Sentiment Analysis of Serbian Old Novels

Ranka Stanković, Miloš Košprdić, Milica Ikonić Nešić, Tijana Radović

Sentiment Analysis of Serbian Old Novels | Ranka Stanković, Miloš Košprdić, Milica Ikonić Nešić, Tijana Radović | Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data, June 2022, Marseille, France | 2022 | |

http://dr.rgf.bg.ac.rs/s/repo/item/0006277

Proceedings of the SALLD-2 Workshop @LREC2020, pages 31–38
Language Resources and Evaluation Conference (LREC 2022), Marseille, 24 June 2022

Sentiment Analysis of Sentences from Serbian ELTeC corpus

Ranka Stanković, Miloš Košprdić, Milica Ikonić Nešić, Tijana Radović
University of Belgrade, Serbia, Studenski Trg 1, Belgrade, Serbia
ranka.stankovic@rgf.bg.ac.rs, tijana.n.radovic@gmail.com,
milica.ikonic.nesic@fil.bg.ac.rs, milos.kosprdic@gmail.com

Abstract

In this paper we present first study of Sentiment Analysis (SA) of Serbian novels from the 1840-1920 period. The preparation of sentiment lexicon was based on three existing lexicons: *NRC*, *AFFIN* and *Bing* with additional extensive corrections. The first phase of dataset refinement included filtering the word that are not found in Serbian morphological dictionary and in second automatic POS tagging and lemma were manually corrected. The polarity lexicon was extracted and transformed into *ontolex-lemon* and published as initial version. The complex inflection system of Serbian language required expansion of sentiment lexicon with inflected forms from Serbian morphological dictionaries. Set of sentences for SA was extracted from 120 novels of Serbian part of ELTeC collection, labelled for polarity and used for several model training. Several approaches for SA are compared, starting with for variation of lexicon based and followed by Logistic Regression, Naive Bayes, Decision Tree, Random Forest, SVN and k-NN. The comparison with models trained on labelled movie reviews dataset indicates that it can not successfully be used for sentiment analysis of sentences in old novels.

Keywords: sentiment lexicon, sentiment analysis, distant-reading, machine learning, old novels

1. Introduction

This paper presents Sentiment Analysis (SA) on a corpus of Serbian novels, from the 1840 – 1920 period, that is being developed under the umbrella of the “Distant Reading for European Literary History” COST Action CA16204, using different methods, including lexicon based SA. The lexicon based approach of SA for Serbian is not much used due to the lack of sentiment lexicons for Serbian. We have decided to work on development of the Serbian Sentiment Lexicon which will contribute in overcoming this gap. This paper presents first results in this research, including publishing lexical resource as Linguistic Linked Open Data in order to provide and enable further research of SA on different corpora written in Serbian.

The inspiration was found in lexicons described in (Iglesias and Sánchez-Rada, 2021), especially on a polarity lexicon of Latin lemmas, called LatinAffectus which is a part of LiLa – Linked Data-based Knowledge Base of Linguistic Resources and NLP tool for Latin language (Sprugnoli et al., 2020). The objective of LiLa was to connect and ultimately exploit the wealth of linguistic resources and NLP tools for Latin created so far, in order to bridge the gap between raw language data, NLP and knowledge descriptions, so in line with that our objective is to expand and enrich tools for NLP in Serbian language by creating this lexicon. Sprugnoli et al. (Sprugnoli et al., 2021) introduced fourth category: *mixed* where the opposite emotions where produced and it is not possible to find a clearly prevailing emotion (between lexicon and evoked images). It could be seen that it is somehow similar to our category ”both”, but we did not go in this direction since there were so few those entries, so we just eliminated them.

Hybrid sentiment analysis framework for a morphologically rich language (SAFOS) (Mladenović et al., 2016) used a sentiment lexicon and Serbian WordNet (SWN) synsets assigned with sentiment polarity scores in the process of feature selection. They expanded the lexicon generated using SWN, by adding morphological forms of emotional terms and phrases using Serbian Morphological Electronic Dictionaries (Krstev, 2008). Testing was performed on news and movie reviews, the best classification accuracy scores were achieved for the combination of unigram and bigram features reduced by sentiment feature mapping (accuracy 78.3 % for movie reviews and 79.2 % for news test set).

The sentiment analysis on Serbian Movie Review Dataset achieved best accuracy 85.5% for 2 classes and 62.2% for 3 classes, by using unigram, bigram and trigram features in a combination of l Naïve Bayes (NB) and Support Vector Machines (SVM) (Batanović et al., 2016).

Improving sentiment analysis for twitter data by handling negation rules in the Serbian language (Ljajić and Marovac, 2019) was based on grammatical rules that influence the change of polarity are processed. A statistically significant relative improvement was obtained (up to 31.16% or up to 2.65%) when the negation was processed using rules with the lexicon-based approach or machine learning methods. By applying machine learning methods, an accuracy of 68.84% was achieved on a set of positive, negative and neutral tweets, and an accuracy of as much as 91.13% when applied to the set of positive and negative tweets.

