Abusive Speech Detection in Serbian using Machine Learning

Danka Jokić[1,3], Ranka Stanković[2], Branislava Šandrih Todorović[3]
[1]University of Belgrade, Serbia
[2]Faculty of Mining and Geology, University of Belgrade, Serbia
[3]NLB DigIT d.o.o., Serbia

danka.jokic@afrodita.rcub.bg.ac.rs, ranka.stankovic@rgf.bg.ac.rs,
branislava.sandrih.todorovic@nlbdigit.rs

Abstract

The increase in the use of abusive language on social media and virtual platforms has emphasized the importance of developing efficient hate speech detection systems. While there have been considerable advancements in creating such systems for the English language, resources are scarce for other languages, such as Serbian. This research paper explores the use of machine learning and deep learning techniques to identify abusive language in Serbian text. The authors used AbCoSER, a dataset of Serbian tweets that have been labeled as abusive or non-abusive. They evaluated various algorithms to classify tweets, and the best-performing model is based on the deep learning transformer architecture. The model attained an F1 macro score of 0.827, a figure that is commensurate with the benchmarks established for offensive speech datasets of a similar magnitude in other languages.

1 Introduction

As the number of Web and social network users increases, abusive speech and its detection are becoming very important (Hardage et al., 2020).
The concept of abusive speech, in the context of this paper, is an umbrella term for phenomena such as profanities or offensive and hate speech. Caselli et al. (2020) defined abusive language as 'hurtful language that a speaker uses to insult or offend another individual or a group of individuals based on their personal qualities, appearance, social status, opinions, statements, or actions. This might include hate speech, derogatory language, profanity, toxic comments, racist and sexist statements.' The definition of abusive speech is very broad, and it makes the problem of its identification and detection even more challenging. Abusive speech, as outlined by its definition, is an intricate phenomenon that encapsulates both social and linguistic dimensions. The computational processing of such language necessitates the deployment of finely-tuned, task-specific language tools and resources. This requirement is particularly prominent for languages such as Serbian, which are morphologically rich, highly inflective, and under-resourced.
In the past, users were usually expected to report abusive speech to the site moderator. It was also often the case that sites used "black" lists to detect and filter the abusive content automatically (Nobata et al., 2016). However, due to the enormous amount of online content generated daily, automatic detection of inappropriate content and even prediction and prevention of flames generation are necessary. The research community supported the initiatives by organizing workshops and tracks on major NLP conferences such as Abusive Language Workshop on ACL 2017 [1], OffensEval on SemEval 2019 (Zampieri et al., 2019b) and 2020 (Zampieri et al., 2020), Toxic spans detection on SemEval 2021 [2], GermEval offensive language detection task (Wiegand et al.,

[1] https://www.aclweb.org/portal/content/1st-workshop-abusive-language-online

[2] https://sites.google.com/view/toxicspans

Proceedings of the 1st International Conference on NLP & AI for Cyber Security, pages 153–163
July 29–30, 2024.

2018b), online sexisms detections SemEval 2023 (Kirk et al., 2023), etc.

Here we present our research on identifying abusive speech in tweets in Serbian language. As a dataset, we used the AbCoSER corpus (Jokić et al., 2021) with the primary focus on detecting whether a tweet contains abusive content or not. To accomplish this task, we employed numerous machine learning algorithms, ranging from traditional machine learning and n-gram features to modern transformer models. The remainder of the paper is structured as follows. Related work is given in Section 2, containing a short overview of the machine learning algorithms and systems used for abusive speech detection. The description of the dataset used in our study is given in Section 3. An overview of the methods used in our research is presented in Section 4. The results of abusive speech detection classification algorithms are presented in Section 5. In Conclusion, we summarize the results of our research and indicate further research directions.

2 Related work

The most common strategy for detecting offensive speech on the Web is to train the system to recognize offensive content, which would then be deleted or forwarded to the site moderators (Zampieri et al., 2019a).

Since the first work on Smokey flame detection system (Spertus, 1997) until nowadays, the majority of the approaches to abusive content detection are based on supervised machine learning. This is done either by using traditional approaches that rely on machine learning models with feature extraction methods, or by applying deep learning architectures that have been predominant in recent years. Some of the systems employ specialized lexica or blacklists of abusive terminology either as the only or as a supplementary tool for the abusive language detection systems in social media (Wiegand et al., 2018a; Chen et al., 2012; Pamungkas et al., 2019; Razavi et al., 2010; Rezvan et al., 2018). These lexica can help to detect explicit swear words and profanities in the text directly (Pedersen, 2019). However, they are not a sufficient resource for hate speech detection.

