Review of the NCD 41 (2022), 58–73

Milena Šošić
Faculty of Mathematics, University of Belgrade

SRPOL – A LEXICON BASED FRAMEWORK FOR SENTIMENT STRENGTH OF SERBIAN TEXTS

Abstract. Determining the polarity of words is an important task in sentiment analysis and its applications. The most comprehensive dictionaries could be found in English, however, many other low-resource languages lack established polarity dictionaries,the existing ones are small in size or contain only polarity identifier. In this study, we propose a new lexicon-based approach for text polarity detection using sentiment triggers that add contextual semantics during the analysis. To this end, the existing word polarity dictionary in Serbian has been extended as to contain approximately 15000 words annotated with polarity strength. Serbian sentiment framework (SRPOL), relying on the new lexicon and proposed sentiment triggers, has shown an overall accuracy score of 79% on validation datasets from different domains annotated for sentiment, which is in the range with the state-of-the-art approaches on this task.

Keywords. Polarity, Strength, Lexicon, Serbian, SRPOL

1. Introduction

Identifying semantic orientation or polarity of words is one of the most important topics in sentiment analysis tasks. Most previous studies on word polarity detection have been carried on for English and make use of language-specific resources such as opinion corpus MPQA (Deng & Wiebe, 2015), Harvard General Inquirer with attached syntactic, semantic, and pragmatic information to part-of-speech (PoS) tagged words (Stone, Dunphy, & Smith, 1966), highly researched Affective Norms for English Words (ANEW) containing words rated by human subjects along three dimensions – valence, arousal, and dominance (Bradley & Lang, 1999) or SentiWords, the lexicon of word polarities derived from their contextual semantics (Gatti, Guerini, & Turchi, 2015). In all of these lexicons, words are associated with their positive or negative sentiment evoked by the word taken out of the context.

In applications, semantic lexical resources are often used as a baseline or as features for machine learning methods for sentiment analysis research (Liu & Zhang, 2012). The advantage of these approaches is that they do not require deep semantic analysis or word sense disambiguation to assign a polarity score to a word and are domain independent - consequently being less precise but more portable across domains.

Another valuable lexical resource that has been in use for more than two decades as the standard lexical database for English is Princeton WordNet - PWN (Fellbaum, 2005). As a result of automatically annotating all PWN lemma-PoS pairs sharing the same meaning, called synsets, a lexical resource for opinion mining SentiWordNet 3.0 is obtained based on their degrees of positivity, negativity, and neutrality (Baccianella, Esuli,

& Sebastiani, 2010). According to the design of SentiWordNet, each lemma-PoS pair can have more than one sense and thus can have different polarities. For that reason, usage of SentiWordNet word polarities is limited to the sense-disambiguated contexts.

Today, most sentiment analysis tools are based on machine learning approaches. In these models, words are represented as functions of the context in which they occur, and by utilising machine learning algorithms, the models are then able to match these generalised functions with tokens that have similar representations. This way they can detect the exact context and sentiment rating of a word with high accuracy (Barnes, Klinger, & Walde, 2017). This arguably gives the models an advantage over the approaches relying on the bag-of-words (BoW) principle, which does not take context into account. However, (Hutto & Gilbert, 2014) argue that machine learning models have several drawbacks, including the fact that they depend on the data sets upon which they are trained. Moreover, these models are much more computationally expensive concerning CPU processing and memory requirements, compared to models based on the BoW principle. This is not just a huge disadvantage when running analyses, but also makes this type of model less convenient for broad usage. Furthermore, a hybrid model incorporating machine learning algorithms in a model relying on the BoW principle and word lexicons has been shown to improve simple BoW models by a margin of 5% in the accuracy scores on binary classification tasks (Kolchyna, Souza, Treleaven, & Aste, 2015).

2. Existing word lexicons in Serbian

Even though some of the languages have their distributions of Princeton WordNet - PWN (Miller, 1995), they are in general not as comprehensive as PWN. Serbian WordNet - SWN (Krstev C. , Pavlović-Lažetić, Obradović, & Vitas, 2003), (Krstev, Pavlović-Lažetić, & Obradović, Using textual and lexical resources in developing serbian wordnet, 2004) has been developed within the BalkaNet -BWN (Tufis, Cristea, & Stamou, 2004) for a cluster of languages from the Balkans, and EuroWordNet - EWN (Vossen, 1998) for a cluster of European languages projects. Alignment between the languages has been achieved with the Inter-Lingual Index (ILI), established to connect similar contexts in different languages on the basis of PWN. The existence of the ILI connector has enabled the possibility for languages with WordNet distribution, such as Serbian, to assign SentiWordNet scores to lemma-PoS pairs belonging to the ILI-connected synsets. Limitations of assigning polarity scores to senses rather than to lemmas still remain for the polarity scores transferred in this way.

In Table 1, the five English corresponding synsets of the Serbian verb 'voleti' (eng. 'to love') present all possible senses in which it can occur. The difference in the score strengths is visible for the same lemma-PoS entry across the senses, including mixed scores (wish#v#1), where both positive and negative scores are assigned to the lemma-PoS (Mladenovic, 2016). Starting from a number of basic sentiment words, and by using a list of synonyms, authors in (Mladenović, Mitrović, Krstev, & Vitas, 2016) have proposed a semi-automated method for sentiment dictionary creation. Authors in (Ljajić & Marovac, 2019) create special sentiment dictionaries of words that appear in the scope of syntactic negations in the Serbian language and evaluate negation impact using an automatically translated Opinion Lexicon (Hu & Liu, 2004).

