Annotated Lexicon for Sentiment Analysis in the Bosnian Language

Sead JAHIĆ

Faculty of Mathematics, Natural Science and Information Technologies, University of Primorska

Jernej VIČIČ

Faculty of Mathematics, Natural Science and Information Technologies, University of Primorska; Fran Ramovš Institute of the Slovenian Language

The paper presents the first sentiment-annotated lexicon of the Bosnian language. The annotation process and methodology are presented along with a usability study, which concentrates on language coverage. The composition of the starting base was done by translating the Slovenian annotated lexicon and later manually checking the translations and annotations. The language coverage was observed using two reference corpora. The Bosnian language is still considered a low-resource language. A reference corpus comprised of automatically crawled web pages is available for the Bosnian language, but the authors had a hard time sourcing any corpora with a clear time frame for the text contained therein. A corpus of contemporary texts was constructed by collecting news articles from several Bosnian web portals. Two language coverage methods were used in this experiment. The first used a frequency list of all words extracted from two reference Bosnian language corpora, and the second ignored the frequencies as the main factor in counting. The computed coverage using the first presented method for the first corpus was 19.24%, while the second corpus yielded 28.05%. The second method yielded 2.34% coverage for the first corpus and 6.98% for the second corpus. The results of the study present a language coverage that is comparable to the state of the

Jahić, S. et al.: Annotated Lexicon for Sentiment Analysis in the Bosnian Language. Slovenščina 2.0, 11(2): 59–83.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2023.2.59-83

art in the field. The usability of the lexicon was already proven in a Twitter-based comparison.

Keywords: Bosnian lexicon, corpus, sentiment analysis, AnAwords, stopwords, log-likelihood, annotation

1 Introduction

Sentiment analysis, also known as opinion mining, is a field of study within the larger discipline of natural language processing (NLP) that aims to determine the sentiment expressed in text, categorizing it as positive, negative, or neutral. The goal of sentiment analysis is to extract meaningful insights from large amounts of unstructured data, such as social media posts (Iglesias and Moreno, 2019) or online product reviews (Wu et al., 2020), in order to understand public opinions and attitudes. In this paper we present the coverage of the first Bosnian sentiment annotated lexicon using two reference corpora.

Although arguably Bosnian is closely related to Serbian and Croatian, there are subtle differences between these three languages that are more evident from the sentiment analysis point of view. The main differences between Bosnian, Serbian, and Croatian lie in the use of vocabulary, grammar, and syntax. Although the three languages share a similar Slavic origin and linguistic heritage, they have evolved differently over time and been influenced by different cultural, historical, and political factors. These differences are particularly pronounced when it comes to sentiment analysis, as the choice of words and the way they are used can significantly impact the sentiment expressed in a text. As such, it is important to consider these differences when developing sentiment analysis tools for the Bosnian language.

The lexicon used in this study has been constructed using two reference corpora and combines NLP and machine learning techniques to assign weighted sentiment scores to the entities within a sentence or phrase. The study covers two approaches to evaluate the performance of the lexicon – the first takes into account the frequencies of the covered and missed words, while the second just counts the words that are covered by the lexicon.

The paper provides a comprehensive overview of the state of the art in NLP and sentiment analysis for the Bosnian language. It explains the methodology used in the process of creating the lexicon, cleaning the corpora, the corpora covered by the lexicon, and annotation. The results of the experiment and the conclusion, along with suggestions for future work, are presented in the last section of the paper.

In summary, the development of the Bosnian sentiment annotated lexicon is a step towards better understanding and analysing public opinion expressed in the Bosnian language. The results of the study suggest that the lexicon has good coverage, and the methodology used in the construction of the lexicon can serve as a reference for future work in this field.

2 State of the art

There has been quite extensive research in the area of sentiment analysis, and many types of models and algorithms have been proposed depending on the final goal of the analysis of the interpretation of user feedback and queries, such as fine-grained sentiment analysis (based on polarity precision) (Chen et al., 2020), emotion detection, aspect-based sentiment analysis (Suciati and Budi, 2020), and multilingual sentiment analysis (Kia et al., 2016). All those algorithms and models can be divided into one of three basic classes: rule-based systems (relying on long-used linguistic methods, rules, and annotated linguistic materials such as annotated lexicons), automatic (corpus-based) systems, and hybrid systems that combine properties from both previous types. In the latter, hybrid systems use machine learning techniques together with NLP techniques developed in computational linguistics, such as stemming, tokenization, part-of-speech tagging, parsing, and lexicons.

Lexicons have been widely used for sentiment analysis. One of the first-known, human-annotated lexicons for sentiment analysis is the General Inquirer lexicon (Hartman et al., 1967), which contains 11,788 English words (2,291 labeled as negative and 1,915 as positive, with the rest, labeled as objective).

Sentiment lexicons exist for most Slavic languages, including Bulgarian (Kapukaranov and Nakov, 2015), Croatian (Glavaš et al.,

2012), Czech (Veselovská, 2013), Macedonian (Jovanoski et al., 2015), Polish (Wawer, 2012), Slovak (Okruhlica, 2013), Slovenian (Kadunc, 2016) and Bosnian (Jahić and Vičič, 2023b), with the last of these containing 1,219 entries labeled as positive and 3,935 as negative.

Important questions for natural language researchers, general linguists, and even teachers and students are how much text coverage can be achieved with a certain number of words from the lexicon in a given language, since the number of terms in the lexicon is smaller by a few magnitudes than the number of terms in the corpus.

Studies of vocabulary coverage have been carried out for many languages, such as German (Jones, 2006), where a study based on the BYU/Leipzig Corpus of Contemporary German has shown that a basic vocabulary of 3,000 high – frequency words can account for between 75% and 90% of the words in the text. Moreover, with Spanish (Davies, 2005) it is claimed that it is enough to know 4,000 words to cover or recognize more than 90% of the words in native texts. Moreno-Ortiz and Pérez-Hernández (2018) presented Lingmotif-lex, a wide-coverage, domain-neutral lexicon for sentiment analysis in English, and stated that it achieves significantly better performance than the other lexicons for English, with coverage of up to 75% and 84% (F1-score) for two datasets.