The NgramSPD (Graovac et al., 2019) explored n-gram models in conjunction with k Nearest Neighbourhood (kNN), Support Vector Machine (SVM) and Maximum Entropy (MaxEnt) algorithms to determine opinion polarity of the seven publicly available movie review benchmarks in Arabic, Czech, English, French, Spanish, Turkish, and Serbian. Formal evaluation confirmed that the proposed byte and character n-gram models outperform word n-gram model, and in conjunction with the presented MaxEnt algorithm outperform other machine learning supervised techniques used with more complex document representation approaches. Despite their simplicity and broad applicability, byte and character n-grams have been shown to be able to capture information on different levels – lexical and syntactic. For *SerbMR-2* best performance was achieved with accuracy 85.54% by maxEnt, while with kNN 81.14% and SVM 83.47%.

2. Sentiment Lexicons

2.1. Existing Sentiment Lexicons

In the lexicon-based approach the polarity of the text is determined on the basis of a set of positive, negative and neutral words (Mostafa and Nebot, 2020). To implement a semantically based approach, lexicons of sentiments are used, in which words are classified as positive, negative or neutral according to its polarity. The polarity of the whole text represent a combination of the polarity of the words that make up the text. Currently, there is large number of different lexicons of sentiments however, three that are most commonly used (Silge and Robinson, 2017) are: *Bing* (Liu et al., 2004), *NRC* (Mohammad and Turney, 2010) i *AFINN* (Nielsen, 2011).

The *NRC* lexicon of Sentiments (Mohammad and Turney, 2010) classify words according to polarity as positive or negative, and according to the category of emotions to which they belong (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust). Determining the polarity and the category of emotions to which words belong was manually done by crowd-sourcing. The *AFINN* lexicon is a list of English terms manually rated for valence with an integer between -5 (negative) and +5 (positive) (Nielsen, 2011). The *Bing* sentiment lexicon is a general purpose English sentiment lexicon that consists of manually categorized words in a binary fashion, either positive or negative (Liu et al., 2004).

These three lexicons can be found as a part of numerous packages that are used for lexicon-based sentiment analysis in R programming language such as *tidytext* described in (Silge and Robinson, 2017) and *syzhet* (Jockers and Thalken, 2020). The tidytext package (Silge and Robinson, 2017) specializes in preprocessing, analyzing, and visualizing textual data. Also, this package provides access to *NRC*, *AFINN* and *Bing* lexicons of sentiments which enables extraction of sentiments in text. Syuzhet package (Jockers and Thalken, 2020) comes with four sentiment dictionaries and provides a method for accessing the robust, but computationally expensive, sentiment extraction tool developed in the NLP group at Stanford. The main functions in the package are quickly extraction of sentiments from your own text files. More precisely, this package serves to extracts sentiment and sentiment-derived plot arcs from text using a variety of sentiment dictionaries conveniently packaged for consumption by R users. Implemented dictionaries include *syuzhet'* developed in the Nebraska Literary Lab, *Bing* (Liu et al., 2004), *NRC* (Mohammad and Turney, 2010) and *AFINN* (Nielsen, 2011). Althought, *Bing*, *NRC* and *AFINN* lexicon are widely use, there are sentiment analysis packages in R that use other sentiment lexicons.

The Sentiment Analysis package (Pröllochs et al., 2018) (Nicolas Proellochs and Stefan Feuerriege, 2021) introduces a powerful toolchain facilitating the sentiment analysis of textual contents in R. This implementation utilizes various existing dictionaries, such as QDAP (Qualitative Data Analysis Package) and, Harvard IV and Loughran-McDonald. Furthermore, it can also create customized dictionaries. The latter function uses LASSO regularization as a statistical approach to select relevant terms based on an exogenous response variable. Finally, all methods can be easily compared using built-in evaluation routines.

Lexicon of sentiments created for this research is based on three existing lexicons: *NRC*, *AFFIN* and *Bing* with additional extensive manual corrections.

2.2. Senti-Pol-sr Sentiment Lexicon

Entries from *NRC*, *AFFIN* and *Bing* lexicons are available in Serbian or Serbo-Croatian but mostly by automatic translation with numerous entries with translation errors and English terms instead of Serbian translation equivalent. The headwords of three lexicons were merged, duplicate entries were removed and union of polarities were assigned in the first step. The shallow lexicon was produced, where not all headwords were assigned all categories. However, polarity -1, 1 was either assigned or possible to derived for all. For AFINN lexicon from -5 to -2 was assigned -1 (negative), -1, 0, +1 were assigned 0 (neutral) and from +2 to +5 (positive).

Several entries in Serbian side of lexicon were multiple since different English words had same translation e.g. *odvratan* is aligned with *depraved, despicable, disgusting, distasteful, distracted, hideous, loathsome, obnoxious, odious, revolting, sickening*, so the new acquiring a new list of entries with distinct headwords in Serbian was produced.

The elimination of words that do not belong to Serbian language was based on Serbian morphological dictionaries (Krstev and Vitas, 2006) that are managed by Leximirka developing environment (Stanković et al., 2019). If headword was not found in lexical database either as lemmatized or inflected form, it was eliminated. If the headword was not found as lemmatized but it was found as inflected form, the lemma was corrected. The part of speech label was also assigned to the new lexical entry. All words that were not in lexicon were removed for this experiment. However, for further research additional exploration off excluded dataset is envisaged.