When building a classifier for abusive speech detection, the researchers usually employed two types of features. The first group of features is based on n-grams, linguistic and syntactic characteristics of text, which are combined with traditional machine learning algorithms like Vowpal Wabbit regression model (Nobata et al., 2016), Logistic Regression classifier (Waseem and Hovy, 2016; Davidson et al. 2017), SVM (Fabio Del Vigna et al., 2017; Malmasi and Zampieri, 2018; Coltekin, 2020). The second group of features relies on word embeddings, obtained by feeding deep neural networks with vast amounts of text data, such as GloVe (Badjatiya et al., 2017), ElMO (Oberstrass et al., 2019) or word2vec (Mitrovic et al., 2019) in combination with gradient boosting decision trees, LSTM or combination of CNN and RNN neural network architectures.

In recent research, transformer based large language models like BERT (Devlin et al., 2018) have been predominantly used for offensive speech detection as they outperform other methods (Zampieri et al. 2019b). In a comparative study of the application the contemporary large language models for the task of offensive language identification (Zampieri et al., 2023), the authors used zero-shot prompting with six models and demonstrated that only Flan-T5 (Chung et al., 2022) reached performance close to but not better than state-of-the-art models from OffensVal competitions. In addition, that was the only model that supported languages other than English.

The first paper dealing with hate speech in Serbian language by Krstev et al. (2007) presented the results of an information search experiment in quest for attacks which are the result of national, racial, or religious hatred and intolerance on a corpus of newspaper articles. The AbCoSER was the first abusive speech dataset in Serbian language (Jokić et al., 2021) presented together with Ontolex lemon lexicon developed to facilitate abusive speech detection.

Vujičić and Mladenović (2023) curated a hate speech lexicon and a dataset in Serbian language to train a classifier for automatic hate speech detection in sports domain. They experimented with BiLSTM deep neural network, and the results showed high precision of detecting Hate Speech in sports domain (96% and 97%) and low recall.

3 Dataset

In this research, we have used the AbCoSER corpus that consists of 6,436 tweets out of which 5,020 with regular speech and 1,416 annotated as abusive speech (Jokić et al., 2021). The AbCoSER corpus contains general abusive speech, meaning

LEVEL A: Abusive speech detection	LEVEL B: Abusive speech category
Abusive (**ABU**): insults, vulgarities, threats, curses, insinuations, irony, sarcasm e.g. "@USER Mnogo ne znaš...Kada neko nema elementarnog znanja, onda je diskusija besmislena. Prijatno."/"@USER You don't know much...When someone does not have elementary knowledge, then the discussion is pointless. Have a good day."	Profanity (**PROF**): the tweet contains simplicity and vulgarity. e.g. "ako ne možeš da mi nabaviš pandu za kućnog ljubimca koji ćeš mi kurac"/"if you can't get me a panda for a pet what the fuck are you going to do"
	Hate speech (**HS**): if a tweet contains an attack, disparagement, or promotion of hatred towards a group of people or members of that group in terms of race, ethnicity, nationality, gender, religion, political orientation, sexual orientation. e.g."Da mi imamo policiju kako treba, ne bi imali migrante. Nijednog. Ali nemamo policiju kako treba. To se vidi."/If we had the police properly, we would not have migrants. No one. But we don't have the police properly. It's obvious."
Not abusive (**NOT**) e.g. "@USER Ne mozes se promeniti, samo prilagoditi 😐😐" /"@USER You can't change, only adapt 😐😐"	Derogatory speech (**DS**): a tweet is used to attack or humiliate an individual or group in a general sense, not like hate speech. e.g."Ne znam sta je neprijatnije: gledati tvoje slike, ili čitati tvoje "tvitove". 😫"/"I don't know what's more unpleasant: looking at your pictures, or reading your "tweets". 😫"
	Other (**OTH**): abusive speech that doesn't belong to the above-mentioned categories e.g., ironic or sarcastic tweets. e.g. "Na izborima bolesni glasaju za bolesne." /"In elections, the sick vote for the sick."

Table 1: The AbCoSER dataset labels with examples.

that it's not prepared with the focus on a specific type of targets such as racial, LGBT or misogyny speech. The corpus resulted from a random sampling of tweets from a timeline of 111 Twitter users, whose profiles were gathered via crowdsourcing and manual search as the ones who are more likely to generate abusive speech. The dataset was annotated by using a hierarchical annotation scheme, similar to Nobata et al. (2016). The scheme is presented in Table 1. In the first level, annotators marked whether a tweet was abusive. On the second level, an abusive tweet was further categorized as profanity, hate speech, derogatory speech, or other. An abusive tweet had to belong to at least one of the categories from the second annotation level. The dataset was annotated by two independent annotators and one resolving annotator. The annotation task was executed manually by a cohort of ten postgraduate students, predominantly holding a degree in Philology. Before the commencement of the task, the annotators were equipped with the training session and annotation guidelines with examples.