In a study aimed at developing resources for sentiment analysis in Slovene, Bučar, Žnidaršič and Povh (2018) collected more than 250,000 news items from five Slovenian online media sources as the basis for their resources, which corpora, annotations, and a lexicon. To evaluate the quality of the annotation process, they used five different measures of correlation. The results showed good internal consistency across all levels of granularity, although the values decreased slightly when applied to the smaller units of text.

Corpus-based and lexicon-based model methods have been increasingly used to compare language coverage, and the comparison of hundreds of thousands or even millions of words/lemmas from a corpus with a few thousand words/lemmas from a lexicon is one of the main types of corpus comparison.

3 Construction of the lexicon

The Bosnian sentiment-annotated lexicon is presented and analysed in this paper. For this purpose, our data consists of the "core" lexicon (Jahić and Vičič, 2023b), a list of stopwords, and a list of AnAwords (Affirmative and Non-affirmative words, such as "ekstremno" ("vrlo") – extremely, "jedva" – barely) (Jahić and Vičič, 2023a), as clarified in Figure 1.

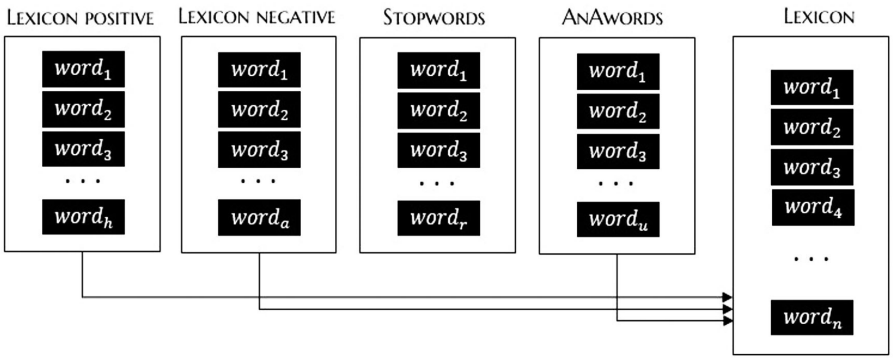

Figure 1: Construction of the lexicon.

The lexicon creation process is comprised of taking entries (word forms) from the Slovene sentiment lexicon KSS 1.1 (Kadunc, 2016) and translating them into Bosnian. We also allow some variance of the same lemma as part of the lexicon.The process of creating the Bosnian translation was undertaken in a dual-phase approach. In the initial phase, the transformation of the Slovenian lexicon into Bosnian took place through well-defined steps. Initially, the Slovenian lexicon underwent translation into English through the utilization of the translators from Google and Microsoft. Subsequently, this intermediary English version was subjected to translation into the Bosnian language, which is visually depicted in Figure 2. Moreover, the first phase involved these steps:

- Translation using Microsoft Translator for the Slovene sentiment lexicon KSS 1.1.
- Translation using Google Translator for the same lexicon.
- Manual comparison and merging of the two lists, removing duplicate entries.

- Manual cross-checking to ensure that words had matching or similar meanings.
- The result was the creation of the Bosnian_MG_Translated Lexicon.

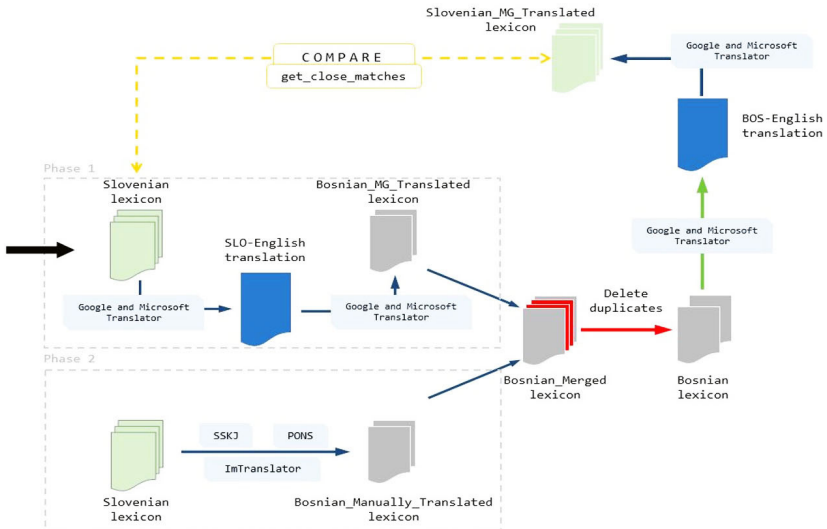

Figure 2: The lexicon creation process.

The last phase witnessed the creation of the lexicon in a two-fold manner. Firstly, word forms from the Slovenian lexicon were manually translated into Bosnian through manual means. This comprehensive process encompassed a verification of each term using various tools, including Pons,[1] Google Translate, ImTranslator,[2] and the Dictionary of Slovenian Literary Language (SSKJ – Slovar slovenskega knjižnega jezika[3]). The results of this process yielded the "Bosnian_Manually_Translated" lexicon.

These two lexicons ("Bosnian_MG_Translated" lexicon and "Bosnian_Manually_Translated" lexicon) were subsequently united and merged into a cohesive entity, referred to as the "Bosnian_Merged" lexicon. The refinement process further entailed the removal of duplicate entries, resulting in the initial iteration of the Bosnian sentiment

1 https://sl.pons.com/
2 https://imtranslator.net/
3 https://fran.si/

lexicon. To ensure the accuracy and robustness of the lexicon, a back--translation procedure was executed. This involved translating the newly constructed Bosnian lexicon back into the Slovenian language, as depicted in Figure 2.