The preliminary inspection ed that words are mostly foreign *Hawking, headdress, idleness*, so proper translation is required. The evaluation of a lexicon was done by two annotators who used English dictionary Morton Benson in order to manually evaluate our new lexicon. While manually evaluating one of the challenges was status of those terms that in English are represented with one word, but translated into Serbian have two words, for example, English word *scapegoat* is in Serbian translated as two words *žrtveno jagnje*, or *hearse* as *mrtvačka kola*. Moreover, the similar problem was when translation equivalent is a phrase in Serbian: *forsooth (ma nemojte mi reći)* or *halfway (na pola puta)*. Also, some adverbs and adjectives in English have the same form and they occurred in the lexicon twice but as different part of speech, for example the word *hilarious* was tagged as an adjective and as an adverb.

The manual disambiguation, correction, exclusion of contradictory (different) polarity of the same word followed. A number of new entries with lexical variants and synonyms of already existing entries was introduced.

The overview of positive, negative, both positive and negative is given, with a total column at the end of Table1. The graphical overview is given in Figure 1.

For further analysis words that had both polarities were excluded.

	pos	neg	both	total
NRC	2231	3243	81	5555
AFFIN	1293	878		2171
Merged	5889	10197	225	16311
Filtered	3387	5058	154	8599
Distinct	2678	3628	148	6454

Table 1: The sentiment lexicon entries statistics table.

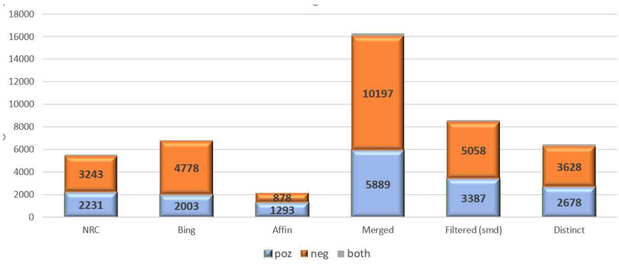


Figure 1: The sentiment lexicon entries statistics graph.

For transformation of produced lexicon *Senti-Pol-sr* into *ontolex-lemon* model (McCrae et al., 2011; McCrae et al., 2017) we adapted procedure in Leximir tool (Stanković and Krstev, 2012), based on approach described in (Ranka et al., 2018) and adapted . The initial form of lexicon (Stanković et al., 2022) is published in: http://llod.jerteh.rs/SA/. An excerpt of lexicon is:

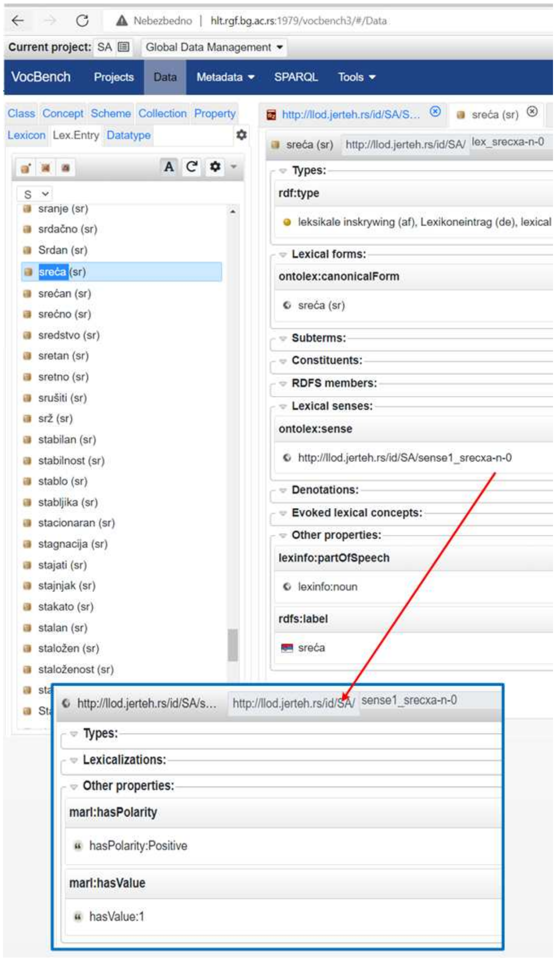
```
:SentiPolLexicon a lime:Lexicon;
  dct:title "SentiPol"@sr;
  lime:entry :lex_folirant;
  lime:language "sr"^^xsd:language .

:lex_folirant a ontolex:LexicalEntry;
   ontolex:canonicalForm :form_folirant;
   rdfs:label "folirant"@sr;
   lexinfo:partOfSpeech "noun"@sr;
   ontolex:sense :sense1_folirant-n-0-sense1.

:form_folirant a ontolex:Form;
   ontolex:writtenRep "folirant"@sr.

:sense1_folirant marl:hasPolarity
   "hasPolarity:Negative";
   marl:hasValue "hasValue:-1".
```