The Senti-Pol-sr (SentiPol.SR) lexicon is another valuable contribution to the field of Serbian sentiment analysis (Stankovic, Kosprdic, Ikonic-Nesic, & Radovic, 2022). The lexicon consists of 4188 words which have been semi-automatically rated with discrete polarity indicators of positive or negative sentiment. Even though it is mainly aimed toward sentiment analysis of Serbian novels from the period 1840–1920, SentiPol.SR has also shown to be useful within other domains such as movie reviews. Nonetheless, a thorough examination of the word list of the SentiPol.SR lexicon revealed several limitations. First, words do not have PoS tag assigned which makes it hard to determine the exact word form and the meaning. Next, words have polarity identifier assigned rather than polarity strength. Moreover, several words were found to be duplicated with the same or opposite scores, e.g. ’neverovatno’ (eng. ’incredibly’/’unbelievably’) has a score of +/-1 in the lexicon. Another matter of concern with SentiPol.SR is that it still has to be further validated and extended.

Name	Synset ID	Synonyms	Positive	Negative
love.v.01	1775164	love	0.5	0.0
love.v.02	1828736	love, enjoy	1.0	0.0
love.v.03	1775535	love	0.625	0.0
wish.v.01	1824339	wish	0.125	0.375
wish.v.02	1824736	wish, care, like	0.125	0.0

Table 1. Polarity scores for the word ‘voleti’ (eng.‘to love’) in the Serbian SentiWordNet v3 lexicon

All of the above facts emphasize the importance of developing a new efficient sentiment analysis tool in Serbian that would help researchers to benefit from the vast amount of textual data written in the Serbian language. In this paper, we will present a computational sentiment analysis framework for the Serbian language (SRPOL) that consists of:

- A word sentiment lexicon with a focus on both the polarity orientation and sentiment strength
- Methods to identify and take lexical properties such as modifiers and negations into account
- Algorithm to calculate the polarity strength of a text written in the Serbian language

The rest of the paper starts with the construction of the SRPOL framework including a new SentiWords.SR lexicon for Serbian words, exploration of sentiment triggers and overall algorithm for text sentiment strength measurement which are all presented in section 3. Methods for approach validation are described in section 4, followed by a discussion in section 5 and next planned research activities, presented in section 6 which concludes this paper.

3. Construction of SRPOL

3.1. SentiWords.SR Lexicon. SentiWords is a lexicon of English words with assigned polarity scores (Gatti, Guerini, & Turchi, 2015). Authors compare the most frequently used techniques in the research studies based on SentiWordNet with the newly suggested ones and blend them in an ensemble method, outperforming all the other SentiWordNet-based methods. As opposed to SentiWordNet, SentiWords assigns scores directly to words rather than to word senses. These scores, usually called prior polarities, denote polarities of the words independent of their context in the contrast to posterior polarities, dependent on the context in which word appears. Since it is derived from WordNet 3.0, this dictionary covers approximately 155,000 words, making it one of the most extensive word polarity dictionaries for English (see Table 2).

PoS	Positive	Negative	Neutral	Total
Noun	13287	14680	89692	117659
Adjective	5814	7598	7835	21247
Verb	3588	3052	4889	11529
Adverb	2271	540	1670	4481
Total	24957	25870	104086	154913

Table 2. Statistics of the polarity scores for words in the SentiWords lexicon. Words with a prior polarity score equal to 0 are marked as 'Neutral' in this table

Like most of the low-resource languages, Serbian does not have a lexicon such as SentiWords to obtain words polarity. For that reason, we have developed a semi-automated method (see Algorithm 1) to produce Serbian versions of the SentiWords dataset, called SentiWords.SR in the following text, with the help of an automatic Google translation tool.

The algorithm takes lemma-PoS entries from the English SentiWords lexicon and automatically translates all single-word lemma entries in the lexicon with the scores $\neq 0$, ending in 41843 lemma-PoS pairs in total. Translation pairs were evaluated for the correct translated Serbian word meaning and PoS tag assignment. During the evaluation, translations have been categorized into completely correct, partially correct - lemma translation or PoS tag assignment is wrong, and completely wrong - translation is missing or corrupted together with the PoS tag. All completely correct translations (20244 or 48.3%) and corrected versions of partially wrong translations (7528 or 18%) were assembled, while completely wrong translations were omitted (14071 or 33.6%).

Correct Serbian lemma$_{Sr}$-PoS combinations (27772 or 66.4%) have been assigned the same polarity score from the original lexicon as the corresponding English lemma-PoS. Finally, scores for each Serbian lemma$_{Sr}$-PoS were averaged and additional statistics such as the standard deviation of the scores and the number of the corresponding English lemmas have been calculated as additional checking points for the validation step. Final scores were evaluated by annotators who checked averaged polarity score strength and the orientation. During annotation, additional statistics such as the words with standard deviation >0 and number of corresponding English lemmas >1 were used as indicators of possible misleading during automatic processing e.g. word has different senses or it is found as a translation of the words from different senses. Established annotation schema considers word polarity values in the range [-1, +1], with the strongest sentiments closer to the range boundaries and weaker sentiments closer to 0.

After the mean sentiment score for each lemma was calculated, all lemmas with an accumulated sentiment score of 0 were omitted from the lexicon. Once the initial lexicon had been composed, it was compared to the list of words in the SentiPol.SR lexicon. The lemmas present in SentiPol.SR but missing in SRPOL, amounting to a total of 1281, were then gathered, inspected and re-annotated with the same coding schema. Once all the lemmas had been scored, the lexicon was assembled. To this end, SentiWords.SR contains 15073 unique lemma$_{Sr}$-PoS combinations, being able to detect 210.000 different inflected forms of Serbian words in total.