The goal of the back-translation procedure was to retranslate the obtained Bosnian lexicon into a Slovenian lexicon and then compare this translated lexicon (in Slovenian) with the initial KSS 1.1 lexicon. What we found during the back-translation process is that many words were translated into a form that is not present in KSS 1.1, while the infinitive form of those words is indeed available in KSS 1.1.

To circumvent this challenge in the evaluation phase and also in the process of using the lexicon in the sentiment analysis process, we used the '*get _close_matches*' function (part of the *difflib* module in *Python*). By using this function we effectively pinpointed the closest approximations to the target string from a pool of candidate strings. This process substantially improved the coverage and reliability of our lexicon, amplifying the precision of our sentiment analysis efforts.

The method works by comparing the target string with each candidate string, using a defined similarity ratio, and then returning the matches with the highest similarity ratio. The number of matches returned and the similarity ratio threshold can be controlled through the *n* and *cutoff* parameters, respectively. The order of close-matched strings is based on the similarity score, so the most similar string comes first in the list.

This function accepts four parameters:

- *word*: This is the string for which we need the close matches.
- *possibilities*: This is usually a list of string values with which the word is matched.
- *n*: This is an optional parameter with a default value of 3. It specifies the maximum number of close matches required.
- *cutoff*: This is also an optional parameter with a default value of 0.6 . It specifies that the close matches should have a score greater than the cutoff.

In our case, we pick the first element from the close-matched strings list (with the highest similarity score). More cutoff values were considered, and the best confidence-accuracy score was reached with a cutoff of 81%.

Table 1*: Comparing the Slovenian lexicon before and after the translation process*

	Slovenian lexicon (lemmas) (Kadunc, 2016)	cutoff	Slovenian MG_Translated lexicon		Comparing accuracy
			translated words	matched words	
Positive	1911	80%	1758	1829	-
		81%	1781	1686	**88.23%**
		82.5%	1790	1627	85.14%
		85%	1806	1550	81.11%
		90%	1838	1369	71.64%
		100%	1858	1235	64.63%
Negative	5125	80%	4572	4999	-
		81%	4654	4604	**89.83%**
		82.5%	4690	4432	86.48%
		85%	4739	4125	80.49%
		90%	4846	3514	68.57%
		100%	4898	3067	59.84%

The accuracy score was counted by comparing the primary lexicon of the Slovenian language (Kadunc, 2016) and the back-translated lexicon of the Slovenian language.

$$\text{Comparing\_accuracy} = \frac{\text{Number of matched words (positive/negative)}}{\text{Number of all words in the lexicon (positive/negative)}}$$

The equation used is as follows:

The Bosnian sentiment lexicon consists of 3,935 negative words (Lexicon negative), and 1219 positive words (Lexicon positive). Besides that, we also added a list of 394 Bosnian stopwords (such as: “gosp.” (“Mr.”), “je” (“is”), “juli” (“July”), and so on), and list of AnAwords. Stopwords usually refer to the most common words in a language, and there is no single universal list of these. The first

Bosnian sentiment lexicon was tested by using it to label tweets written in the Bosnian language (Jahić and Vičič, 2023a, 2023c).

4 Methodology and work

The core emphasis of this paper is on assessing the coverage achieved by the lexicon, rather than on its creation, although a comprehensive account of this is also presented. More specifically, the focus of is on evaluating how many lemmas the lexicon covers in bsWac and bsNews, as detailed below.

The language coverage of the lexicon was evaluated through two different corpora:

- The Bosnian web corpus bsWaC 1.1 (Ljubešić and Klubička, 2014). The bsWaC 1.1 corpus was part of a collection of corpora, named the {bs, hr, sr}WaC – Web corpora of Bosnian, Croatian, and Serbian languages. The number of seed URLs (crawled web pages) was 8,388 for bsWaC, 11,427 for srWaC, and 14,396 for hrWaC.
  The bsWaC corpus consists of more than 285 million tokens (286,865,790, to be precise) written in Bosnian. The corpus is also morphosyntactically annotated and lemmatized. At the time of writing, this corpus was the *de facto* reference corpus for the Bosnian language.
- The Bosnian news corpus 2021 bsNews 1.0 (Vičič, 2021), which is a collection of web news articles crawled at the start of 2021. The corpus contains a balanced set of at most 2,000 of the most recent news articles from each identified web news portal in Bosnia and Herzegovina. The list of portals is maintained by Press Council in Bosnia and Herzegovina.[4]
  The corpus contains news articles from 46 portals. This corpus was used as a contemporary and balanced source. The sentence tokens are morpho-syntactically annotated with MULTEXT-East morpho-syntactic annotations for Croatian, Version 6[5]. The corpus was morpho-syntactically annotated and lemmatized with ToTaLe (Erjavec et al., 2015). It consists of more than 36 million tokens in the Bosnian language.

4 Vijeće za štampu u Bosni i Hercegovini: https://www.vzs.ba/index.php/vijece-za-stampu/internet-portali-u-bosni-i-hercegovini.

5 http://nl.ijs.si/ME/V6/

Two different approaches are applied:

- First, all lemmas with their frequencies were considered,
- Second, the frequencies for lemmas were ignored.

A list of lemmas with frequencies was extracted from each corpus and cut off at five occurrences to avoid clutter.

The list of lemmas extracted from the first corpus (Ljubešić and Klubička, 2014) consisted of 348,988 different lemmas with frequency. The lemmas are ordered in increasing order by frequency, where the lowest value is five (cutoff) ("batkovi - drumsticks" ...) and the highest value is 16,652,066 for the lemma "biti - to be".

The list of lemmas extracted from the second corpus (BsNews 1.0 corpus (Vičič, 2021)) consisted of 101,771 lemmas ordered in decreasing order, the most frequent lemma again being "biti – to be" with the frequency of 2,350,487, and with the lowest frequency of five for lemmas such as "polegnuti – lay down".

Not all lemmas can be included in the analysis. Symbols, equation marks, and numbers, even if part of the corpus, cannot be part of the lexicon, especially a sentiment annotated lexicon.