Senti-Pol-sr ontolex-lemon version is loaded in Vocbench (Stellato et al., 2015) for further exploration and refinement and for retrieval via SPARQL endpoint. Figure 2 presents a list of enties starting with *s* focused on *sreća* (happiness) with positive polarity. (Armando Stellato et al., 2021)



hlt.rgf.bg.ac.rs:1979/vocbench3/#/Data

sranje (sr)
srdačno (sr)
Srdan (sr)
sreća (sr)
srećan (sr)
srećno (sr)
sredstvo (sr)
sretan (sr)
sretno (sr)
srušiti (sr)
srž (sr)
stabilan (sr)
stabilnost (sr)
stablo (sr)
stabljika (sr)
stacionaran (sr)
stagnacija (sr)
stajati (sr)
stajnjak (sr)
stakato (sr)
stalan (sr)
staložen (sr)
staloženost (sr)

http://llod.jerteh.rs/id/SA/S... | sreća (sr)

sreća (sr) http://llod.jerteh.rs/id/SA/ lex_srecxa-n-0

Types:
rdf:type
leksikale inskrywing (af), Lexikoneintrag (de), lexical

Lexical forms:
ontolex:canonicalForm
sreća (sr)

Subterms:
Constituents:
RDFS members:

Lexical senses:
ontolex:sense
http://llod.jerteh.rs/id/SA/sense1_srecxa-n-0

Denotations:
Evoked lexical concepts:

Other properties:
lexinfo:partOfSpeech
lexinfo:noun
rdfs:label
sreća

http://llod.jerteh.rs/id/SA/s... http://llod.jerteh.rs/id/SA/ sense1_srecxa-n-0

Types:
Lexicalizations:

Other properties:
marl:hasPolarity
hasPolarity:Positive
marl:hasValue
hasValue:1

Figure 2: The sentiment lexicon in VocBench.

3. Labeled Dataset Preparation

3.1. Annotation guidelines

The annotations were done on a sentence level. The annotator's job was to determine is the given sentence positive, negative or neutral. In order to determine the polarity of the sentence annotator should rely on its intuition as a native speaker of a given language. The lack of these approach is that for some sentences such as sarcastic sentences or sentences when one side wins against another may be hard to determine the polarity of the sentence without given specifications about what is positive, what is negative and what is neutral (Mohammad, 2016).

In order to determine the polarity of sentence the annotator should consider positive all sentences that express support, admiration, positive attitude, forgiveness, fostering, success, positive emotional state. Negative sentences are those that include expressions of criticism, judgment, negative attitude, questioning validity/competence, failure, negative emotion. Finally, when the speaker is neither using positive language nor using negative language only giving the description of some event or place or talking about facts those sentences are marked as neutral (Mohammad, 2016).

While annotating, it was important that agreeing or disagreeing with the speaker's views should not have a bearing on annotator's response. The job of the annotator is to assess the language being used (not the views of the speaker). For example, the sentence, 'Evolution makes no sense', should be marked as negative since the speaker's words are criticizing or judging negatively something (in this case the theory of evolution). Note that the answer is not contingent on whether you believe in evolution or not. This approach groups the speaker's emotional state, speaker's opinion, and description of valanced events all into one category and aims simply to determine the dominant sentiment inferable from the sentence. For example, 'Yay! Novak beats Nadal 3–2' will be marked as positive because the speaker is using the positive expression 'Yay!'. Also, in the example 'Serbia lost to Montenegro' it may be difficult to annotate with respect to the opinion of the speaker towards the Serbian team, but the framing of the event as a loss is easily identified as negative expression (Mohammad, 2016).

3.2. Sentiment Dataset

The ELTeC[1] (Odebrecht et al., 2021) multilingual corpus of novels written in the time period 1840–1920 is built to test various distant reading methods among them sentiment analysis. Serbian part of ELTeC corpus (Krstev, 2021), dubbed *SrpELTeC*, comprises 100 novels in main collection and 20 in extended collection. The novels have structural annotations, sentence splitting, words are POS-tagged, lemmatized and seven classes of named entities are annotated (Stanković et al., 2022a).

From srpELTeC novels collection set of 30K sentences was extracted, relaying on sentence segmentation encoded in TEI XML, with <s> XMLS element. The <s> element was used to mark orthographic sentences, or any other segmentation of a text, provided that the segmentation is end-to-end, complete, and non-nesting. The number of positive and negative words was computed and assigned to each sentence. Sentences with different size and with different number of positive and negative words (according to lexicon presented in 2.2) were chosen. The set of sentences for manual evaluation was selected in several runs. The evaluation started with set where 5 or more positive words were found, than where 5 or more negative words was found. In second run set of sentences with at least one occurrences form sentiment lexicon was found and at the end without word from lexicon. The goal was to produce balances set with equal number of positive, neutral and negative sentences.

Four evaluators evaluated 1320 sentences and each sentence was evaluated by two evaluators. For 1089 evaluators had an agreement while 231 sentences were labeled differently. Inter-annotator agreement was calculated using ReCal2 tool (Deen Freelon, 2011) that show: Percent Agreement 82.5%, Scott's Pi 0.737, Cohen's Kappa 0.739, Krippendorff's Alpha (nominal) 0.737.

