A Twitter Corpus and Lexicon for Abusive Speech Detection in Serbian

Danka Jokić, Ranka Stanković, Cvetana Krstev, Branislava Šandrih

A Twitter Corpus and Lexicon for Abusive Speech Detection in Serbian | Danka Jokić, Ranka Stanković, Cvetana Krstev, Branislava Šandrih | 3rd Conference on Language, Data and Knowledge (LDK 2021) | 2021 | |

10.4230/OASIcs.LDK.2021.13

http://dr.rgf.bg.ac.rs/s/repo/item/0005133

A Twitter Corpus and Lexicon for Abusive Speech Detection in Serbian

Danka Jokić
University of Belgrade, Serbia

Ranka Stanković
Faculty of Mining and Geology, University of Belgrade, Serbia

Cvetana Krstev
Faculty of Philology, University of Belgrade, Serbia

Branislava Šandrih
Faculty of Philology, University of Belgrade, Serbia

Abstract

Abusive speech in social media, including profanities, derogatory and hate speech, has reached the level of a pandemic. A system that would be able to detect such texts could help in making the Internet and social media a better and more respectful virtual space. Research and commercial application in this area were so far focused mainly on the English language. This paper presents the work on building AbCoSER, the first corpus of abusive speech in Serbian. The corpus consists of 6,436 manually annotated tweets, out of which 1,416 were labelled as tweets using some kind of abusive speech. Those 1,416 tweets were further sub-classified, for instance to those using vulgar, hate speech, derogatory language, etc. In this paper, we explain the process of data acquisition, annotation, and corpus construction. We also discuss the results of an initial analysis of the annotation quality. Finally, we present an abusive speech lexicon structure and its enrichment with abusive triggers extracted from the AbCoSER dataset.

2012 ACM Subject Classification Computing methodologies → Natural language processing

Keywords and phrases abusive language, hate speech, Serbian, Twitter, lexicon, corpus

Digital Object Identifier 10.4230/OASIcs.LDK.2021.13

Funding Linked data development is supported by the COST Action CA18209-NexusLinguarum "European network for Web-centred linguistic data science". Access to SketchEngine and Lexonomy is provided by the ELEXIS project funded by the European Union's Horizon 2020 research and innovation program under grant number 731015.
Branislava Šandrih: COST Action CA18209-NexusLinguarum "European network for Web-centred linguistic data science"

Acknowledgements We would like to acknowledge the team of annotators that provided their time and efforts to help us build AbCoSER v1.0 corpus of abusive speech in Serbian.

1 Introduction

1.1 Motivation and research background

With the development of the Internet and the increasing use of online mass media and social networks, detection of inappropriate content and incitement to violence have gained importance. The concept of abusive speech, in the context of this paper, is an umbrella term for phenomena such as profanities, derogatory, and hate speech. One of the most cited definitions of hate speech comes from John T. Nockleby [44, 4], who perceives hate speech as "any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic".

3rd Conference on Language, Data and Knowledge (LDK 2021).
Editors: Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch; Article No. 13; pp. 13:1–13:17
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

According to a survey from 2014, 60% of Internet users witnessed name-calling, 25% saw that someone was physically threatened, and 24% noticed abuse over a long period [14]. According to more recent research from 2017, two-thirds of Americans stated that they had experienced some kind of harassment on the Internet [39]. Studies also show that 18% of children are involved in cyber-abuse, which leads to serious depression, and even suicide [12]. As far as Serbian law is concerned, any discrimination, endangering security, persecution, insults, and harassment on social networks are punishable [26, 4, 30]. Hate speech and flames are present in Serbian media and public discourse especially towards the LGBT population, Roma people, women and migrants [19].

Hate speech has become a major problem for all types of online platforms where an increasing amount of user-generated content appears: from comments on the web news portals, through social networks, to chats on real-time games [37]. Users are usually expected to report abusive speech, and then the site or social network moderators manually review the report. More advanced platforms use systems with regular expressions and "black" lists of words and expressions, to catch abusive language and remove posts [25]. There are also online portals such as HateBase.org that collect examples of online hate speech in all languages that can be used as trigger words for hate speech detection ([46, 41, 14, 13]). However, detecting hate speech by simply filtering by keywords is not a satisfactory solution, as interpretation can be influenced by the domain of the conversation, the context of the discourse, the objects that accompany the conversation (images, video, and audio materials), the time of publication and ongoing world events and the recipient of the message [38]. Given the huge amount of online material that is created every day, automatic methods are needed to detect and process this type of content.

One of the biggest problems that researchers have to solve before building the automatic hate speech detection systems is finding as many as possible publicly available annotated data sets of a considerable size, especially if the system will be based on deep learning [39]. Another problem researchers face is the non-existence of generally accepted definitions of hate speech and related phenomenon ([14, 38]), which leads to the use of different annotation schemes and categories definitions in various data set making it impossible to compare results of different systems [40]. An additional problem is that the available datasets usually focus on specific topics like misogyny or racist speech and do not cover all types of hate speech. In the last few years, hate speech has gained more attention from the research community, which led to the organization of several workshops, both independently or at international conferences that address problems of hate speech and related topics such as GermEval2018, Offenseval2019 and Offenseval 2020 ([47, 51, 52]).