The following items were thus removed from both corpora in the cleaning process: emoticons, punctuation like quotes, exclamation marks, etc., numbers, and hyperlinks.

4.1 The first approach: lemmas with their frequency were included in the analysis (all appearances of lemmas were used for each corpus)

In the first approach used in our analysis, we considered lemmas along with their frequency as the basis for our investigation. This means that we included all instances of lemmas found in each corpus for our analysis.

Figure 3 shows the procedure for checking the existence of given words from the lexicon in the corpus.

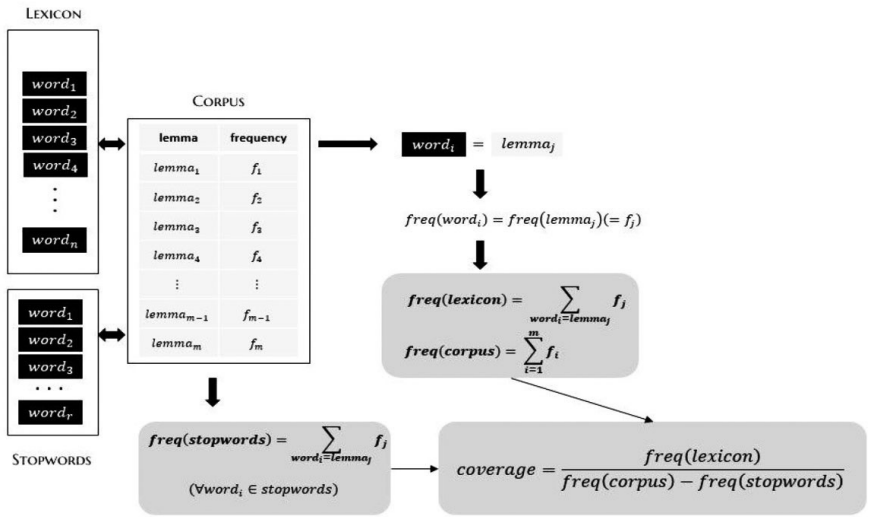

Figure 3: Process of matching words from the corpus with words from the lexicon.

If this statement is true and the word exists in the corpus, the value of freq(lexicon) is accumulated for the value of the frequency of each word, otherwise, the value 0 is added to freq(lexicon).

The sum of all word's frequencies from the corpus is given as freq(corpus) and freq(stopwords) presents the sum of all frequencies of the stopwords that appears in the corpus.

The coverage is counted as:

$$coverage = \frac{freq(lexicon)}{freq(corpus) - freq(stopwords)} \quad (1)$$

where all stopwords were excluded from the corpus.

4.2 The second approach: using the accuracy of words without the influence of frequency

Given that the sentiment value is not at the forefront at this stage of research (we are looking for language coverage), lists of 1,219 positive and 3,935 negative words were united in a unique lexicon.

In addition to lexicons, two other groups of words — stopwords (394 of them) and AnAwords (Affirmative and Non-affirmative words) – play a significant role in this process. Jahić and Vičič (2023a) pointed out that stopwords usually refer to the most common words in a language and that there is no single universal list of these.

However, 139 words from the AnAwords list were created by Jahić and Vičič (2023a), and it has been proven (Osmankadić, 2003) that most of these are intensifiers.

The consideration of words from the AnAwords list significantly impacted the corpus coverage, as elucidated in the second stage of the second approach. These words were evaluated in a manner similar to stopwords, given their absence of inherent sentiment value. Consequently, they were excised from the corpora, in line with the objective of eliminating non-sentiment-bearing terms.

Taking this into account, the AnAwords were also subject to examination, considering their lack of any discernible sentiment value. Consequently, they were treated analogously to stopwords, leading to their exclusion from the analysis

The process of annotating the lexicon went through several stages, and they were all based on the following equation:

$$\frac{FOUND}{NOT\_FOUND} \qquad (2)$$

where FOUND presented the list of all words in the corpus that were matched with words from the lexicon, NOT_FOUND opposite.

In more details, these stages are as follows:

- Simple coverage of the corpus by a lexicon was shown in the first stage. The stopwords were part of the corpus at this stage.
- In the first stage, stopwords were integrated into the corpus, contributing to the text's initial structure. However, as the coverage process unfolded in the second stage, the corpus coverage was achieved without the incorporation of stopwords, in addition to the exclusion of AnAwords words. The rationale behind these

decisions stems from the fact that the number of stopwords and AnAwords is almost negligible in comparison to the total number of elements in the corpus. As such, a substantial variance in coverage during this stage was not anticipated.

- Guided by the results of research conducted for the corpus-based lexical analysis of subject-specific university textbooks in English (Hajiyeva, 2015), in the third stage coverage was observed by the frequency distribution of words.
- In the fourth stage, the question arises as to whether it is possible to group similar words (such as “anđeo” and “anđel” (angel)) and view them as a single word. As Davies (2005) stated, one of the solutions to this problem is grouping words according to word families. Given this possibility of grouping, matching functions were applied between corpus words and lexicon words.
- In the fifth stage the log-likelihood was computed for each word in the lexicon. Following Rayson and Garside (2000), the word frequency list is then sorted by the resulting log-likelihood values. This gives the effect of placing the largest log-likelihood value at the top of the list representing the word that has the most significant relative frequency difference between the two corpora. This method enables a comparison of the most indicative (or characteristic) words in both corpora.

5 Results

This section showcases the results of the two approaches described earlier. We started by cleaning the corpora, which led to the inclusion of 263,969 words from the first corpus and 84,859 words from the second in our subsequent analysis (see Table 2).

Table 2: *Number of lemmas left after pre-processing the corpora*

	CORPUS1	CORPUS2
The overall number of lemmas	348,988	101,771
Cleared lemmas	263,969	84,859
Percent (%)	75.64	83.38

In the first approach (influence of frequency was considered), freq(corpus), the sum of stopwords frequencies freq(Stopwords)) and the overall sum of all frequencies of the words from the lexicon (freq(lexicon)) were computed.