Algorithm 1: Creation of SentiWords.SR from the English SentiWords lexicon

```
FindPolarity
    inputs: $SentiWords; GoogleTranslate$
    output: $SentiWords.SR$
    foreach $lemma, PoS \in SentiWords$ do
        $score \leftarrow score(lemma, PoS)$;
        $lemmas_{Sr} \leftarrow clean(GoogleTranslate(lemma, PoS))$;
        $lemmas_{Sr}, PoS \leftarrow score(lemma, PoS)$;
        $SentiWords.SR \leftarrow evaluate(lemma_{Sr}, PoS, score)$;
    foreach $lemma_{Sr}, PoS \in SentiWords.SR$ do
        $score \leftarrow mean(lemma_{Sr}, PoS, score)$;
        $std \leftarrow std(lemma_{Sr}, PoS, score)$;
        $count \leftarrow count(lemma_{Sr}, PoS, lemma)$;
    return $SentiWords.SR$;
```

Algorithm 1. Creation of SentiWords.SR lexicon from the English SentiWords lexicon

As evident in Figure *1*, the lexicon has a slight bias towards lemmas with a positive sentiment (7803, corresponding to 51.7%) compared to the negatively annotated lemmas (7270, corresponding to 48.3%). Additionally, the majority of the lemmas are centered around 0, with approximately 57% of the lemmas carrying a sentiment score between -0.25 and 0.25. Conversely, it is observed that the lexicon contains slightly more highly negative lemmas than highly positive lemmas, with 6.2% of the lemmas having a score between -1 and -0.5, whereas positive scores between 0.5 and 1 account for 5.4% of the lemmas. Regarding the PoS distribution, the lexicon contains approximately 56% nouns, 29% adjectives, 12% verbs and 2% adverbs annotated for polarity strength.

3.2. Sentiment Triggers. Following the lexicon construction process, the polarity analysis tool has been expanded beyond the simple BoW technique. Several simple rules were constructed with an attempt to simulate some degree of context awareness:

- Adverb modifiers changing the perceived sentiment value of the following word in the text
- Negations reversing the polarity of the perceived sentiment
- Exclamation marks increasing the perceived sentiment of sentences

- Elongated words increasing the perceived sentiment of a word which is elongated
- Emojis and emoticons changing the perceived sentiment of a sentence in which they appear
- Text segmentation into meaningful morphological units such as sentences or part of the sentences

3.2.1. Adverb Modifiers. Modifiers are words such as ’vrlo’ (eng. ’very’), ’neverovatno’ (eng. ‘incredibly’) or ’zaista’ (eng. ’really’). Researchers in (Dragut & Fellbaum, 2014) have investigated the impact of adverbs in sentiment analysis, providing intensity modifying scores for several different modifiers. These words have a purpose to modify the polarity of an upcoming sentiment-laden word, but not to change its orientation.

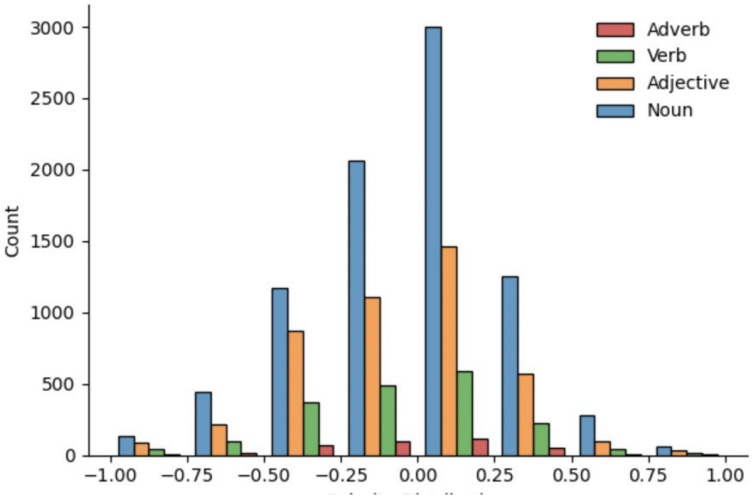

Figure 1. Word polarity distribution in the SentiWords.SR lexicon onthe bins of 0.25 lengthgrouped by PoS tag

Intensity modifiers proposed in (Dragut & Fellbaum, 2014) have been translated and incorporated into SRPOL by creating adverbs modifiers dictionary, while additional adverbs, such as ’veoma’ (eng. ’very’) and ’malo’ (eng. ’slightly’), specific for the usual use in the Serbian language, have been added. The decision to use translation of proposed modifiers was made based on the comparison of the usage of these words in two languages. Moreover, the translated intensity modifiers carry the same intensity score, whereas those added have been scored upon evaluation. The entire list of adverb modifiers and their intensity scores can be found in the Table *3*. Below is an example of how adverb modifier affects the polarity score of the containing phrase:

$$\textit{”Veoma (}\rightarrow \textit{MOD=1.2) dobar (p=+0.43) film...”} \xrightarrow{1.2\times(+0.43)} \textit{+0.52}$$

$$\textit{”Very (}\rightarrow \textit{MOD=1.2) good (p=+0.43) movie...”} \xrightarrow{1.2\times(+0.43)} \textit{+0.52}$$

In this case, the word ’dobar’ (eng. ’good’), with the score of +0.43, is multiplied by the intensity score (1.2) of the modifier ’veoma’ (eng. ’very’) to create a final score of the phrase equal to +0.52.

3.2.2. Negations. One of the most dominant flaws in the lexicon-based approach is the failure to deal with negations. Usually, the sentiment of individual words isused and averaged for the final sentiment score of the phrase, which could lead to the wrong sentiment assignments in the case of negated terms.