Despite careful data collection, the data set was unbalanced, which was reported as one of the major challenges in the similar studies (Zampieri et al., 2019a; Davidson et al., 2017).

In this paper our objective is to detect abusive speech in general, therefore we will focus on binary classification of tweets into two categories – a tweet contains abusive speech, the tweet doesn't contain abusive speech. In addition to the tweet content, tweet number, and class label, the dataset contains additional tweet metadata such as tweet author, number of replies, number of retweets, number of favorites, etc.

4 Research methodology

The pre-processing of text data in our dataset is an important step to make it easier to extract information and apply machine learning algorithms. Twitter data differs significantly from other types of texts, e.g., books or newspaper articles, meaning that there are specific issues that have to be considered when processing non-standard Serbian language present in Twitter (Jokić et al., 2021).

For all the models we applied the following preprocessing steps:

- Alphabets unification to Latin script,
- Mentions, starting with @, were removed as they don't give much information about the content of a tweet,
- Punctuation, such as ", special characters like new line or numbers were removed as well as double spacing,
- Emoticons as well as punctuation representing emoticons were removed;
- In hashtags, sign '#' was removed, and the remaining text left since it could contain useful information about the content;
- The whole text was lowercased to avoid diverse treatment of the same word written in different case or false casing;
- For each model, we performed evaluation with and without restoration of diacritics as described in (Krstev and Stankovic, 2019).

Data pre-processing resulted in 62 empty tweets, mainly those that contained just mentions and emoticons. Those tweets were removed and that resulted in 6,373 tweets in our final dataset, with 4,958 tweets annotated as NOT and 1,416 annotated as ABU.

After these pre-processing steps, we performed tokenization and lemmatization of the text. These steps were executed with classla[3] library for NLP tasks for Slovenian, Croatian, Serbian, Macedonian and Bulgarian languages (Ljubešić and Dobrovoljc, 2019a; Terčon and Ljubešić, 2023). The authors used a big Web corpus when performing training for Serbian language. In our research, we used settings for non-standard Serbian language based on the nature of utterances in the Twitter dataset.

4.1 BoW and tf-idf vector representation

In order to perform classification using machine learning, the pre-processed text needs to be converted into a feature vector representation. One of the basic techniques to get text features is Bag of Words (BoW). The BoW model with unigrams is used as a baseline classification model in our research. Subsequently, we converted text into a document-term matrix to get TF-IDF model. As terms, we tested unigrams, bigrams, combination of unigrams, bigrams and trigrams as well as characters n-grams. The resulting sparse matrix was utilized as input to the selected machine learning algorithms.

We created the Bag-of-Words text representation using sklearn's CountVectorizer function. The parameters were set to leave stop words, to take into account terms that appear at least in 2 documents and to discard terms that appear in more than 95% of documents.

4.2 FastText embeddings as features

FastText embeddings for Serbian (Grave et al., 2018) were used to get averaged fastText embedding of a cleaned tweet and then used as an input for harnessed classification algorithms and neural networks as an input layer.

4.3 Feature set for feature engineering approach

Based on the conducted literature review and the categorization of features provided in (Schmidt and Wiegand, 2017; Nobata et al., 2016; Šandrih, 2020), we selected and implemented a set of 26 features potentially relevant for abusive speech detection.

Simple surface features

These features include bag of words - n-grams of words and characters, tf-idf, frequency of URLs and punctuation marks, text and word lengths, capital letters, unknown words in the dictionary, etc. We used:

Word Count: total number of words in a tweet;

Length: total number of characters in a tweet before data pre-processing;

Number of characters after data pre-processing;

Sentence count total number of sentences in a tweet;

Number of abbreviations used in a tweet;

[3] https://pypi.org/project/classla/

Number of long words might indicate writer skillfulness and education and could be connected with absence of abusive speech. This feature represents the number of words longer than certain threshold (in our study it was set to 11 after experimenting with a few different values);

Number of long sentences, similar to long words, this feature may also indicate higher education level of the tweet author. The value for this feature is calculated as the number of sentences longer than a certain threshold divided by the total number of sentences in a tweet. In our study, this threshold was set to 16 after empirical examination of the impact of different threshold values;

Number of punctuations in tweet text, normalized by the total number of words in the cleaned tweet. Separately, we checked if there are **exclamation marks** and **question marks** in tweets and these two features were of Boolean data type TRUE/FALSE;

Parts of speech count. Following the work of Wassem and Hovy (2016) and Robinson et al. (2018), we counted various parts-of-speech (POS tags): verbs, nouns, adjectives, adverbs and conjunctions. These features were calculated with POS tagger for non-standard Serbian language from the previously mentioned classla[3] library. These values were then normalized by dividing them with the total number of words in a tweet.