By using equation (1) coverage of the corpus1 is 19.24%, and coverage of the corpus2 is 28.05% (see Table 3).

Table 3: *Coverage of the corpora's lemmas with words from sentiment lexicon*

	Freq(corpus)	Freq(lexicon)	Freq(stopwords)	Coverage
CORPUS1	187,957,442	28,174,959	41,542,468	19.24%
CORPUS2	3,0168,771	6,371,417	7,456,808	28.05%

The second approach (when the influence of frequency is ignored) was to compute the overall coverage of the corpora without using word frequencies. The motivation behind this approach was to count how many different lemmas from the corpus are already present in the sentiment lexicon. There are several stages in this approach.

- *First stage*: In this first stage, 1.523% coverage of the first corpus and 4.098% for the second corpus was achieved.

  Table 4: *Coverage of corpora's lemmas with words from the sentiment lexicon (without any additional changes being made)*

	CORPUS1	CORPUS2
FOUND	3,959	3,341
NOT_FOUND	260,010	81,518
Coverage (%)	1.523	4.098

  Table 4 presents lemmas that were matched with words from the lexicon (FOUND) and that were absent from the lexicon (NOT_FOUND).

  Maximum coverage of corpora is possible if all words from the lexicon are included in the corpora. This means that the maximum coverage for the first corpus is 1.99% and for the second corpus is 6.47%. In contrast, the coverage of the lexicon by the corpora is 76.81% and 64.82%. This means that of 5,154 words from the lexicon, 3,959 were presented in corpus1, indicating 76.81% use

of the lexicon, and 3,341 were presented in corpus2, indicating 64.82% use.

- The *second stage* increases the number of words in FOUND since all words that are stopwords or AnAwords have been detected in the corpus. In this case, coverage of corpora is increased to 1.7% and 4.62% for corpus1 and corpus2,respectively.

Table 5: *Coverage of corpora's lemmas with word from the sentiment lexicon*

	CORPUS1	CORPUS2
FOUND	4,406	3,747
NOT_FOUND	259,533	81,112
Coverage (%)	1.7	4.62

- The *third stage* is distributing words by frequency and counting the number of lemmas that were or were not covered by words from the lexicon.

  From a total of 5,154 words from the lexicon,3,257 (63.19%) were included in the 50,000 most frequent lemmas from corpus1 (see Figure 4 (left)). Meanwhile, of the 15,000 most frequent lemmas from corpus2, 3,071 were in the lexicon.Since the overall number of words from the lexicon is 5,154,this means that gave 59.58% of all words from the lexicon are found in the 15,000 most frequent lemmas from corpus2 (see Figure 4 (right)).

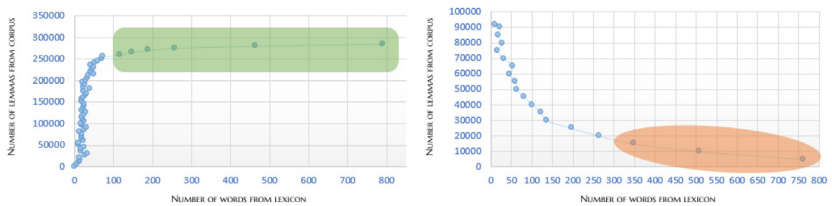

Figure 4: Annotated lexicon by distributed lemmas from corpus1 (left) and corpus2 (right).

- *Fourth stage*: In the fourth stage the lexicon annotation was increased to more than 2.2% for the first corpus. Even though it looks like this contradicts the claim that the maximum coverage for the first corpus is about 1.54%, it does not. The reason for this is

because the *get_close_matches* function was applied with several cutoffs and n=1 (one possibility).

$$get\_close\_matches(word, possibilities[, n][, cutoff])$$

The function works in such a way that all almost similar words (82.5% and 85% matching in this case for the first and second corpora, respectively) are considered as one word. For example, *anđel* (Engl. angel), *anđelko* (Engl. little angel), and *anđela* ("*I saw an* ***angel***"), all three words were replaced with the word *anđeo*. We have found that for a cutoff lower than of the 82.5% matching function returns words that are not matched or related to the root word.

The impact of *get_close_matches* is presented in Figure 5 on a small part of corpus2 with the matching word "anđeo" (angel).

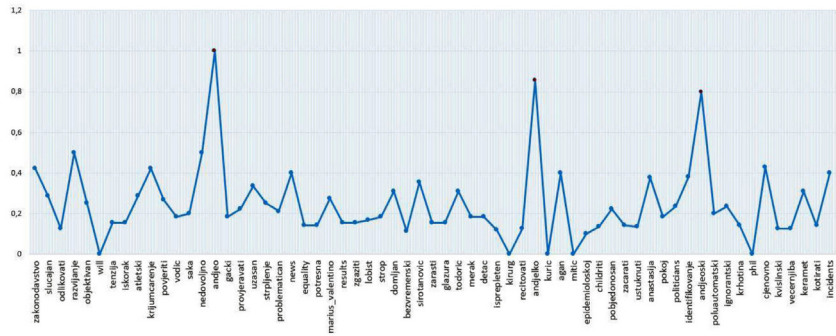

Figure 5: Implementation of get_close_matches on corpus1 of the word "anđeo" (an angel).

In the figure shown above there are some words whose matching factor with the word "anđeo" was greater than 0.85. Those words (anđeo, anđelko, anđeoski) were replaced with the word "anđeo". Having that in mind, the number of words (for corpus1) in the corpora decreased (see Table 6). If an 85% cutoff is applied, the overall number of words in corpus1 is reduced to 242,615, and the annotation increases to 2.22%. This means that out of 242,615 words in corpus1, 5,280 were found in the lexicon, stopwords, and AnAwords groups. The same applies to corpus2, where an annotation of 4,706 words from the lexicon in corpus2

was detected. Moreover, when a cutoff of 82.5% is applied, we get an annotation of 2.34% and 6.98% for the first and second corpora, respectively.