Adverb	Intensity
malo, neznatno, nesto (eng. slightly)	0.8
apsolutno (eng. absolutely)	1.2
veoma, vrlo (eng. very)	1.2
ekstremno (eng. extremely)	1.2
totalno (eng. totally)	1.2
neverovatno (eng. incredibly)	1.2
sasvim, stvarno, posteno, iskreno (eng. fairly)	1.2
pravicno, pravedno, zakonito (eng. rightful)	1.2
zaista (eng. really)	1.2
ozbiljno (eng. seriously)	1.4
strasno, uˇzasno (eng. awfully, horribly)ˇ	1.6

Table 3. Effect of adverbs on sentiment ratings

Serbian, as a morphologically rich language, recognizes lexical, morphological and syntactic negations. As elaborated by (Ljajić & Marovac, 2019), syntactic negations like the Serbian words ’ne’ or ’ni’ (eng. ’not’) tend to reverse the score of the upcoming sentiment-laden word. Moreover, the authors have analysed the effect and the scope of the syntactic negations on the part or the whole sentence sentiment by establishing a set of rules. Lexical negation is related to the use of a word whose meaning has a negative orientation e.g. ’mrznja’ (eng. ’hate’). Morphological negation is accomplished by using prefixesˇ such as “ne-” (non-), “bez-” (no-), “ni-” (not-), “a-” (a-), “dis-” (dis-) and “in-” (in-). Both lexical and morphological negations in SRPOL were processed throughSentiWords.SR lexicon construction by assigning appropriate polarity sign and the strength during annotation. For syntactic negations, SRPOL is following the structure proposed by (Ljajić & Marovac, 2019) for negated signals, which are replaced by the special token ’NEG’, and negation modifiers which are incorporated into the SRPOL modifiers list.

Specifically, SRPOL considers negations not only for the upcoming sentiment-laden word or by a set of established rules for the negation scope, but for the upcoming phrase which could include adverb and negation modifiers in addition to the first upcoming standard sentiment-laden word. When using SRPOL, the function will detect the special token ’NEG’ and automatically reverse the scores of the following word or phrase. In the simplest case the function works as it follows:

”Film nije (→ NEG) zanimljiv (p=+0.53)” $\xrightarrow[\times(-1)]{+0.53}$ *-0.53*

”The movie is not (→ NEG) interesting (p=+0.53)” $\xrightarrow[\times(-1)]{+0.53}$ *-0.53*

This example shows how the word ’nije’ (eng. ‘not’), which belongs to the negation signals list and is treated as a negation token, reversesthe score of the next word ’zanimljiv’ (eng. ’interesting’) (p=+0.53) to create a final score of -1∗0.53=-0.53.

In the combination with adverb modifiers, negation signals are reciprocally reversing intensity of the following adverb modifier, multiply it with the polarity of the sentiment-laden word and reverse it for the final polarity score of the phrase:

$$\textit{”Nije (→ NEG) mnogo (→ MOD=1.2) lose(p=-0.45)...”} \xrightarrow[\times(-1)]{-\frac{0.45}{1.2}} \textit{+0.37}$$

$$\textit{”Not (→ NEG) very (→ MOD=1.2) bad(p=-0.45)...”} \xrightarrow[\times(-1)]{-\frac{0.45}{1.2}} \textit{+0.37}$$

Lang	Phrase	Polarity
Sr	”**Nije** (→ NEG) uradio(p=+0.2)...”	-0.2
En	”**Not** (→ NEG) done(p=+0.2)...”	
Sr	”**Nije** (→ NEG) **zaista**→ MOD=1.2) uradio(p=+0.2)...”	-0.17
En	”**Not** (→ NEG) **really**(→ MOD=1.2) done(p=+0.2)...”	
Sr	”**Niko**(→ MOD=1.2) **nije** (→ NEG) uradio(p=+0.2)...”	-0.24
En	”**Nobody** did [not (→ NEG)] do(p=+0.2)...”	
Sr	”**Niko**(→ MOD=1.2) **nikada**(→w MOD=1.2) **nije** (→ NEG) uradio(p=+0.2)...”	-0.29
En	”**Nobody** has **never**(→ MOD=1.2) [not (→ NEG)] done(p=+0.2)...”	

Table 4. Effect of negation signals in combination with adverb and negation modifiers on sentiment ratings

Negation modifiers have been used similarly, except that their intensities were not diminished with intensity reversal when used with the negation signals. Moreover, their intensities have been multiplied in the case of multiple negation modifiers in the sequence, making the phrases gradually comparable on their polarity intensities (see Table 4).

3.2.3. Exclamation Marks. When dealing with sentiment, some of the most modern tools have started to include the influence of exclamation marks. In the paper (Teh, Rayson, Pak, & Piao, 2015) authors researched the effect of exclamation marks on perceived sentiment. Their research shows that an exclamation mark increases the perceived sentiment by an average of 6% for one, and 18% for the sequence of more than two exclamation marks. These effects have been incorporated into the SRPOL algorithm, boosting the score of the phrase as it is shown in the following example:

$$\textit{”Odlican film!”} \xrightarrow[\times 1.06]{+0.57} \textit{+0.60}$$

$$\textit{”Excellent movie!”} \xrightarrow[\times 1.06]{+0.57} \textit{+0.60}$$

In this example, the sentiment score of the phrase (p=+0.57) is multiplied by the exclamation mark intensifier (i=1.06), creating a final score of +0.60.