For this experiment we proceeded with sentences where evaluators had an agreement and the rest of the sentences will be later harmonised. At the end, in each class: positive, neutral and negative, there was 363 sentences. For 2 class classification only 726 senteces was used were data set is named *SrpELTeC-2C* and for 3 class classification 1089 named *SrpELTeC-3C*.

4. Sentiment Analysis

4.1. Experimental Approach

The sentiment data set from Section 3.2 with sentences from SrpELTeC novel collection in this section will be analysed by several models. In the first approach we analysed lexicon based model on both data sets *SrpELTeC-2C* and *SrpELTeC-3C* using different experiments with lexicon based models, including also combination of lexicon based models with other approaches witch will be describe briefly in Section 4.2.

In Section 4.3 will be given binary classification on *SerbMR-2C* and *SrpELTeC-2C* dataset using different methods for binary classification. By using Logistic Regression, Decision Tree, Random Forest and k-NN we trained models on datasets: *SerbMR-2C* (The Serbian movie review dataset, 2 classes) (Batanović et al., 2016) and on dataset *SrpELTeC-2C* (The Serbian ELTeC novels dataset, 2 classes). The models were evaluated and the results were compared . Logistic Regression and SVM using n-grams shown that results can be different quality using different n-grams vectoriza-

[1] ELTeC: European Literary Text Collection

tion. For the purpose of this research we compared trained models on *SerbMR-2C* dataset and evaluated on *SrpELTeC-2C* as vice verse. In further work is planed to do the same on data set with tree classes.

Classification of novel's sentences based on their sentiment show that the results on lexicon based approach are better than trained models. The outcome was expected, since in case of SerbMR-2C, the dataset used for training was on movies review and lexica and language style are different than in old novels, while for *SrpELTeC-2C* the dataset was too small.

4.2. Lexicon Based Classification

In this section we present different lexicon based models approaches using our produced lexicon on the *SrpELTeC-3C* dataset, and we give the brief comparison with the same models on the *SrpELTeC-2C* dataset. For the purpose of this work we done few experiments inspired by solution published in (Mitrović, 2021):

- **Experiment 1**: Solution based only on the sentiment lexicon. The model is comprised of one parameter only - limit. The average polarity of each sentence (sample) is calculated (essentially whether there are more positive or negative words). The prediction takes into consideration the calculated average only if it is grater than the limit or not. For the three class data, the model is comprised of two parameters, the positive and the negative limit. The parameters determine whether the sentence is positive (if the mean word polarity is grater than positive limit), negative (if the mean word polarity is less than negative limit) or neutral (if the mean word polarity is between those limits).

- **Experiment 2**: Solution based only on the sentiment lexicon, however this time it takes into consideration the ratio of positive / negative and the total number of words in the sentence.

- **Experiment 3**: Baseline model using Multinomial Naïve Bayes (MNB) with features only devised from the sentiment lexicon.

- **Experiment 4**: Baseline model using MNB with Bag-of-Words approach combined with the features of the sentiment lexicon.

Table 2 represent accuracy for four lexicon based experiments explained above with comparison on two evaluation datasets. The results that are much worse on SrpELTeC data are probably caused by lexical variety in novels and the fact that the novels might have lexica that is not in common used nowadays.

The best results were achieved in Experiment 4 and confusion matrix for two classes is presented in Figure 3 and Figure 4.

Accuracy	SerbMR–2C	SrpELTeC–3C
Experiment 1	0.864	0.649
Experiment 2	0.849	0.576
Experiment 3	0.848	0.657
Experiment 4	**0.878**	**0.719**

Table 2: Accuracy of SA on evaluation dataset for lexicon based experiments

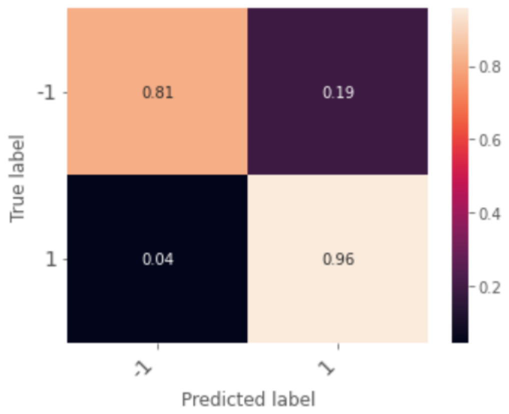


Figure 3: Confusion matrix for Experiment 4 with SrpELTeC-2C

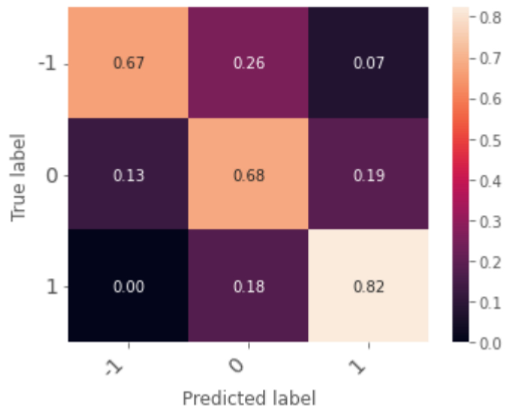


Figure 4: Confusion matrix for Experiment 4 with SrpELTeC-3C

4.3. Binary Classification on *SerbMR-2C* and *SrpELTeC-2C* dataset

In this section, we will approach the task of analyzing sentiments as a task of binary classification. We will get acquainted with three algorithms for classifying classical machine learning: logistic regression, random forests and k-nearest neighbors, and at the end, we will compare the performance of the model on two data sets *SerbMR-2C* and *SrpELTeC-2C* dataset.