Table 6: *Coverage of corpora's lemmas with words from the sentiment lexicon*

Cutoff	82.5%		85%	
Corpus:	CORPUS1	CORPUS2	CORPUS1	CORPUS2
No. of lemmas	233,157	74,134	242,615	77,437
FOUND	5,327	4,838	5,280	4,706
NOT FOUND	227,830	69,296	224,939	72,731
Coverage (%)	2.34	6.98	2.22	6.47

- *Fifth stage*: In the fifth stage the log-likelihood was computed for each of the 5,000 most frequent lemmas from both corpora (10,000 overall) and only those that were common for both corpora were counted (4,207 in total).

  In this way, the opportunity was given to compare the frequencies of word form occurrences in two texts (for this purpose two corpora) and obtain a statistical measure of the significance of the differences.

  To compute the log-likelihood, a *two-by*-two *contingency table* (see Table 7) of frequencies for each word was constructed.

Table 7: *Contingency table for word frequencies*

	CORPUS1	CORPUS2	TOTAL
Freq. of word	a	b	a+b
Freq. of other words	c-a	d-b	c+d-a-b
TOTAL	c	d	c+d

In Table 7, *c* and *d* present the number of words in the corpora. In this case, c = 183m481,818 and d = 28m690m802. *c* and *d* were obtained by summing all the frequencies of all 4,207 words.

Following Rayson and Garside (2000), the equations:

$$E_1 = \frac{c \cdot (a+b)}{(c+d)}, E_2 = \frac{c \cdot (a+b)}{(c+d)}$$

were used to calculate the expected values, and:

$$LL(word) = 2 \cdot \left(a \cdot ln \frac{a}{E_1} + b \cdot ln \frac{b}{E_2} \right),$$

to calculate the log-likelihood for each word.

Using these equations a word frequency list (LL_list) was created, and the words were sorted from the smallest to largest values, where the largest value represents the word that has the most significant relative frequency difference between the two corpora.

As such, the most characteristic words of one corpus, as compared to the other corpus, were listed at the bottom of the given list, while words with almost the same relative frequency were listed at the top.

To evaluate the result and identify the N number of words that have similar interpreted values, we needed another method. As stated by Kilgarriff in *Comparing Corpora* (Kilgarriff, 2001), the simplest method that could be used for this is applying Sketch Engine. For each word the quotient of frequency was computed, and if the value of the quotient is 1 it indicates that its frequency is identical in both corpora. The higher the score in the Sketch Engine frequency word list (SE_list), the greater the difference between corpora. However, it should be noted that the given score can only be used for comparing differences, and it does not give clues as to what exactly is different between the corpora.

Because of this, the identification of words with similar interpreted values was done. This means that the percentage of coverage of N's highest keyness score words with words from the lexicon for both lists LL_list and SE_list was computed.

The 500 most relevant words for both lists were identified. These words distinguish one corpus from the other, and also present the strengths of one corpus over the other. First the log-likelihood was computed and the LL_list was created. A list of 500 words with the biggest frequency differences between the two corpora was created. Then the two corpora were compared

by using Sketch Engine and the SE_list was created. As for the LL_list, the 500 words with the biggest frequency differences between the two corpora were identified.

The aim was not to compare corpora but to check the coverage of the most relevant words – those that distinguish the two corpora from each other – with words from the lexicon.

Using these comparison methods produced a matching factor of 55.2% between LL_list and SE_list of the 500 most relevant words.

Table 8: *Coverage of 500 most relevant words from the lexicon group and distribution of words from the lexicon, stopwords, and AnAwords lists (LSAnA group)*

	LL_list				SE_list			
FOUND words	156				136			
Coverage (%)	31.2%				27.2%			
By words from	Lexicon		Stopwords	AnAwords	Lexicon		Stopwords	AnAwords
	pos	neg			pos	neg		
	30	50	62	14	44	42	35	15

As can be seen in Table 8, from the 500 words there were 156 from the LSAnA group that matches them (31.2% for the LL_list) and 136 words from the LSAnA,representing 16.8% of coverage from the SE_list.

Words from the lexicon had 51.28% coverage ((30+50)/156) and annotation of the lexicon in those 500 words from LL_list had 16% coverage (80/500). For the SE_list this annotation had about 27.2% coverage, and the overall impact of lexicon words with regard to all the words covered by the SE_list was about 63.24% ((44+42)/136).

Even though the third and fifth stages present insights into the annotation of the most frequent words, for overall annotation the most important stages were the first, second and fourth ones, since they produce the overall coverage of the corpora by lexicon (see Table 9).

Table 9: *Annotation of corpora*

Approach:	Coverage of corpora	
	CORPUS1 (bsWaC)	CORPUS2 (bsNews)
First	by using the accuracy of words with the influence of frequency	
	19.24%	28.05%
Second	by using the accuracy of words without the influence of frequency	
First	1.523%	4.098%
Second	1.7%	4.62%
Fourth	2.22%-2.34%	6.47%-6.98%

6 Conclusion

Although Bosnian is arguably closely related to Serbian and Croatian, there are subtle differences between these three languages that are more evident from the sentiment analysis point of view. This paper presents the annotation of the first Bosnian sentiment lexicon that has been proven on a sentiment basis in earlier work. The lexicon includes about 5,500 words (1,219 positive, 3,935 negative, 394 stopwords, and 139 AnA words) and covers more than 19% of the words in the first observed corpus (corpus1) (Ljubešić & Klubička,2014), and more than 28% of words in the second corpus, BsNews 1.0 corpus (corpus2), (Vičič,2021). If the emphasis is on coverage of different words from the corpus by the lexicon, then coverage is 1.7% for corpus1 and 4.62% for corpus2. This coverage will increase by applying some matching functions between the corpora's words and lexicon's words (as described in the fourth stage of the second approach in Section 4). In that case, the coverage rises to 2.34% for corpus1 and 6.98% for corpus2. From 85.07% to 93.67% of words from the lexicon were found in corpus1 (between 360 and 849 words from the lexicon were not found) and between 82.75% and 92.84% in corpus2, which means that between 407 and 981 words from the lexicon were not found in corpus2.