Linguistic features

Average word length expressed in number of characters and average sentence length expressed in number of words can be an indicator for a degree of complexity a writer can master;

Upper case words expressed as number of words typed in upper case normalized by total number of words;

Vocabulary related features are included in this study in order to investigate their relatedness to a tweet abusiveness.

Rare words. We assume that rich vocabulary and usage of uncommon words indicate better writing quality and imply regular speech as the opposite to non-standard language. After text cleaning and removing stop words, we took the list of words that appeared only once in corpus (1490) to identify if any of them is present in the tweet. Any rare attribute is binary yes/no attribute;

Unique words on the other hand resulted from tokenizing the text, removing stop words, and counting the number of unique words that are then normalized by the total number of words in the tweet. The larger the unique words feature value, the richer the vocabulary used in a tweet;

Most frequent words is another feature based on BoW and related to vocabulary. We count the number of words in a tweet that are among 100 most frequent words in the corpus.

Metadata includes information about the author of the text (gender, history of hate speech, online activity, etc.) or data pertaining to the tweet. In our research we used the following metadata:

- favorites count: number of times a tweet got favorited;
- retweet count: number of times a tweet got retweeted;
- mentions count: number of other users mentioned in a tweet (@user id); hashtags count: number of hashtags in a tweet.

Lexical features

Hate speech is full of curses and insults, which can be easily recognized with the help of dictionaries and lexica of a general type or specially developed for this purpose (Razavi et al., 2010; ElSherief et al. 2018). A lexical resource was designed to trigger the recognition of abusive language in Serbian and included phrases and figurative speech (Stanković et al, 2020a). This abusive lexicon was further expanded by incorporating a list of abusive triggers, often referred to as a "black words list", and a coarse list obtained via crowdsourcing. The final list was composed of 1,434 unique lemmas.

HateLex feature: This feature corresponds to the number of lemmas from lemmatized tweets that are found in the abusive speech lexicon.

4.4 Prediction models

In this research 19 traditional machine learning algorithms are evaluated such as: SVM, Random Forest, Logistic Regression, Passive-Aggressive Classifier.

With BoW unigram features, the best results were achieved with Stochastic gradient descent configured to work as a logistic regression classifier, which was finally selected as the baseline model. Diacritics restauration and lemmatization didn't improve the results and therefore were omitted.

When experimenting with TF-IDF word and character n-grams as characteristics, the best results were obtained with 3–5-character n-grams with restored diacritics, trained with Passive Aggressive classifier (PAcharacter-ngram classifier in (PAcharacter-ngram classifier in Table

2). The result is in line with Nobata et al. (2016), who got the best results with 3-5 char n-grams among all other features with an F1 macro score of 0.726 and 0.769 for two examined datasets respectively.

When averaged FastText embeddings are used as features, the best result was achieved with K-nearest neighbors' algorithm with 5 neighbors.

The experiments with 26 features dataset were done as well, and here Quadratic Discriminant Analysis Classifier performed the best and without diacritics restauration. A feature selection experiment on the feature set, unsurprisingly resulted in top three features: tweet length, hate_lex and word count as most discriminatory, which corresponds to dataset statistics presented in Jokić et al. (2021).

Besides, we tested the following deep neural networks models:

- Recurrent neural networks (RNN) and their modalities such as LSTM (long-short term memory) and GRU (Gated Recurrent Unit) networks, that are widely used in the area of NLP. Here we leveraged LSTM, biLSTM, GRU, biGRU;
- Convolutional neural networks (CNN) (Kim, 2014; Zhang et. al., 2018) and
- Combination of CNN and RNN models.

The best performing model was biGRU with self-initialized word embeddings with vector dimension 256, 64 neurons in GRU layer, 128 in hidden layer and 1 neuron in output layer. Random input embeddings were additionally trained during the network training. As a regularization technique, dropout (0.5) was applied before each dense layer. Activation function relu was applied in hidden and sigmoid in output layers, having optimizer RMSprop. This configuration resulted in 5,259,905 network parameters. The results of other models were close to the BiGRU best result. Even fast embeddings didn't contribute much more to improve F1 macro score. Due to the specific nature of tweets, it seems that word vectors trained on regular datasets don't contribute much compared to self-initialized embeddings trained on Twitter dataset in question.

