gram analiza tekstualnih dokumenata na srpskom jeziku“,*Proceedings of TELFOR 2012*, 2012, pp. 1385-1388.

[12] W. Medhat, A. Hassan, H. Korashy , „Sentiment analysis algorithm and applications: A survey“, *Ain Shams Engineering Journal*, 72(2), 2014, pp. 305-328.