3.2.4. Elongated Words. An elongated word is defined as a word that contains a repeating character or group of characters more than two times, for example, 'cjajnoooo' (eng. 'awesoooome'). Character repetition as a means of emphasizing a particular word has been explored and identified as an important feature in the sentiment analysis tasks (Mohammad, Kiritchenko, & Zhu, 2013).

$$\text{”}\textit{Tako dooooosadan (p=-0.24)…}\text{”} \xrightarrow[1.27]{\times 1.05^{chr(o)}} \text{”}\textit{Tako dosadan (p=-0.30)...}\text{”}$$

$$\text{”}\textit{So boooooring (p=-0.24)…}\text{”} \xrightarrow[1.27]{\times 1.05^{chr(o)}} \text{”}\textit{So boring(p=-0.30)...}\text{”}$$

In this example, the sentiment score of the word 'dosadan' (eng. 'boring) (p=-0.24) is multiplied by the letter intensifier (i=1.05) powered with the number of letter repetitions, creating a final score of the elongated word equal to -0.30.

3.2.5. Emoticons and Emojis. With the era of social networks and instant messaging, emoticons have become one of the most dominant ways to express sentiment. For that reason, many researchers have started analyzing the effect of emoticons on the overall sentiment that a user is expressing through the message.

We have included a range of most used emoticon character sets with assigned sentiment strength for the usage in the SRPOL framework. Emoticons were grouped into named categories with the sentiment range of [-0.5, +0.5]. For example, emoticons such as '<3' or '♡' from the category 'love' got assigned the highest positive polarity of +0.5, emoticon ':)' from the category 'smile' has a polarity score of +0.25, while emoticon ':'(' from the category 'crying' has the lowest assigned polarity score of -0.5.

In the recent period, emojis are taking precedence over emoticons in social media messaging. Research on the interpretation of emoji is constantly growing exploring different functions and use of emoji (Bai, Dan, Mu, & Yang, 2019). For example, (Ljubešić & Fišer, 2016) has explored the influence of national development indicators on emoji usage, showing, among other facts, that in Eastern Europe emojis are used in an emotionally clear way. Researches in (Kralj Novak, Smailović, Sluban, & Mozetič, 2015) has also analysed the sentiment of emojis by manually annotating 70,000 tweets written in 13 European languages including Serbian. Their work has resulted in the Emoji Sentiment Ranking (ESR) lexicon consisting of 751 emoji characters with their corresponding sentiment distribution. ESR lexicon has been incorporated in the SRPOL framework in the original form of emoji sentiment assignments and used in sync with the SentiWords.SR polarity lexicon as it is presented in the following example:

”Divan (p=+0.4) film (p=+0.14) 😍 (p=+0.678)” →+0.41
”Lovely (p=+0.4) movie (p=+0.14) 😍 (p=+0.678)” →+0.41

In this example, the sentiment score of the word 'divan' (eng. 'lovely') (p=+0.4), the word 'film' (eng. 'movie')(p=+0.14) and emoji 😍 (p=+0.678) were averaged, creating a final score of the phrase word equal to +0.41.

3.2.6. Segmentation. The primary goal of splitting text into segments is to help in improving the polarity scoring for the long text with mixed sentiments detected on the containing segments. Splitting the text into segments can be performed in several different ways and it represents a field of wide range of research studies (Pak & Teh, 2018). In the sentiment analysis task, authors in (Hutto & Gilbert, 2014) use the contrastive conjunction 'but' as a signal shift in sentiment polarity, with the sentiment of the text following the conjunction being dominant and dictating the overall polarity rating. In this work, we assess the polarity score for each sentence of a particular text and using a majority rule approach we predict a new sentiment score. The majority rule is based on the segment polarity sign (see Equation 1),

$$P_{text} = \frac{\sum_i^S w_i * P_s^i}{\sum_i^S w_i}, w_i = \sum_m^S \left|sign(P_s^i) = sign(P_s^m)\right| \qquad (1)$$

$$P_s = \frac{\sum_i^k P_w^i}{k} \qquad (2)$$

where S is the number of sentences in the text, P_i is the polarity of the i-th sentence and w_i is the weight factor of the i-th sentence representing the number of occurrences of sentence polarity sign across the sentences. In this way, we are able to measure the polarity of the longer texts with quite opposite polarity scores retrieved in its segments. Sentence polarity score is the sum of word polarity scores divided by the number of words contributing to the score (see Equation 2).

3.3. Algorithm. In the application, SRPOL takes a text and splits it into sentences. Sentences have been identified by using the nltk package's PunktSentenceTokenizer having for endpoints punctuation marks such as '.', '...', '?', '!' and emoticon signs. Each sentence is then tokenized with RegexTokenizer and stop words have been removed. Stop words list is a carefully constructed list of words containing mostly words which do not bring any sentiment to the text such as prepositions and conjunctions. Any word which contains a negation signal or modifier has been excluded from the stop words list as their presence is important for sentiment triggers. Stop words removal is not mandatory in this approach but is a valuable additional check point for the words which will be further processed through the flow.

Upon tokenization and cleaning, tokens are then lemmatized and tagged for PoS tags with the help of the Serbian language lemmatizer and tagger models (Stanković, Šandrih, Krstev, Utvić, & Skoric, 2020). Consistent with the BoW technique, each lemma-PoS is then matched against the SRPOL sentiment lexicon SentiWords.SR, and if the lemma-PoS is present, the corresponding score is added to the sentence scores assignments. If a negation appears, the next assignment score is reversed. If one of the words is in the SRPOL modifier list (see 3.2.1), the following word is multiplied by the corresponding intensifier. In more complex cases, with both negations and modifiers, a set of modifying rules has been applied (see 3.2.2). Similarly, if the text includes an exclamation mark, the sentence score is multiplied by 1.06 (<2) or 1.18 (>=2) depending on the number of exclamation marks. Elongated words have been replaced with the standard form of the word it is derived from and the score is multiplied by1.05 powered by the number of repeated letters in the elongated word. Finally, the polarity of the text has been calculated as a weighted average of the polarity assignments on the identified sentences inside the text which is limited to the [-1, +1] range (see 3.2.6).