A CNN text classification model (Kim, 2014) was constructed with kernel size 5, 128 filters and RMSprop optimizer. These parameters were found by applying RandomizedCV hyperparameter search. The model was trained in 10 epochs, having batch size of 10 samples. We also tested different combinations of CNN and GRU and biGRU networks (Zhang et al., 2018; Mitrovic et al., 2019), with self-initialized and fasttext embeddings. This has recently become a very popular approach where the CNN model serves for feature extraction and the LSTM model for interpreting the features across time steps. Unfortunately, these otherwise promising models didn't perform any better than regular CNN. It might be that we reached top performances with this dataset when CNN was used. That might be due to the size of the dataset since deep learning models require much more training data.

4.5 Transformers architecture

Following the recent advances in deep learning architectures and their application for abusive speech detection and classification problems in general (Zampieri et al., 2023; Batanović, 2020), we evaluated nine transformer models, fine-tuned with annotated data from AbCoSER dataset. The following models were evaluated:

- XLM-T (Barbieri et al., 2022) as a fine-tuned version of XLM-R (Conneau et al., 2020) with millions of tweets in over thirty languages, among them also Serbian, which was the rationale to evaluate this model;
- Multilingual BERT cased (Devlin et al., 2019), which supports 104 languages and was trained on Wikipedia data. The model has 12 layers, while vectors have 768 dimensions and 12 heads. Total number of 110M parameters;
- Multilingual DistilBERT model (Sanh et al. 2019), as a compressed version of BERT, has 6 layers, 768 dimension and 12 heads, totalizing 134M parameters;
- BERTić (Ljubešić et al., 2021), a pre-trained BERT model with 8 billion tokens with text written in Bosnian, Croatian, Montenegrin or Serbian, based on ELEKTRA transformer architecture and with 110M parameters;
- BERTić frank hate model, the fine-tuned BERTić model with FRANK dataset (Ljubešić et al., 2019b) of LGBT and migrant hate speech in Croatian language, which was the ground to test this model;
- XLM-R-BERTić (Ljubešić et al., 2024), bigger XLM-R based model (Conneau et

al., 2020) pre-trained on the same datasets as BERTić;
- Jerteh-81 (Škorić, 2024), based on RoBERTa-base architecture and with 81 million parameters trained with corpuses created and curated by Language Resources and Technologies Society Jerteh[4];
- SRoBERTa-base and SRoBERTa-F, models based on RoBERTa architecture trained on 3GB and 43GB datasets with texts in Serbian and Croatian (Cvejić, 2022).

All the models are fine-tuned for classification task for four epochs (batch size = 8, learning rate = 4e-5), with tweet text retained in original form but with unified alphabet.

We expect that models pre-trained with corpuses in Serbian language will perform better than multilingual large language models trained to support hundreds of languages.

4.6 Evaluation strategy

The dataset was divided into training and testing subsets in a 70:30 ratio, utilizing stratified sampling to guarantee a uniform class distribution in both subsets. Given the imbalanced label distribution, we employed the macro-averaged F1-score for the evaluation and comparison of various model performances. The macro-averaged F1-score, which calculates the average F1 score across all classes, is a commonly used metric in most reference papers on this topic. In addition, we compared the performance of the models against the BoW model and majority class baselines.

5 Results

To get an observable picture of the results, we present results of a dummy (All OFF) classifier that assigns to each record the label of a most frequent class, that could serve as a default baseline model.

Although we executed a vast number of model-classifier experiments, for the sake of the scope of this paper, the results are presented in Table 2 for each type of model together with the best performing classifier as explained in the *Prediction models section*.

The best performing classifier was BERTić that achieved an F1 score of 0.827 and accuracy 0.89. The confusion matrix in Figure 1, depicts better the

System	F1-score	Accuracy
All OFF baseline	0.4375	0.7778
BoW + SGD baseline	0.6190	0.7439
$PA_{character\text{-}ngram}$	0.7124	0.8259
FastText + KNN(5)	0.6076	0.7308
26 features set+QDA	0.6166	0.7091
biGRU	0.6401	0.7731
CNN	0.6489	0.7820
XLM-T	0.7270	0.8230
BERTić	**0.8270**	**0.8900**
$BERTic_{\text{-frenk-hate}}$	0.7760	0.8540
XLM-R-BERTić	0.4380	0.7800
Jerteh-81	0.7480	0.8380
SRoBERTa-base	0.6820	0.7860
SRoBERTa-F	0.7710	0.8540
$BERT_{base\text{-}multiling\text{-}c}$	0.7090	0.8290
$DistilBERT_{base\text{-}multiling\text{-}c}$	0.7150	0.8240

Table 2: Results on the test dataset.