The results show that about a quarter of the words from the corpora have their sentiment value annotated in the lexicon, which greatly helps in the sentiment annotation of the sentences (tweets or regular text).

Stopwords and AnAwords were also included in the analysis, which leads to the possibility that the LSAnA group becomes a representative group for sentiment words, stopwords, and intensifiers (all written in Bosnian).

The language coverage of the lexicon is comparable with the current state of the art, and the values can be compared (Moreno-Ortiz & Pérez-Hernández, 2018).

During the process of creating our lexicon, we were aware that there would be deviations during the translation. The Slovene sentiment lexicon KSS 1.1 also includes multi-part words, which are words composed of multiple individual words joined by "_", such as "dobro_sprejet", "dobro_upravljan", "dobro_voden", "dobro_vzgojen", "energetsko_varčen", "funkcijsko_bogat", and similar terms. Most of these types of words do not have an equivalent in the Bosnian language. However, during the manual review of our lexicon we noticed that some of these words could be included, such as "prekomerna_teža" (Bosnian: predebelo, English: too fat) or "srce_parajoč" (Bosnian: srceparajuće, English: heartbreaking). Despite this, these words did not make it into the primary version of our lexicon.

Table 10: *Comparison of lexicon terms*

	Positive		**Negative**	
Language	core terms	terms with "_"	core terms	terms with "_"
Slovenian	1,911	61	5,152	276
Bosnian	1,126	-	3,868	-

According to Table 10, it is evident that the Bosnian lexicon can be updated by finding appropriate translations for multi-part words from KSS 1.1. This update would have an immediate impact on the lexicon's annotation, as incorporating more terms would allow for more comprehensive annotation. Furthermore, during our analysis we discovered that among these multi-part words some contained elements from the AnAwords list, which we treated separately. Examples of such cases include "hudo_bolan" (Bosnian: veoma bolan, English: very painful), "zelo_poceni" (Bosnian: veoma jeftin, English: very cheap), "povsem_prava" (Bosnian: potpuno pravo (tačno), English: completely right), and others.

Additionally, we found that there were entire expressions in the Slovenian lexicon that were not included in the Bosnian lexicon. Some examples of these are “nič_hudega_sluteč” (Bosnian: ne slutiti ništa loše, English: unaware of any harm), “obesiti_na_klina” (Bosnian: objesiti o klin, English: hang on a nail), and “veliko_hrupa_za_nič” (Bosnian: mnogo buke oko ničega, English: much ado about nothing), “zvit_kot_lisica” (Bosnian: lukav kao lisica, English: sly as a fox), among others.

The focus in future work will be on developing and improving the LSAnA group. All members of the group should be extended, which means that we expect to have more items/words labelled as positive or negative in our “core” lexicon, as well as extending lists of stop--words and AnAwords. To increase coverage, we will try to create a lexicon with all possible words, and in doing so we will contain all the grammatical rules found in the Bosnian language itself (declination, conjugation, change of words by gender, number, and so on). Although the process of annotation, as well as improvement of the first Bosnian lexicon (Jahić and Vičič, 2023b), is still in development, the results shown here are comparable with those reported for other related languages, and also for language families, as shown in Davies (2005) and Bučar, Žnidaršič and Povh (2018).

Acknowledgments

The authors gratefully acknowledge the European Commission for funding the InnoRenew CoE project (Grant Agreement #739574) under the Horizon 2020 Widespread-Teaming programme and the Republic of Slovenia (investment funding of the Republic of Slovenia and the European Union of the European Regional Development Fund).

References

Bučar, J., Žnidaršič, M., & Povh, J. (2018). Annotated news corpora and a lexicon for sentiment analysis in slovene. *Language Resources and Evaluation, 52*, 895– 919. doi:10.1007/s10579-018-9413-3

Chen, C., Hu, X., Zhang, H., & Shou, Z. (2020). Fine grained sentiment analysis based on Bert. *Journal of Physics: Conference Series, 1651*.

Davies, M. (2005). Vocabulary range and text coverage. insights from the forthcoming routledge frequency dictionary of spanish. *Selected Proceedings of the 7th Hispanic Linguistics Symposium* (pp. 106–115).

Erjavec, T., Ignat, C., Pouliquen, B., & Steinberger, R. (2015). Massive multi lingual corpus compilation: Acquis communautaire and totale. *Archives of Control Sciences 15*.

Glavaš, G., Šnajder, J., & Bašić, B. D. (2012). Semi-supervised acquisition of croatian sentiment. *Proceedings of the International Conference on Text, Speech and Dialogue*, *7499* (pp. 166–173). Brno, Czech Republic. doi:10.1007/978- 3- 642- 32790- 2_20

Hajiyeva, K. (2015). *A corpus-based lexical analysis of subject-specific university textbooks for english majors*, *2*, 136–144. doi:https://doi.org/10.1016/j.amper.2015.10.001

Hartman, J. J., Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1967). The General Inquirer: A Computer Approach to Content Analysis. *American Sociological Review, 4*. doi:10.2307/1161774

Iglesias, C., & Moreno, A. (2019). Sentiment Analysis for Social Media. *Sentiment Analysis for Social Media*, 1–4. Retrieved from https://www.mdpi.com/journal/applsci/special

Jahić, S., & Vičič, J. (2021). Determining sentiment of tweets using first Bosnian lexicon and (AnA)-affirmative and non-affirmative words. *Advanced technologies, systems, and applications V, 142*, 361–373. doi:https://doi.org/10.1007/978-3-030-54765-3_25

Jahić, S., & Vičič, J. (2023a). *Lists of stopwords and AnAwords of Bosnian language (1.00) [Data set].* doi:10.5281/zenodo.8021150

Jahić, S., & Vičič, J. (2023b). Sentiment polarity lexicon of Bosnian language. 361–373. Univerza na Primorskem; CERN. Retrieved from https://zenodo.org/record/7520809#.Y8-4L3bMLi0

Jahić, S., & Vičič, J. (2023c). Impact of Negation and AnA-Words on Overall Sentiment Value of the Text Written in the Bosnian Language. *Applied Science, 13*, 7760. doi:10.3390/app13137760

Jones, R. L. (2006). An analysis of lexical text coverage in contemporary German. In *Brill, Language and Computers* (pp. 115–120). Leiden, The Netherlands: Brill. doi:https://doi.org/10.1163/9789401202213_010.