The first machine learning algorithm we will encounter is logistic regression, as one of the basic algorithms of binary classification, so it is often encountered in comparative analyzes of model performance as a base model. The following Figure 5 shows the vocabulary words most deserving of the classification of texts by sentiment.

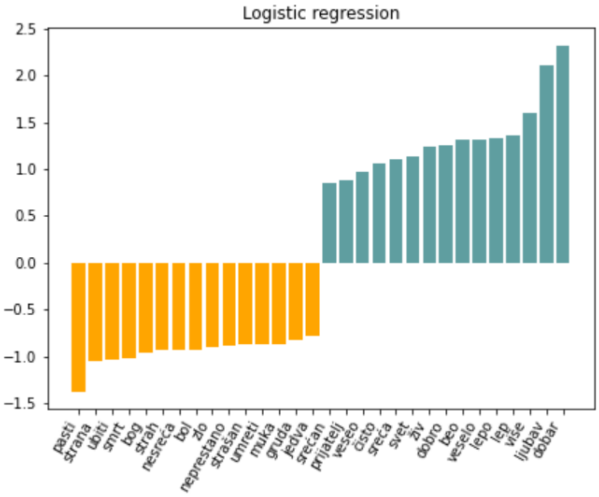


Figure 5: The vocabulary words.

The second algorithm was a decision tree, as an algorithm that learn a set of rules that can determine whether an instance is positive or negative. The Figure 6 presents the tree where in each node of the tree, the test is stated, then the value of the homogeneity measure used, the total number of instances analyzed, as well as the number of instances by classes.

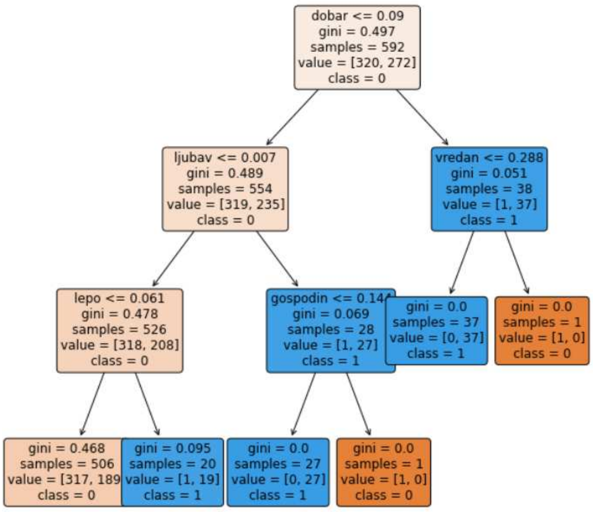


Figure 6: The decision tree subset.

For the next type of training we used Random Forest and k-nearest neighbors (KNN) algorithm for k = 3.
The Table 3 presents accuracy for 4 methods on two datasets: trained on 80%, tested 10% and evaluated on 10% of the same dataset. For each dataset, the training model performance is evaluated for original and lemmatized text and the best accuracy for each dataset is emphasized. The part of speech tagging and lemmatization was performed using tagger for Serbian (Stanković et al., 2022b) that is using (Krstev et al., 2021; Škorić and Stanković, 2021). Results clearly show that lemmatized option achieve better accuracy.

Method	SerbMR-2C		SrpELTeC-2C	
accuracy	token	lemma	token	lemma
Log. Regr.	0.828	**0.831**	0.768	**0.878**
Dec. Tree	0.590	0.597	0.561	0.621
Rand. for.	0.692	0.733	0.698	0.681
k-NN	0.656	0.674	0.657	0.757

Table 3: Evaluation of four SA models on *SerbMR-2C* and SrpELTeC-2C

Logistic Regression gave the best accuracy (Table 3), so in addition to previously evaluated model using tf–idf representation, we proceeded with further training on unigrams, bigrams and trigrams using lemmatized *SrpELTeC-2C* text. However, we included also SVM in this phase and the results are presented in Table 4.