performances of our model, which in 140 out of 425 cases misclassified abusive tweet as non-abusive. Further analysis of the misclassified tweets indicated that the model was not able to recognize:

- subtle language nuances such as "jao nano, kol'ka mu glava" (eng. "oh boy, how big is his head");
- irregular language such as "Du vaj kitu" (eng. „blow the dick" but deliberately misspelled);
- sarcastic implicit insults "Nemaš za terapeuta, ali tu je tviter. Dobro, šta sad" (eng. "You don't have money to pay a therapist, but there's Twitter. Okay, so what now.");
- some explicit insults such as "Nije on misteriozan, nego je glup pa stalno ćuti." (eng. "He is not mysterious, rather, he is stupid, so he keeps silent.").

As for the other transformer models, the performance of the multilingual models such as $BERT_{base\text{-}multiling\text{-}c}$, $DistilBERT_{base\text{-}multiling\text{-}c}$ and even XML-T, which was finetuned with Twitter datasets, were worse than BERTić and comparable to the best traditional $PA_{character\text{-}ngram}$ model. Out

[4] https://jerteh.rs/index.php/en/

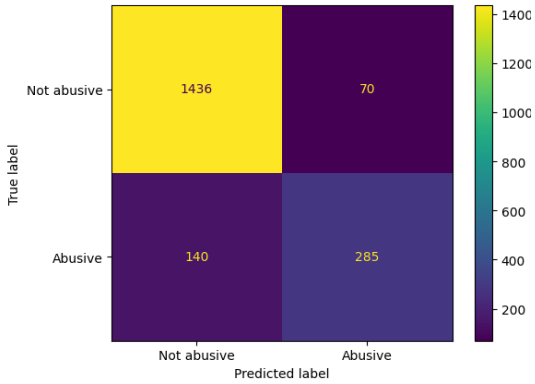

Figure 1: Confusion matrix for the best performing model.

of other models, which were pretrained with datasets only in Serbian or in the regional closely related languages, Jerteh-81, SRoBERTA-F and BERTić$_{\text{-frenk-hate}}$ had better performance than traditional models, still far behind BERTić model. BERTić$_{\text{-frenk-hate}}$, as the only model fine-tuned with hate speech in Croatian, a language close to Serbian, didn't improve the results compared to BERTić, although it was expected as per the research conducted for HateBERT dataset for English (Caselli et al., 2021). The reason might be hate speech domain of FRANK dataset (Ljubešić et al., 2019b).

At the moment there is no benchmark available for the AbCoSER dataset. Therefore, without intention to do a comparison across languages, we compared our results with the benchmarks for offensive speech datasets available in other languages, evaluated with similar methodology in SemEval2019 for English (Zampieri et al., 2019b), and SemEval2020 for multiple languages (Zampieri et al., 2020). As presented in Table 3, it can be observed that our best model achieved the performance comparable to the results on datasets for English (SemEval2019 benchmark), Danish and Turkish (SemEval2020).

6 Conclusion

In this paper, we presented the results of various systems performance on the automated abusive speech detection task in Serbian language. A number of models were evaluated, ranging from traditional ones using BoW, TF-IDF and text features combined with machine learning classifier algorithms, over word embeddings and deep learning architectures, to state-of-the-art transformer models.

	Dataset statistics			
Language	OFF	NOT	Total	F1 score
English	4,640	9,460	14,100	0.8290
Arabic	1,991	8,009	10,000	0.9017
Danish	425	2,865	3,290	0.8119
Greek	2,911	7,376	10,287	0.8522
Turkish	6,847	28,441	35,288	0.8258
Serbian	1,416	5,020	6,436	0.8270

Table 3: Comparing results with other datasets.

By far the best algorithm was obtained by fine-tuning BERTić (Ljubešić et al., 2021) for classification of abusive tweets. The best traditional model in our study was acquired by using TF-IDF 3-5 character n-grams and Passive aggressive classifier on the dataset with restored diacritics. The surprise was the excellent result of the Passive aggressive classifier, which has not been mentioned in relevant literature. Deep learning models had lower performances possibly due to the small size of our dataset for these models.

In future work, we plan to extend the AbCoSER corpus with new tweets and short texts from other sources e.g. online news comments, while addressing the issue of labels imbalance on both annotation levels. In addition, we would focus on application of extra methods for text preprocessing such as conversion of abbreviations and emoticons, application of better lemmatizer for Serbian (Stankovic et al., 2020b), processing of negation in Serbian language (Ljajić and Marovac, 2019) etc. In order to improve the recall rate, which currently stands at 0.6520 for the abusive category, it's important to understand that abusive comments can also include implicit bullying through the use of irony or sarcasm (Dadvar et al., 2013). Therefore, employing a separate classifier, like the one suggested by Mladenovic et al. (2017), specifically trained to detect irony and sarcasm, could prove to be advantageous. Based on error analysis, we envision that a hybrid classification system model which combines traditionally crafted text features, abusive speech lexicon that includes MWEs (Stanković et al., 2020a), with a modern transformer model would provide most robust solution for an abusive speech detection system for Serbian language.

