Towards the Lexical Resources for Sentiment-Reach Informal Texts - The Serbian Language Case

Ulfeta Marovac[*], Adela Ljajić, Ejub Kajan, and Aldina Avdić

State University of Novi Pazar, Republic of Serbia

Abstract

In this paper, the sentiment dictionary is presented as one of the lexical resources necessary for the sentiment analysis. The use of this resource is related to the existence of other specific resources that are adapted to the ultimate goal - in this case, sentiment analysis. Thus, in order to normalize documents for sentiment analysis, a specific stop word dictionary is needed (for example, it does not contain negation signals), then a stemmer or a lemmatizer to transform the word to the basic form, to which the words from the sentence must be reduced. Informal writing on social networks leads to the impossibility of applying standardized lexical resources. Therefore, in order to improve the results parallel lexical resources are made.

Keywords— sentiment analysis, social media, n-grams, lexicon

1 Introduction

People living in the 21st century are witnessing and experiencing the impact of social media on their private and public lives. According to Chaniotakis et al. [3], every minute, almost 250,000 tweets are posted on Twitter, almost 300,000 Facebook statuses are updated, and about 136,000 photos are uploaded to Facebook. The phenomenon is called "social fever". It brings a large amount of available data, the meaning of which has not known in advance, usually unstructured, written in any spoken language, and full of sentiments. These sentiments may be inserted by tagging, like, post, comments, share, follow, and other actions that social media platforms allow. We called that *sentiment-reach environment.*

Living in the digital age and competitive market require faster access to information about a new product, public personality, or a popular topic. Many

[*] Corresponding author, email: umarovac@np.ac.rs

applications like e-commerce, e-health, e-learning, e-government, etc., try to benefit of using aforementioned data. Due to characteristics of these data, it is not easy to achieve. Diamantini et al. [5], emphasize that, despite images and videos that people and organizations post to social media, textual data is the main source of information often contains opinions and feelings (sentiments), whose contents is considered authentic due to the freedom of expressing the thoughts of users by themselves. In that way, in [5] two main text analysis tasks are entity extraction (recognize relevant peace of information) and sentiment analysis (to determine user's opinion about a particular topic).

Several resources have been created for the processing of texts written in Serbian, such as: wordnet [9], morphological dictionary [22], stemer[14], the stop word dictionary [12], etc. However, text processing in the Serbian language is hampered by the lack of adequate specific lexical resources as sentiment dictionary.

Our contribution is manifold: (1) requirements analysis for creating sentiment lexicons under diversity of informal data; (2) proposing methodology and algorithms for creating such lexicons and (3)validation of the lexicon by the set of several thousands of tweets.

The remainder of this paper is organized as follows. Section 2 describes the problem, section 3 is devoted to background that brings some relevant definitions and related work on sentiment analysis and lexicons, section 4 presents experiment with some statistical data obtained, and finally, some concluding remarks are given.

2 Problem description

Tweets are informal short texts by which people often give their opinion about certain phenomena, things, personalities. The definition of the sentiment that is expressed by the tweet is specific due to the length of the text and the informal language in which they are written. The sentiment in tweets is most often expressed by sentiment words. For the sentiment of tweets determination, we developed the lexicon with positive and negative sentiment words [11]. As tweets are short texts, the number of occurrences of sentiment words in them is small, that makes their presence more significant. To determine the sentiment only on the basis of such words, there are aggravating circumstances such as the presence of negation, the use of irony, etc.

In order to process any text, it must pass the process of normalization. Normalization consists in reducing words with the same meanings of the same form and remove words that do not affect the determination of sentiment. To remove words that do not affect sensitivity analysis, we used the lexicon of stop words [12] by excluding words that participate in a negation forming.

Reducing words to the same form can be done by cutting them to n-grams of a certain length, stemming or lemmatization. All three forms have their advantages and disadvantages depending on the nature of text on which they are applied and on the purpose of the text analysis. Informal writing can lead to the poor application of either lemmatizer or stemmer, or both. In sentiment analysis, very often a word with a sentiment was not found in the dictionary due to an error in writing, improper use of the stemmer or lemmatizer.

To solve aforementioned problems, we evaluate the influence of various types of normalization on mapping sentiment words. The parallel resources, depending on the type of normalization that has been applied, could improve mapping of sentiment words in tweets.