Model accuracy	Log. Reg.	SVM
tf-idf vec.	**0.878**	**0.891**
unigram vec.	0.877	0.876
bigram vec.	0.592	0.601
trigram vec.	0.521	0.531

Table 4: SrpELTeC accuracy of SA on evaluation dataset for Logistic Regression and SVM

The research question was: can we use model trained on one dataset for SA of another? Namely, can *SerbMR-2C* (more that double in size in number of samples, but much more in number of words) be used for *SrpELTeC-2C* SA (and vice versa)?
Two experiments were conducted using different datasets for training and evaluation:

- **Experiment 5**: Trained model on *SrpELTeC-2C* dataset and evaluated on 10% of *SerbMR-2C* dataset (169 reviews)

- **Experiment 6**: Trained model on *SerbMR-2C* dataset and evaluated on 10% of *SrpELTeC-2C* dataset (66 sentences)

Table 5 shows that the accuracy is much lower that those presented in Table 3 and Table 4. We suspect that the reason is the difference in lexica and language style between the datasets *SerbMR-2C* and SrpELTeC-2C. So we conclude that in order to achieve better performance we have to proceed with enlarging SrpELTeC dataset for model training.

Model accuracy	Experiment 5	Experiment 6
Log. Reg.	0.550	**0.681**
Dec. Tree	**0.556**	0.454
Random forest	**0.556**	0.575
k-NN	0.474	0.467

Table 5: Accuracy of SA on cross-dataset evaluation

5. Conclusion

We outlined the research on development and application of sentiment lexicon, (sentence) dataset labelling and training of the models for sentiment analysis. The challenges in these tasks were discussed, as well as statistics of developed resources and performance of the training models. The first presented approach was with lexicon based model using four different experiments, with the best accuracy 87.8% on the *SrpELTeC-2C* and 71.9% on the *SrpELTeC-3C* using MNB with Bag-of-Words approach combined with the features of our sentiment lexicon (experiment 4). The second approach was based on trained models Logistic Regression, Decision Tree, Random Forest and k-NN using labeled datasets. The Logistic Regression gave the best accuracy 87.8%. By preliminary comparison of miss-classified sentences we have fond missing entries in a lexicon: *zavoleti (fall in love)*, *milina (grace, enjoyment)*, *nesrećnik (unfortunate person)*, *sirotinja (poor people)* etc. The current activities are focused on producing larger set of manually evaluated sentences that will enable more suitable training dataset. The analysis of miss-classified sentences with lexicon-based approach will be used for lexicon improvement. Final version of lexicon will be published also in ELG portal (Rehm et al., 2020) and in a public SPARQL endpoint. Plan is also to add examples to the lexicon using FrAC - frequency and attestations for *ontolex-lemon* (Chiarcos et al., 2020). First steps towards RDF editions of the ELTeC corpus are publishing two Serbian novels Ivkova slava : pripovetka (Ivko's patron saint's day: a short story) and Nečista Krv (Impure blood), POS-tagged, lemmatized, with NER and NEL with Wikidata, available in NIF (Ikonić Nešić and Stanković, 2022), so integration and futher conversion is envisaged.

Further research will be guided towards 1) fine-tuning the lexicon: adding synonyms and antonyms, adding words found in positive and negative sentences that were ”missed” by dictionary approach 2) including word embeddings in model training, 3) analyse sentences with negation in context that is related to sentiment.

6. Acknowledgements

The presented work is supported by the the COST Action CA18209 – NexusLinguarum “European network for Web-centred linguistic data science” and the COST Action 16204 – Distant Reading for European Literary History support. The authors would like to thank Mihailo Škorić for lemmatization, and Biljana Rujević for participation in the sentiment annotation.

7. Bibliographical References

Batanović, V., Nikolić, B., and Milosavljević, M. (2016). Reliable baselines for sentiment analysis in resource-limited languages: The serbian movie review dataset. In *in Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2688–2696, Portorož, Slovenia.

Chiarcos, C., Ionov, M., de Does, J., Depuydt, K., Khan, F., Stolk, S., Declerck, T., and McCrae, J. P. (2020). Modelling frequency and attestations for ontolex-lemon. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 1–9.

Graovac, J., Mladenović, M., and Tanasijević, I. (2019). Ngramspd: Exploring optimal n-gram model for sentiment polarity detection in different languages. *Intelligent Data Analysis*, 23(2):279–296.

Iglesias, C. Á. and Sánchez-Rada, J. F. (2021). Sentiment analysis meets linguistic linked data: An overview of the state of the art. In *Workshop on Sentiment Analysis & Linguistic Linked Data, September 01, 2021, Zaragoza, Spain.*

Jockers, M. L. and Thalken, R. (2020). Sentiment analysis. In *Text Analysis with R*, pages 159–174. Springer.

Krstev, C. (2008). *Processing of Serbian. Automata, Texts and Electronic Dictionaries*. Faculty of Philology of the University of Belgrade.

Krstev, C. (2021). The serbian part of the eltec collection through the magnifying glass of metadata. *Infotheca - Journal for Digital Humanities*, 21(2):26–42.

Liu, B., Li, X., Lee, W. S., and Yu, P. S. (2004). Text classification by labeling words. In *Aaai*, volume 4, pages 425–430.

Ljajić, A. and Marovac, U. (2019). Improving sentiment analysis for twitter data by handling negation rules in the serbian language. *Computer Science and Information Systems*, 16(1):289–311.