Jovanoski, D., Pachovski, V., & Nakov, P. (2015). Sentiment analysis in Twitter for Macedonian. *Proceedings of the International Conference Recent Advances in Natural Language Processing* (pp. 249–257). Hissar, Bulgaria: INCOMA Ltd. Shoumen. Retrieved from https://aclanthology.org/R15-1034

Kadunc, K. (2016). *Določanje sentimenta slovenskim spletnim komentarjem s pomočjo strojnega.* Ljubljana: Fakulteta za računalništvo in informatiko Univerze v Ljubljani. Retrieved from https://repozitorij.uni-lj.si/IzpisGradiva.php?lang=eng&id=91182

Kapukaranov, B., & Nakov, P. (2015). Fine-grained sentiment analysis for movie reviews in Bulgarian. *Proceedings of the International Conference Recent Advances in Natural Language Processing* (pp. 266–274). Hissar, Bulgaria: INCOMA Ltd. Shoumen. Retrieved from https://aclanthology.org/R15-1036

Kia, D., Soujanya, P., Amir, H., Erik, C., Ahmad, H. Y., Alexander, G., & Qiang, Z. (2016). Multilingual Sentiment Analysis: State of the Art and Independent Comparison of Techniques. *Springer Link – Cognitive Computation, 8*, 757–771. doi:10.1007/s12559-016-9415-7

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics, 6*(1), 97–133. doi:https://doi.org/10.1075/ijcl.6.1.05kil

Ljubešić, N., & Klubička, F. (2014). bs,hr,srWaC - web corpora of Bosnian, Croatian and Serbian. *Proceedings of the 9th Web as Corpus Workshop (WaC-9)* (pp. 29–35). Gothenburg, Sweeden: Association for Computational Linguistics. doi:10.3115/v1/W14- 0405

Moreno-Ortiz, A., & Pérez-Hernández, C. (2018). Lingmotif-lex: a wide-coverage, state-of-the-art lexicon for sentiment analysis. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 2653–2659). Miyazaki, Japan: European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L18-1420

Okruhlica, A. (2013). *Slovak sentiment lexicon induction in absence of labeled data, Master's Thesis.* Comenius University Bratislava.

Osmankadić, M. (2003). A Contribution to the Classification of Intensifiers in English and Bosnian. 50–62.

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. *Proceedings of the Workshop on Comparing Corpora WCC'00. 9* (pp. 1–6). USA: Association for Computational Linguistics. doi:10.3115/117729.117730

Suciati, A., & Budi, I. (2020). Aspect-Based Sentiment Analysis and Emotion. *(IJACSA) International Journal of Advanced Computer Science and Applications, 11*(9), 179–186.

Veselovská, K. (2013). Czech subjectivity lexicon : A lexical resource for czech polarity classification. *Proceedings of the 7th international conference Slovko* (pp. 279–284). Bratislava.

Vičič, J. (2021). Bosnian news corpus 2021. Retrieved from http://hdl.handle.net/11356/1406

Wawer, A. (2012). Extracting emotive patterns for languages with rich morphology. *International Journal of Computational Linguistics and Applications*, 11–24.

Wu, F., Shi, Z., Dong, Z., Pand, C., & Zhang, B. (2020). Sentiment Analysis of Online Product Reviews Based On SenBERT-CNN. *International Conference on Machine Learning and Cybernetics (ICMLC)* (pp. 229–234). Adelaide, Australia: IEEE. doi:10.1109/ICMLC51923.2020.9469551

Razpoloženjsko označeni leksikon v bosanskem jeziku

Prispevek predstavlja prvi razpoloženjsko označeni leksikon bosanskega jezika. Postopek in metodologija označevanja sta predstavljena skupaj s študijo uporabnosti, ki se osredotoča na jezikovno pokritost. Sestava izhodišča je bila izvedena s prevajanjem slovenskega označenega leksikona in kasnejšim ročnim preverjanjem prevodov in oznak. Jezikovna pokritost je bila preverjana z uporabo dveh referenčnih korpusov. Bosanski jezik še vedno velja za jezik z malo jezikovnimi viri. Za bosanski jezik je na voljo referenčni korpus, ki ga sestavljajo samodejno preiskane spletne strani, vendar so avtorji ugotavljamo, da korpus z jasnim časovnim okvirom vsebnega besedila ni dosegljiv. Z zbiranjem novic z več bosanskih spletnih portalov je bil sestavljen korpus sodobnih besedil. V raziskavi sta bili uporabljeni dve metodi jezikovnega pokrivanja. Pri prvi je bil uporabljen frekvenčni seznam vseh besed, ekstrahiranih iz dveh referenčnih korpusov bosanskega jezika, druga metoda pa je prezrla frekvence kot glavni dejavnik pri štetju. Izračunana pokritost po prvi predstavljeni metodi za prvi korpus je bila 19,24 %, drugi korpus pa 28,05 %. Druga metoda daje 2,34 % pokritost za prvi korpus in 6,98 % za drugi korpus. Rezultati študije predstavljajo jezikovno pokritost, ki je primerljiva s znanimi metodami na tem področju. Uporabnost leksikona je bila dokazana že s primerjavo na Twitterju.

Ključne besede: Bosanski leksikon, korpus, analiza sentimenta, potrdilne in nepotrdilne besede (PnPbesede), ustavne besede, logaritemska verjetnost, označevanje