3 Background

This section provides an overview of the related work on sentiment analysis and sentiment lexicons supported with some definitions that support background.

3.1 Some definitions

Sentiment lexicon resources. Resource in ICT related fields means different things. *"Despite the multiple uses of resources, they abstract some entities, whether physical or logical, that could be discovered, composed, and consumed so that certain business goals are achieved"* [1]. In the context of this paper a lexicon resource is a linguistic repository in machine-readable format that allows applications to process informal data in a meaningful manner by consulting such a resource and help companies and government to evaluate the performance of products or services. Lexicon resources are created either manually, semi-automatic, or automatic. Sentiments are words that are commonly used to express positive or negative opinions. A list of such words and phrases is called a *sentiment lexicon.*

Natural language processing(NLP). A field of research in artificial intelligence, information retrieval, Web mining, and similar disciplines, that process information given in written language by some criteria. In [20], several paradigm shifts of NLP are explained, starting with bag-of-words assumptions, latent semantic analysis, topic modeling (a.k.a. shallow NLP), etc., and towards to deep NLP with information extraction: a technique for automatic recognition of named entities from text using supervised learning.

Sentiment Analysis(SA). An ongoing field of research *"aims at determining opinions, emotions, and attitudes reported in source materials like documents, short texts, sentences from reviews, blogs, and news, among other sources"* [4]. SA process can be applied on a variety of textual sources on different granularity levels (an entire document, phrases, separate words) [6] and may include several steps, like tokenization, POS tagging and lemmatization, that are usual in any NLP task, but also may include word disambiguation [5].

3.2 Related work

3.2.1 Sentiment analysis

Like other areas of the NLP, SA is the mostly developed for the English language. Information on social networks is related to a specific geographic and/or speaking area, and so developing tools need to be tailored to specific languages. Most SA techniques are either machine-learning oriented or dictionary-based. According to [16], the former approaches have made significant advances but require labeled training data sets whose compilation is time consuming and seeks for additional efforts. The later approaches are based on lexicons. Examples include, but not limited to, WordNet [13], MPQA [23], etc. Despite the well-recognized feature that once they have built, no extra training sets are required, certain drawbacks are exist, too. In fact, they are usually based on a common language (e.g. lack of technical, medical, and other domain-specific terms) and suffer to catch up desired accuracy in multi-domain scenarios [16].

Defining sentiments is not just a straightaway look into a sentiment lexicon. Some other problems in SA are the ambiguity of words or syntagms, non-standard writing, use of slang, neologisms, segmentation problems, irony, metaphor, and negation. Pandey et al. [17] classify SA methods to those which are either lexicon-, or machine-learning-, or hybrid -based. The lexicon-based methods only, are used to sentiment identification in unsupervised cases to assign polarity scores to individual words for detecting the overall sentiment of a document. The overview of lexicon-based methods may be found in [21]. However, they suffer from several reasons that may be summarized as follows:

- *Low coverage*: i.e. documents may contain non of the lexicon's worlds, emoticons, etc. [6]. This aspect is also recognized by Saif et al. as the full dependence on the presence of words or syntactical features that explicitly reflect a sentiment [18].

- *They are restricted by their lexicons* [18]. These restrictions are multifold: (1) most are English-oriented [19]; (2) most of them are unigram-based [19]; (3) many of them are created manually and usually constrained by the target domain [19].

In contrast, in favor to lexicon-based methods for sentiment analysis, Taboada et al. [21] claim that they are robust, may be result with good cross-domain performance, and could be easily enhanced with multiple knowledge sources.

There are several solutions which attempt to classify the text in the Serbian language, based on the sentiments carry by text. They differ in the domain to which had applied to, the method of text normalization used, and the methods of the classifications used. In [14], implementation of sentiment analyzer for Serbian language using Naive Bayes algorithm of machine learning, is described, and a hybrid stemmer for Serbian language that works on principles of suffix stripping and dictionary, as well. Batanović et al. [2] presents a dataset balancing algorithm that minimizes the sample selection bias by eliminating irrelevant systematic differences between the sentiment classes. They are used to create the Serbian movie review dataset SerbMR the first balanced and topically uniform sentiment analysis dataset in Serbian. Mladenović [**?**]**E2U: staviti ovu referencu u bib**, analyze emotions in text written 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. Grljević [15] 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.