4. Method validation

4.1. Sentiment annotated corpora. To validate SRPOL, a corpus of sentiment-rated texts is needed to compare the human annotators sentiment scores with the SRPOL generated scores. For that reason, two different corpora from different domains were constructed simultaneously with the creation of SRPOL. The first corpus consists of social media messages written in the Serbian language extracted from pre-selected Twitter accounts for the period of one day (Tweets.SR) containing 7668 messages. The second corpus is a collection of randomly selected sentences from the newspaper texts of Leipzig Corpora Collection in Serbian language (Serbian-News, 2019). From that collection, we have selected sentences from the portal 'rts.rs' – the main TV station and news portal in Serbia, publishing news from wide range of topics such as politics, culture or sport (RTS.SR) which are written in the standard form of contemporary Serbian language. Total number of sentences in RTS.SR corpus is 7197. Both corpora have been annotated for positive, negative and neutral sentiment by three annotators of different age, gender and occupation.

The third corpus consists of user-generated reviews which are widely used for testing sentiment analysis tools as a measure of how well a tool can predict human sentiment. For the analysis of SRPOL, 3490 movie reviews were taken from the main SentiComments.SR dataset (Batanović, Cvetanović, & Nikolić, 2020) with sentiment annotation scheme containing six sentiment labels of +/-1 denoting predominantly positive/negative sentiment, +/-M denoting an ambiguous sentiment or a mixture of sentiments, but leaning more towards the positive/negative sentiment and +/-NS denoting non-sentiment, but still leaning more towards the positive/negative sentiment.

The SentiComments.SR was then reorganised into the following novel test corpora:

- The Corpus I includes all reviews which are categorized into positive and negative reviews based on the label sign. This corpus adds up to 3490 reviews.
- The Corpus II includes +/-1 and +/-M reviews, which are categorized into positive and negative reviews (-1 & -M = negative, +1 & +M = positive). This corpus adds up to 2871 reviews.
- The Corpus III includes +/-1 and +/-NS reviews, which are categorized into positive and negative reviews (-1 & -NS = negative, +1 & +NS = positive). This corpus adds up to 2876 reviews.
- The Corpus IV includes +/-1 reviews, which are categorized into positive and negative reviews (-1 = negative, +1 = positive). This corpus adds up to 2257 reviews.
- The Corpus V includes all reviews, which are categorized into positive, negative and neutral reviews (-1 & -M = negative, +1 & +M = positive, +/-NS = neutral).

In the statistical testing of the tool, we use SentiPol.SR as a benchmark to provide exhaustive validation of the SRPOL results.

4.2. Results Evaluation. SRPOL was tested against SentiPol.SR on the ability to predict the polarity of texts using logistic regression (LR) as a predictive analysis method. This type of binary classification task is a standard way of testing sentiment analysis tools (Hu & Liu, 2004). We interpreted the polarity of the texts as a binary dependent variable and then used SRPOL's and SentiPol.SR's sentiment scores of the text as predictor variables to create a logistic regression model for each lexicon. Both the SRPOL and the SentiPol.SR models were trained on 80% of the corpora and then subsequently tested on the remaining 20%. The underlying models used to test the accuracy of both lexicons were to compare the polarity of the text with the mean polarity score of the words contained in the text.

In both SRPOL and SentiPol.SR models, change in the mean sentiment score of a text hasan effect in predicting the polarity: χ^2 = 813.5, p <.001 and 339.8, p <.001 respectively. A positive change in the mean sentiment score had a significant positive impact on predicting a positive sentiment with coefficients of 8.15 and 0.95 in these models. The results suggest that both models predict the polarity of the text significantly. However, they differ in how much variance they explain with a difference between the Chi-squared values of the models of 473.7, indicating that SRPOL is performing better (see Table 5).

Model	χ_2	coef	Std Err	z	P> |z|	Odds Ratio
SRPOL	813.5	8.15	0.03	28.5	0.000	3464.2
SentiPol.SR	339.8	0.95	0.05	18.4	0.000	2.58

Table 5. Model and predictor significance in the LR polarity prediction model

SRPOL and SentiPol.SR were further evaluated based on their performance in accurately categorizing the 20% of the data that were not included in the model training as either positive or negative. The logistic regression models were applied to the untrained data to provide probability scores. If the probability was above 50% the text was classified as positive, whereas if it was below 50% it was classified as negative. On average SRPOL correctly predicted 78.8% of the texts to be either positive or negative, whereas SentiPol.SR on average correctly predicted 70.4% of the texts across different domains (see Table 6).

		SRPOL					SentiPol.SR				
Dataset		Accuracy	F1	P	R	Spearman	Accuracy	F1	P	R	Spearman
Tweets.SR		80.5%	76%	76%	77%	0.56	71.2%	68%	67%	70%	0.30
RTS.SR		81.3%	79%	81%	78%	0.58	75.2%	73%	75%	73%	0.42
SentiComments.SR	Corpus I	76.2%	70%	72%	70%	0.46	67.1%	57%	64%	58%	0.29
	Corpus II	75.8%	71%	73%	71%	0.49	66.8%	58%	64%	59%	0.31
	Corpus III	78.2%	72%	75%	71%	0.48	70.5%	61%	67%	61%	0.33
	Corpus IV	80.5%	77%	79%	76%	0.53	71.3%	62%	68%	61%	0.38
Total Avg		78.8%					70.4%				

Table 6. Comparison of binary sentiment classification (positive/negative) accuracy and correlation scores of SRPOL and SentiPol.SR models on datasets from different domains

Results presented in Table 6 and Table 7 have revealed several valuable facts. In the binary classification task, SRPOL shows consistent accuracy score of above 80% in all datasets with the clear sentiment connotation. Especially, in the SentiComments.SR dataset, SRPOL results are in the range of the results reported by the authors for the binary classification task (Batanović, Cvetanović, & Nikolić, 2020). Our results outperform reported results on the BERT embedding models, such as distil-BERT or multilingual-BERT, while still being below the standard BoW technique and XLM-BERT model. SentiComments.SR is user generated informal language dataset containing non-standard word forms or word derivations from English language, especially for the movies or character names, where models such as cross-lingual XLM-BERT or BoW could take their advantage over other models.