References

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. In Proceedings of the 26th international conference on World Wide Web companion (pp. 759-760).

Barbieri, F., Anke, L. E., & Camacho-Collados, J. (2021). XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 258–266, Marseille, France.

Batanović, V. (2020). Metodologija rešavanja semantičkih problema u obradi kratkih tekstova napisanih na prirodnim jezicima sa ograničenim resursima. Универзитет у Београду.

Caselli, T., Basile, V., Mitrović, J., & Granitzer, M. (2021). HateBERT: Retraining BERT for Abusive Language Detection in English. In Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021) (pp. 17-25).

Caselli, T., Basile, V., Mitrović, J., Kartoziya, I., & Granitzer, M. (2020). I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In Proceedings of The 12th Language Resources and Evaluation Conference (pp. 6193-6202).

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In 2012 international conference on privacy, security, risk and trust and 2012 international conference on social computing (pp. 71-80). IEEE.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... & Wei, J. (2024). Scaling instruction-finetuned language models. Journal of Machine Learning Research, 25(70), 1-53.

Çöltekin, Ç. (2020). A corpus of Turkish offensive language on social media. In Proceedings of the twelfth language resources and evaluation conference (pp. 6174-6184).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzman, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451,Online, July.

Cvejić, A. (2022). "Prepoznavanje imenovanih entiteta u srpskom jeziku pomoću transformer arhitekture." Zbornik radova Fakulteta tehničkih nauka u Novom Sadu 37 (02): 310–315.

Dadvar, M., Trieschnigg, D., Ordelman, R., & De Jong, F. (2013). Improving cyberbullying detection with user context. In Advances in Information Retrieval: 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings 35 (pp. 693-696). Springer Berlin Heidelberg.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In Proceedings of the international AAAI conference on web and social media (Vol. 11, No. 1, pp. 512-515).

Vigna, F.D., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate Me, Hate Me Not: Hate Speech Detection on Facebook. Italian Conference on Cybersecurity.

Kenton, J. D. M. W. C., & Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of naacL-HLT (Vol. 1, p. 2).

ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E. (2018). Hate lingo: A target-based linguistic analysis of hate speech in social media. In Proceedings of the international AAAI conference on web and social media (Vol. 12, No. 1).

Grave, É., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning Word Vectors for 157 Languages. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Hardage, D., & Najafirad, P. (2020). Hate and toxic speech detection in the context of covid-19 pandemic using xai: Ongoing applied research. In Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020.

Jokić, D., Stanković, R., Krstev, C., & Šandrih, B. (2021). A Twitter Corpus and lexicon for abusive speech detection in Serbian. In 3rd Conference on Language, Data and Knowledge (LDK 2021). Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, pages 1746–1751.

Kirk, H., Yin, W., Vidgen, B., & Röttger, P. (2023). SemEval-2023 Task 10: Explainable Detection of Online Sexism. In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023) (pp. 2193-2210).

Krstev, C., Gucul, S., Vitas, D., & Radulović, V. (2007). Can we make the bell ring. In Proceedings of the Workshop on a Common Natural Language

Processing Paradigm for Balkan Languages (pp. 15-22).

Krstev, C., & Stanković, R. (2020). Old or new, we repair, adjust and alter (texts). Infotheca - Journal for Digital Humanities, v. 19, n. 2, p. 61-80.

Ljajić, A., & 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.

Ljubešić, N., & Dobrovoljc, K. (2019a). What does neural bring? Analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. In Proceedings of the 7th workshop on balto-slavic natural language processing (pp. 29-34).

Ljubešić, N., Fišer, D., & Erjavec, T. (2019b). The FRENK datasets of socially unacceptable discourse in Slovene and English. In Text, Speech, and Dialogue: 22nd International Conference, TSD 2019, Ljubljana, Slovenia, September 11–13, 2019, Proceedings 22 (pp. 103-114). Springer International Publishing.

Ljubešić, N., & Lauc, D. (2021). BERTić-The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. In Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing (pp. 37-42).

Ljubešić, N., Suchomel, V., Rupnik, P., Kuzman, T., & van Noord, R. (2024). Language Models on a Diet: Cost-Efficient Development of Encoders for Closely-Related Languages via Additional Pretraining. In Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024 (pp. 189-203).