3.2.2 Sentiment lexicons

SA can be done either as a classical classification of text by using a classiffier-based, or lexicon-based approach [10]. Dor the former, sentiment lexicons are important resources. In fact, general purpose dictionaries and domain-specific, are exist . They differ in whether the words in it have a degree of polarity (-n to n) or simply contain a polarity (positive or negative). Semi-automatically created dictionaries start from a number of basic sentiment words and upgrade the vocabulary to synonyms; This approach is presented in [**?**]. Some authors create special sentiment dictionaries of words that appear in the scope of some linguistic phenomena (negation, irony, sarcasm, etc.). In [8] a special sentiment dictionary for words that are negated (participating in negation) is created. There is no publicly available sentiment dictionary for the Serbian language. In [11] authors have constructed a sentiment dictionary for testing the impact of negation on SA. The dictionary is based on the Opinion Lexicon [7] for English and is expanded by using the synonyms. It consists of 6183 sentiment words.

4 Experiment

Few words

4.1 Normalization

The sentiment lexicon contains words that carry a positive or negative sentiments. It would be expected that positive (negative) sentiment words appear in tweets with a positive (negative) sentiment more than in others By applying different types of normalization, the validity of this assertion was verified as well as the effect of normalization on the mapping of sentiment words in the tweets, as show in Fig 1.

To perform the analysis, a set of 5011 tweets in the Serbian language was collected, in which popular personalities from public life were tagged (athletes, artists, politicians, etc.). Due to the nature of the content to which they relate, there is a large number of collected negative tweets 4193 in relation to the tweets with positive sentiment 818.

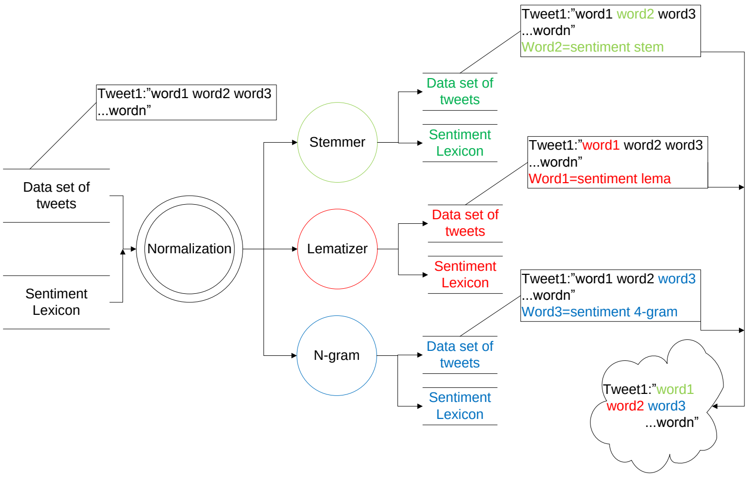
Tweet1:"word1 word2 word3 ...wordn"
Data set of tweets
Sentiment Lexicon
Normalization
Stemmer
Lematizer
N-gram
Data set of tweets
Sentiment Lexicon
Data set of tweets
Sentiment Lexicon
Data set of tweets
Sentiment Lexicon
Tweet1:"word1 word2 word3 ...wordn" Word2=sentiment stem
Tweet1:"word1 word2 word3 ...wordn" Word1=sentiment lema
Tweet1:"word1 word2 word3 ...wordn" Word3=sentiment 4-gram
Tweet1:"word1 word2 word3 ...wordn"

Figure 1: Mapping sentiment words in tweets**E2U-dodati na sliku Twitter-API**

The sentiment lexicon as well as the set of tweets has been normalized in three ways:

1. by stemming the word using stemmers [14],
2. by lematizing the word using the morphological dictionary [22],
3. by taking N-grams (4 grams)[12].

Three parallel sentiment lexicons were obtained by applying different normalization. They contained:

1. the stems of sentiment words(sentiments stem lexicon-SSL);
2. lemas of sentiments words(sentiments lema lexicon-SLL);
3. and 4-grams of sentiment words(sentiments n-gram lexicon-SNL).