A Spearman rank-order correlation test was conducted, as this is another widely used method in validating sentiment analysis tools (Hutto & Gilbert, 2014). The correlation was computed between the human annotated polarity score and the mean sentiment scores assigned by SRPOL and SentiPol.SR models. According to the conventional approach for the interpretation of correlation coefficients suggested by authors in (Schober, Boer, & Schwarte, 2018), SRPOL, with the range of scores 0.46-0.58 in 2-class and 0.50-0.52 in 3-class classification, is in the area of moderate correlation and outperforms SentiPol.SR on obtained correlation coefficients in all domains. Authors in (Hutto & Gilbert, 2014) conducted similar correlation tests on social media, NY Times Editorials and movie reviews datasets, which are highly comparable to the Tweets.SR, RTS.SR and SentiComments.SR datasets respectively, used in this research. The correlation test was conducted on eight most prominent word polarity tools, including GI (Stone, Dunphy, & Smith, 1966), ANEW (Bradley & Lang, 1999) and SentiWordNet (Baccianella, Esuli, & Sebastiani, 2010). The comparison shows that SRPOL is in the range of the most prominent English tools in three different domains.

		SRPOL					SentiPol.SR				
Dataset		Accuracy	F1	P	R	Spearman	Accuaracy	F1	P	R	Spearman
Tweets.SR		59.1%	43%	50%	44%	0.51	57.9%	40%	39%	43%	0.23
RTS.SR		59.7%	48%	53%	46%	0.52	57.2%	39%	45%	42%	0.31
SentiComments.SR	Corpus V	60.1%	49%	56%	51%	0.50	53.1%	35%	34%	40%	0.32
Total Avg		59.6%					56.1%				

Table 7. Comparison of 3-class sentiment classification (positive/negative/neutral) accuracy and correlation scores of SRPOL and SentiPol.SR models on datasets from different domains

5. Discussion

SRPOL outperforms SentiPol.SR in all tests performed, while also obtains a correlation score that is in the range with prominent English tools in comparable tests on datasets from different domains.

It is a clear indication that SRPOL is domain independent and highly applicable across different domains. By utilizing features from different existing English sentiment tools and related Serbian language studies such as lemmatized word list constructed by (Stanković, Šandrih, Krstev, Utvić, & Skoric, 2020), SRPOL has proven to be promising Serbian tool for catching the sentiment. The tool differs from the traditional BoW technique by incorporating functions that handle negations, adverb modifiers, exclamation marks, elongated words, emoticons and text segmentation. The construction of SRPOL reflects the complexity of language, as sentiment is not only found in words themselves but also to a high degree in which context they appear. Even when taking the latter into account by employing a degree of context awareness, SRPOL is far from perfect, showing how hard it is to precisely capture the sentiment of a text.

Although SRPOL incorporates context sensitive measures such as negations, adverb modifiers and segmentation, the main methodological limitations arise from the assumptions following the BoW technique. By analysing words in a text individually this technique potentially compromises aspects of meaning as it fails to take into account the context in which the words appear. SRPOL also has a problem with interpretation of statements expressing emotions metaphorically, sarcastically or ironically. In addition, similar problem is encountered with the interpretation of homographs.

With lemmatization, it is important to note that it relies on the assumption that all inflections of a word contain the same sentiment. This can be a problem when handling inflected adjectives such as the word ’loš’ (eng. ’bad’) that receives the same sentiment score for all inflections, even though the inflected superlative degrees of ’gori’ (eng. ’worse’) and ’najgori’ (eng. ’worst’) have relatively more negative connotations. In addition, SRPOL highly depends on the lemmatization and PoS tagger models accuracy. Another limitation arises in the validation of SRPOL which relies on the congruency between the text and annotation scheme used as it is observed in the results obtained for the SentiComments.SR dataset.

6. Conclusion and future work

The main goal of this paper was to fill the observed gap of missing tools within the field of Serbian sentiment analysis. A framework named SRPOL, including a new word polarity lexicon and algorithm to derive sentiment from the context, was constructed and validated in comparison with SentiPol.SR. While both models achieved significant results, SRPOL outperformed SentiPol.SR in all tests performed. Likewise, a comparison of validation results with prominent English tools further ensures the competence of SRPOL.Future work on the framework is considered to enrich the number of words in the SentiWords.SR lexicon by using advanced machine learning methods to achieve that goal. Even more, incorporation of additional sentiment triggers and evaluation of different segmentation techniques are among the next planned research activities.

SRPOL has been made with the aim of providing an accessible tool without the usage of advanced machine learning approach. Even with the excellent results achieved by machine learning models, there is still constant trade-off between computational efficiency and accuracy which should be considered. With SRPOL providing significant results in different domains, the tool could be valuable in the areas like social media behaviour analysis or market research. Moreover, it can be used within a broad range of Serbian language research studies, following remarkable results obtained by using tools developed for English language.