Malmasi, S., & Zampieri, M. (2018). Challenges in discriminating profanity from hate speech. Journal of Experimental & Theoretical Artificial Intelligence, 30(2), 187-202.

Mitrović, J., Birkeneder, B., & Granitzer, M. (2019). nlpUP at SemEval-2019 task 6: A deep neural language model for offensive language detection. In Proceedings of the 13th international workshop on semantic evaluation (pp. 722-726).

Mladenović, M., Krstev, C., Mitrović, J., & Stanković, R. (2017). Using lexical resources for irony and sarcasm classification. In Proceedings of the 8th Balkan Conference in Informatics (pp. 1-8).

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive language detection in online user content. In Proceedings of the 25th international conference on world wide web (pp. 145-153).

Oberstrass, A., Romberg, J., Stoll, A., & Conrad, S. (2019). HHU at SemEval-2019 Task 6: Context does matter-tackling offensive language identification and categorization with ELMo. In Proceedings of the 13th international workshop on semantic evaluation (pp. 628-634).

Pamungkas, E. W., & Patti, V. (2019). Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop (pp. 363-370).

Pedersen, T. (2019). Duluth at SemEval-2019 Task 6: Lexical Approaches to Identify and Categorize Offensive Tweets. In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 593-599).

Razavi, A. H., Inkpen, D., Uritsky, S., & Matwin, S. (2010). Offensive language detection using multi-level classification. In Advances in Artificial Intelligence: 23rd Canadian Conference on Artificial Intelligence, Canadian AI 2010, Ottawa, Canada, May 31–June 2, 2010. Proceedings 23 (pp. 16-27). Springer Berlin Heidelberg.

Rezvan, M., Shekarpour, S., Balasuriya, L., Thirunarayan, K., Shalin, V. L., & Sheth, A. (2018). A quality type-aware annotated corpus and lexicon for harassment research. In Proceedings of the 10th acm conference on web science (pp. 33-36).

Robinson, D., Zhang, Z., & Tepper, J. (2018). Hate speech detection on twitter: Feature engineering vs feature selection. In The Semantic Web: ESWC 2018 Satellite Events: ESWC 2018 Satellite Events, Heraklion, Crete, Greece, June 3-7, 2018, Revised Selected Papers 15 (pp. 46-49). Springer International Publishing.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In Proceedings of Thirty-third Conference on Neural Information Processing Systems (NIPS2019).

Šandrih, B. (2020). Impact of text classification on natural language processing applications. Универзитет у Београду.

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. In Proceedings of the fifth international workshop on natural language processing for social media (pp. 1-10).

Škorić, M. (2024). Novi jezički modeli za srpski jezik. Infotheca - Journal for Digital Humanities (paper accepted for publishing in in Vol 24, No.1).

Spertus, E. (1997). Smokey: Automatic recognition of hostile messages. In Aaai/iaai (pp. 1058-1065).

Stanković, R., Mitrović, J., Jokić, D., & Krstev, C. (2020a). Multi-word expressions for abusive speech detection in Serbian. In Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons (pp. 74-84).

Stankovic, R., Šandrih, B., Krstev, C., Utvić, M., & Skoric, M. (2020b). Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian. In Proceedings of The 12th Language Resources and Evaluation Conference (pp. 3954-3962).

Terčon, L., & Ljubešić, N. (2023). CLASSLA-Stanza: The next step for linguistic processing of South Slavic Languages. arXiv preprint arXiv:2308.04255.

Vujičić Stanković, S., & Mladenović, M. (2023). An approach to automatic classification of hate speech in sports domain on social media. Journal of Big Data, 10(1), 109.

Waseem, Z., & Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL student research workshop (pp. 88-93).

Wiegand, M., Ruppenhofer, J., Schmidt, A., & Greenberg, C. (2018a). Inducing a lexicon of abusive words–a feature-based approach. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (pp. 1046-1056).

Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018b). Overview of the germeval 2018 shared task on the identification of offensive language.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019a). Predicting the Type and Target of Offensive Posts in Social Media. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 1415-1420).

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019b). SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 75-86).

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., ... & Çöltekin, Ç. (2020). SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In Proceedings of the Fourteenth Workshop on Semantic Evaluation (pp. 1425-1447).

Zampieri, M., Rosenthal, S., Nakov, P., Dmonte, A., & Ranasinghe, T. (2023). OffensEval 2023: Offensive language identification in the age of Large Language Models. Natural Language Engineering, 29(6), 1416-1435.

Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on twitter using a convolution-gru based deep neural network. In The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15 (pp. 745-760). Springer International Publishing.