4.2 Evaluation method of sentiment lexicon

The validity of the vocabulary is based on a set of tweets that are normalized by the corresponding normalization. For each form of sentiment word (stem, lemma or n-gram) the score (the number of occurrences of that word in positive and neglected tweets) is calculated. The number of words occurring in positive (negative) tweets is diminished if a word is in the range of negation and changes

Table 1: Number of sentiments words and contradictory data in different normalized sentiment lexicons

normalization	#sentiment words	#contradictory data
stemming	5951	22
lematization	5473	20
n-gram	2697	289

the polarity (it would be added to the opposite scores). Normalization of the score is done by dividing the score with the number of tweets from that class (the data set is unbalanced). For example, if we have the word "zvezda" (*"star"*) that it appears n times in positive and m times (n»m) in negative tweets, then the score will be calculated as:

$$score = \frac{n}{num\_positive\_tw} - \frac{m}{num\_negativ\_tw}.$$

The number of sentiment words is reduced by normalizations. The small amount of words with a different sentiment is reduced to the same basis, due to which it becomes contradictory. Normalizations based on language rules produce a lower number of such words. Contradictory data will be excluded from the sentiment lexicons.

The number of occurring sentiment words in the data set is calculated for all three lexicons. The total number of occurring sentiment words is: stems 10338; lemmas 8463; 4-grams 22458. This distribution indicates that words cut to n-grams are better mapped in tweets, which is expected due to the number of different form of words which start with the same 4-gram. We checks how these numbers influence the sentiment with rule "whether the sentiment words appearing in tweets of the corresponding polarity". Number of words in different lexicons and number of sentiment words with the contradictory sentiment are shown in Table 1.

By classifying the sentiment words based on whether they appear more in positive or negative tweets, the effect of normalization on sentiment analysis is tested. Table 2 shows next parameters: **E2U: provjeriti**

- fp- the number of negative sentiment word which appear more in positive than in negative tweets.
- tp- the number of positive sentiment word which appear more in positive than in negative tweets.
- tn- the number of negative sentiment word which appear more in negative than in positive tweets.
- fn- ?.

Table 2: Results of sentiment words classification using different normalization

normalization	fp	tp	tn	fp	Acc
stemming	49	143	1501	611	71%
lematization	30	114	1121	523	69%
4-gram	59	155	2516	791	76%

- ACC- the obtained accuracy calculated on the formula:

$$Acc = \frac{tp + tn}{tp + tn + fp + fn} \tag{1}$$

From the obtained results, it can be concluded that the sentiment is the best joined to n-grams. The reason is that a larger number of n-grams were found in the set of tweets compared to stemmed words and lemmas. As informal texts, tweets often contain misspell that are rarer at length up to 4 letters. On the other hand, the Serbian language as a morphologically rich language is difficult to processing and a large number of words are found in forms that are not adequately processed by stemmer and lemmatizer, so such sentiment words cannot be found in the sentiments lexicon.

Results indicate that n-gram analysis brings a better accuracy in join the 4-gram with the corresponding polarity. Testing the improvement of the classification of sentiment words by cutting it to 4-gram versus stemmer and lemmatizer is done by applying MC Nemar's test**E2U: add reference**. We made a correlation matrix for classification by using n-gram analysis and lemmatization and n-gram analysis and stemming, and in both cases, it was found that the value of p <0.000, i.e. that the n-gram analysis statistically significantly influenced the improvement of the sentiment word classification.

EJUB STOPPED HERE

Given results should be carefully applied. The n-gram analysis is not a precise algorithm for normalization. Whenever it is possible, methods that are based on grammatical rules should be applied first. The results show that cutting off sentiment words on 4-gram gives good results in the polarization of the sentiment words in tweets, especially due to the informal form of writing in tweets. Therefore, we recommend the use of parallel resources. Language dependent resources should be applied first and then n-gram analysis so that mapping sentiment words in tweets should be improved.

5 Conclusion and future work

Sentiment analysis is significantly dependent on the use of sentence sentences. The use of such a lexical resource requires normalization of it as well as the text for which the sentiment is being examined. The obtained results indicated that with the use of different normalizers we obtain a different mapping of sentiment words in tweets. It has been shown that the n-gram analysis keeps the polarity sentiment words. We have proposed a use of parallel resources which would contribute to better mapping sentiment words in informal texts.

References

[1] Thar Baker, Emir Ugljanin, Noura Faci, Mohamed Sellami, Zakaria Maamar, and Ejub Kajan, *Everything as a resource: Foundations and illustration through internet-of-things*, Computers in Industry **94** (2018), 62–74.

[2] V. Batanovi, B. Nikoli, and M. Milosavljevi, *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) (Portoro, Slovenia), pp. 2688–2696.