McCrae, J., Spohr, D., and Cimiano, P., (2011). *Linking Lexical Resources and Ontologies on the Semantic Web with Lemon*, pages 245–259. Springer Berlin Heidelberg, Berlin, Heidelberg.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., and Cimiano, P. (2017). The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.

Mladenović, M., Mitrović, J., Krstev, C., and Vitas, D. (2016). Hybrid sentiment analysis framework for a morphologically rich language. *Journal of Intelligent Information Systems*, 46(3):599–620.

Mohammad, S. and Turney, P. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34.

Mohammad, S. (2016). A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 174–179.

Mostafa, M. M. and Nebot, N. R. (2020). The arab image in spanish social media: A twitter sentiment analytics approach. *Journal of Intercultural Communication Research*, 49(2):133–155.

Nielsen, F. Å. (2011). A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*.

Pröllochs, N., Feuerriegel, S., and Neumann, D. (2018). Statistical inferences for polarity identification in natural language. *PloS one*, 13(12):e0209323.

Ranka, S., Cvetana, K., Biljana, L., and Mihailo, Š. (2018). Electronic dictionaries-from file system to lemon based lexical database. In *Proceedings of the 11th International Conference on Language Resources and Evaluation-W23 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science (LDL-2018), LREC 2018, Miyazaki, Japan, May 7-12, 2018*, pages 48–56.

Rehm, G., Berger, M., Elsholz, E., Hegele, S., and et al., F. K. (2020). European language grid: An overview. *CoRR*, abs/2003.13551.

Silge, J. and Robinson, D. (2017). *Text mining with R: A tidy approach.* " O'Reilly Media, Inc.".

Sprugnoli, R., Mambrini, F., Moretti, G., and Passarotti, M. (2020). Towards the modeling of polarity in a latin knowledge base. In *WHiSe@ ESWC*, pages 59–70.

Sprugnoli, R., Mambrini, F., Passarotti, M., and Moretti, G. (2021). Sentiment analysis of latin poetry: First experiments on the odes of horace. In *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021). Milan, Italy, January 26-28, 2022*, pages 1–7. CEUR Workshop Proceedings (CEUR-WS. org).

Stanković, R., Krstev, C., Šandrih Todorović, B., and Škorić, M. (2022a). Annotation of the serbian eltec collection. *Infotheca - Journal for Digital Humanities*, 21(2):43–59.

Stanković, R., Škorić, M., and Šandrih Todorović, B. (2022b). Parallel bidirectionally pretrained taggers as feature generators. *Applied Sciences*, 12(10).

Stellato, A., Rajbhandari, S., Turbati, A., Fiorelli, M., Caracciolo, C., Lorenzetti, T., Keizer, J., and Pazienza, M. T. (2015). Vocbench: A web application for collaborative development of multilingual thesauri. In Fabien Gandon, et al., editors, *The Semantic Web. Latest Advances and New Domains*, pages 38–53, Cham. Springer International Publishing.

8. Language Resource References

Armando Stellato et al. (2021). *VocBench*. University of Rome Tor Vergata, Italy, `http://vocbench.uniroma2.it/`, 3.0.

Deen Freelon. (2011). *Reliability Calculator for 2 coders*. Deen Freelon, `http://dfreelon.org/utils/recalfront/recal2/#doc`, 1.0.

Milica Ikonić Nešić and Ranka Stanković. (2022). *srpNIF*. `http://llod.jerteh.rs/ELTEC/srp/NIF/`.

Cvetana Krstev and Duško Vitas. (2006). *SrpMD - Serbian morphological dictionaries*. ELG, `https://live.european-language-grid.eu/catalogue/lcr/17355`, 1.0.

Cvetana Krstev and Duško Vitas and Ranka Stanković and Mihailo Škorić. (2021). *SrpMD4Tagging - Serbian Morphological Dictionaries for Tagging*. ELG, `https://live.european-language-grid.eu/catalogue/lcr/9294`, 1.0.

Miljan Mitrović. (2021). *Sentiment Analysis SerbMR*. github.

Nicolas Proellochs and Stefan Feuerriege. (2021). *R Package 'SentimentAnalysis': Dictionary-Based Sentiment Analysis*. github, `https://github.com/sfeuerriegel/SentimentAnalysis`, 1.3-4.

Carolin Odebrecht and Lou Burnard and Christof Schöch. (2021). *European Literary Text Collection (ELTeC): April 2021 release with 14 collections of at least 50 novels*. Zenodo, `https://github.com/COST-ELTeC`.

Ranka Stanković and Cvetana Krstev. (2012). *LeXimir - Tool for lexical resources management and query expansion*. 1.0.

Ranka Stanković and Cvetana Krstev and Mihailo Škorić and Biljana Lazić. (2019). *Leximirka - lexicographic database and a web application for developing, managing and exploring lexicographic data*. 1.0.

Ranka Stanković and Tijana Radović and Miloš Kosprdić. (2022). *Senti-Pol-sr - Lexicon for sentiment analisys, draft version*. 1.0.

Mihailo Škorić and Ranka Stanković. (2021). *SrpKor4Tagging-TreeTagger*. ELG, `https://live.european-language-grid.eu/catalogue/ld/9296`, 1.0.