References

1. Stone, Philip J., Dexter C. Dunphy, and Marshall S. Smith. 1966. "The general inquirer: A computer approach to content analysis." (MIT press).
2. Kralj Novak, Petra, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. "Emoji sentiment ranking 1.0." (Jožef Stefan Institute).
3. Mohammad, Saif M., Svetlana Kiritchenko, and Xiaodan Zhu. 2013. "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets." *arXiv preprint arXiv:1308.6242.*
4. Stanković, Ranka, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Skoric. 2020. "Machine learning and deep neural network-based lemmatization and morphosyntactic tagging for serbian." *Proceedings of The 12th Language Resources and Evaluation Conference.* 3954–3962.
5. Fellbaum, Christiane. 2005. "WordNet and wordnets."
6. Dragut, Eduard, and Christiane Fellbaum. 2014. "The role of adverbs in sentiment analysis." *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore (1929-2014).* 38–41.
7. Batanović, Vuk, Miloš Cvetanović, and Boško Nikolić. 2020. "A versatile framework for resource-limited sentiment articulation, annotation, and analysis of short texts." *PLoS One* (Public Library of Science San Francisco, CA USA) 15: e0242050.
8. Ljajić, Adela, and Ulfeta Marovac. 2019. "Improving sentiment analysis for twitter data by handling negation rules in the Serbian language." *Computer Science and Information Systems* 16: 289–311.
9. Bradley, Margaret M., and Peter J. Lang. 1999. "Affective norms for English words (ANEW): Instruction manual and affective ratings." Tech. rep., Technical report C-1, the center for research in psychophysiology ….
10. Tufis, Dan, Dan Cristea, and Sofia Stamou. 2004. "BalkaNet: Aims, methods, results and perspectives. a general overview." *Romanian Journal of Information science and technology* (Citeseer) 7: 9–43.
11. Serbian-News, Leipzig Corpora Collection. 2019. "Serbian news corpus based on material from 2019." *Serbian news corpus based on material from 2019.*
12. Stankovic, Ranka, M. Kosprdic, M. Ikonic-Nesic, and Tijana Radovic. 2022. "Sentiment Analysis of Sentences from Serbian ELTeC corpus." *Proceedings of the SALLD-2 Workshop at Language Resources and Evaluation Conference (LREC), Marseille, France.* 31–38.
13. Barnes, Jeremy, Roman Klinger, and Sabine Schulte im Walde. 2017. "Assessing state-of-the-art sentiment models on state-of-the-art sentiment datasets." *arXiv preprint arXiv:1709.04219.*
14. Liu, Bing, and Lei Zhang. 2012. "A survey of opinion mining and sentiment analysis." In *Mining text data*, 415–463. Springer.
15. Gatti, Lorenzo, Marco Guerini, and Marco Turchi. 2015. "SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis." *IEEE Transactions on Affective Computing* (IEEE) 7: 409–421.
16. Ljubešić, Nikola, and Darja Fišer. 2016. "A global analysis of emoji usage." *Proceedings of the 10th web as corpus workshop.* 82–89.

17. Krstev, Cvetana, Gordana Pavlović-Lažetić, Ivan Obradović, and Duško Vitas. 2003. "Corpora issues in validation of Serbian WordNet." *International Conference on Text, Speech and Dialogue*. 132–137.
18. Schober, Patrick, Christa Boer, and Lothar A. Schwarte. 2018. "Correlation coefficients: appropriate use and interpretation." *Anesthesia & Analgesia* (Wolters Kluwer) 126: 1763–1768.
19. Teh, Phoey Lee, Paul Rayson, Irina Pak, and Scott Piao. 2015. "Sentiment analysis tools should take account of the number of exclamation marks!!!" *Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services*. 1–6.
20. Hu, Minqing, and Bing Liu. 2004. "Mining and summarizing customer reviews." *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 168–177.
21. Deng, Lingjia, and Janyce Wiebe. 2015. "Mpqa 3.0: An entity/event-level sentiment corpus." *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*. 1323–1328.
22. Miller, George A. 1995. "WordNet: a lexical database for English." *Communications of the ACM* (ACM New York, NY, USA) 38: 39–41.
23. Mladenović, Miljana, Jelena Mitrović, Cvetana Krstev, and Duško Vitas. 2016. "Hybrid sentiment analysis framework for a morphologically rich language." *Journal of Intelligent Information Systems* (Springer) 46: 599–620.
24. Vossen, Piek. 1998. "A multilingual database with lexical semantic networks." *Dordrecht: Kluwer Academic Publishers. doi* (Springer) 10: 978–94.
25. Mladenovic, Miljana. 2016. "Informatički modeli u analizi osećanja zasnovani na jezičkim resursima." *Univerzitet u Beogradu* (Univerzitet u Beogradu, Matematički Fakultet).
26. Kolchyna, Olga, Tharsis T. P. Souza, Philip Treleaven, and Tomaso Aste. 2015. "Twitter sentiment analysis: Lexicon method, machine learning method and their combination." *arXiv preprint arXiv:1507.00955.*
27. Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. 2010. "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining." *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).*
28. Bai, Qiyu, Qi Dan, Zhe Mu, and Maokun Yang. 2019. "A systematic review of emoji: Current research and future perspectives." *Frontiers in Psychology*(Frontiers Media SA) 10: 2221.
29. Hutto, Clayton, and Eric Gilbert. 2014. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." *Proceedings of the international AAAI conference on web and social media.* 216–225.
30. Pak, Irina, and Phoey Lee Teh. 2018. "Text segmentation techniques: a critical review." *Innovative Computing, Optimization and Its Applications* (Springer) 167–181.
31. Krstev, Cvetana, Gordana Pavlović-Lažetić, and I. Obradović. 2004. "Using textual and lexical resources in developing Serbianwordnet." *Romanian Journal of Information Science and Technology* 7: 147–161.

milena.sosic@gmail.com