[3] E. Chaniotakis, C. Antoniou, and F. C. C. Pereira, *Mapping Social Media for Transportation Studies*, IEEE Intelligent Systems **31** (2016), no. 6.

[4] Na F.F. da Silva, Eduardo R. Hruschka, and Estevam R. Hruschka, *Tweet sentiment analysis with classifier ensembles*, Decision Support Systems **66** (2014), 170 – 179.

[5] Claudia Diamantini, Alex Mircoli, Domenico Potena, and Emanuele Storti, *Social information discovery enhanced by sentiment analysis techniques*, Future Generation Computer Systems (2018).

[6] Maria Giatsoglou, Manolis G. Vozalis, Konstantinos Diamantaras, Athena Vakali, George Sarigiannidis, and Konstantinos Ch. Chatzisavvas, *Sentiment analysis leveraging emotions and word embeddings*, Expert Systems with Applications **69** (2017), 214 – 224.

[7] Minqing Hu and Bing Liu, *Mining and summarizing customer reviews*, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (New York, NY, USA), KDD '04, ACM, 2004, pp. 168–177.

[8] Svetlana Kiritchenko, Xiaodan Zhu, and Saif M. Mohammad, *Sentiment analysis of short informal texts*, J. Artif. Int. Res. **50** (2014), no. 1, 723–762.

[9] Cvetana Krstev, Gordana Pavlović-Lazetic, Duško Vitas, and Ivan Obradović, *Using textual and lexical resources in developing serbian wordnet*, Romanian Journal of Information Science and Technology **7(1-2)** (2004), 147–161.

[10] Bing Liu, *Sentiment analysis - mining opinions, sentiments, and emotions*, Cambridge University Press, 2015.

[11] Adela Ljajić, Ulfeta Marovac, and Aldina Avdić, *Processing of negation in sentiment analysis for the serbian language*, IcETRAN 2017 Conference proceedingsAt (Serbia), June 2017.

[12] U. Marovac, A. Pljaskovic, A. Crnisanin, and E. Kajan, *N-gram analysis of text documents in serbian language*, In Proceedings of the 20th Telecommunications Forum (TELFOR) (Belgrade, Serbia), 2012, pp. 1385–1388.

[13] George A. Miller and Christiane Fellbaum, *Wordnet then and now*, Language Resources and Evaluation **41** (2007), no. 2, 209–214.

[14] Nikola Milosevic, *Stemmer for serbian language*, arXiv preprint arXiv:1209.4471., 2012.

[15] Miljana Mladenović, *Sentiment in social networks as means of business improvement of higher education institutions.*

[16] A. Moreo, M. Romero, J.L. Castro, and J.M. Zurita, *Lexicon-based comments-oriented news sentiment analyzer system*, Expert Systems with Applications **39** (2012), no. 10, 9166 – 9180.

[17] Avinash Chandra Pandey, Dharmveer Singh Rajpoot, and Mukesh Saraswat, *Twitter sentiment analysis using hybrid cuckoo search method*, Information Processing and Management **53** (2017), no. 4, 764 – 779.

[18] Hassan Saif, Yulan He, Miriam Fernandez, and Harith Alani, *Contextual semantics for sentiment analysis of twitter*, Information Processing and Management **52** (2016), no. 1, 5 – 19, Emotion and Sentiment in Social and Expressive Media.

[19] Androniki Sapountzi and Kostas E. Psannis, *Social networking data analysis tools and challenges*, Future Generation Computer Systems (2016).

[20] Marcus Spies, *Towards an open software architecture for interleaved knowledge and natural language processing*, Scientific Publications of the State University of Novi Pazar Series A: Applied Mathematics, Informatics and mechanics **7** (2015), no. 1, 7–18.

[21] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly D. Voll, and Manfred Stede, *Lexicon-based methods for sentiment analysis*, Computational Linguistics **37** (2011), no. 2, 267–307.

[22] Duško Vitas and Cvetana Krstev, *Restructuring lemma in a dictionary of serbian*, Zbornik 7. mednarodne multikonference Informacijska druzba IS 2004 (Ljubljana, Slovenija), 2004.

[23] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann, *Recognizing contextual polarity in phrase-level sentiment analysis*, Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (Stroudsburg, PA, USA), HLT '05, Association for Computational Linguistics, 2005, pp. 347–354.