Abusive language and its detection have also gained more attention recently. Casseli et al. [6] define 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." From the definition itself, it is evident that abusive speech is a complex social and linguistic phenomenon [42]. Computational processing of such language requires the usage of finely-tuned, task-specific language tools and resources, especially for morphologically rich and low-resource languages such as Serbian. The main contribution of this work is the creation of the AbCoSER, the first abusive speech corpus in Serbian, that will, together with abusive speech lexicon, enable the development of automatic abusive speech detection systems for the Serbian language. In the course of this work, we leveraged existing annotation schemes and abusive term definitions as much as possible with the aim of creating a general data set convenient for the detection of a broad range of abusive topics. We already used this resource for the detection of abusive triggers and the augmentation of the abusive language lexicon.

1.2 Related work

In the past two decades, several methods and models for the detection of hate speech, abusive speech, toxic comments, and aggression on the Internet have been presented. From the natural language processing (NLP) perspective, the detection of hate speech can be viewed as a problem of classification: for a given statement, the system needs to determine whether it contains hate speech or not [36]. To achieve this goal systems usually apply text mining techniques. The majority of current hate speech, offensive, and abusive language detection systems in social media are based on lexicons or blacklists ([7, 10, 28, 34]). Their importance lies in the fact that a vast number of swear words and offences can be detected by using only lexicons. Razvan et al. [35], created an offensive word lexicon and then collected Twitter messages that contain at least one word from it. They concluded that the presence of a word in a tweet just indicates the possibility of offensive speech, and manual annotation is necessary to guarantee accurate tweets classification. The same lexicon was used in [48] to extract toxic conversations among adolescents on Twitter. While Pedersen [32] reported high accuracy of hate speech detection when using only a lexicon, the accuracy and F1 score were still lower compared to the state of the art [52] and the number of false positives was high, indicating that lexicons are not a sufficient resource for hate speech detection.

High-quality corpora of hate speech, offensive speech, and abusive language are very important as a first step in building an automated system for the detection of these phenomena ([51, 52, 1, 6]). Warner and Hirschberg [44] presented their research on hate speech toward minority groups in online text, with the main focus on anti-semitic language. Three annotators manually annotated a corpus of 1000 paragraphs taken from offensive websites and Yahoo user comments, with Fleiss kappa inter-annotator agreement at 0.63. Waseem and Hovy [45] created the renowned corpora of hate speech, consisting of 16,000 tweets grouped into 3 categories: racist (racism), sexist (sexism), neither. The corpus was built using bootstrapping, an iterative keyword search method. They created a decision-making list for the annotators, which helped them achieve an agreement score $\kappa = 0.84$ among the annotators. One of the most cited papers in this field was written by Nobata et al. [25]. They worked on several data sets consisting of comments from Yahoo Finance and Yahoo News pages. They performed an annotation experiment giving the same data set to trained users and untrained raters on Amazon's Mechanical Turk crowdsourcing platform and showed that better inter-annotator agreement was achieved by trained internal annotators. For binary categorization ("Clean" vs. "Abusive"), trained raters achieved an agreement rate 0.922 and Fleiss's Kappa 0.843 while Turkers agreement rate was 0.867 and Fleiss's Kappa 0.401. The agreement rate decreased when annotating abusive speech subcategories to 0.603 and 0.405 respectively. Davidson et al. [13] created a corpus of around 25,000 tweets with the idea of separating hate speech from other offensive speech. They used three categories to annotate the corpus: hate speech, just offensive speech, and neither using crowd-sourcing. The inter-annotator agreement score was 92%.

In the last few years, various data sets with multi-layered annotation appeared, which enabled both coarse and fine-grained classification. While Wiegand et al. [47] in the second layer classifies the type of insult (vulgar speech, insult, attack), Zampieri et al. [50] and Fisher et al. [15] emphasize the type and the target of offensive speech. The OLID data set and the scheme proposed by Zampieri et al. [50], used also at the SemEval2019 and SemEval 2020 competitions, gained popularity among researchers leading to the production of Turkish and Danish data sets that use this scheme ([11, 40]), as well as two new datasets, AbusEval and SWAD, which improve or use this data set to annotate a new one ([6, 27]). The multi-level universal annotation scheme, which includes the target of hatred or a type of abusive speech, has many advantages. First of all, the classification can be done in several steps. Traditional machine learning can also be used at every step, which in the case of a smaller number of class examples gives better results than deep learning [31]. Another advantage is a simpler structure of the annotation decision tree, which can contribute to a better annotator agreement (the difference between the levels is clearer). The main advantage is that the same scheme can be used for general-purpose hate speech corpora, which includes several types of hate speech, and for specific corpora, which usually cover only one type of hate speech (racial hatred, misogyny, hatred of migrants, etc.).

The first system that dealt with hate speech detection in the Serbian language was described in [18]. The aim of this system was to detect newspaper articles that report on attacks and improper behaviour that are the result of national, racial, or religious hatred and intolerance. The system relied on electronic dictionaries of Serbian and local grammars that covered various patterns of hate speech and ways they were covered in newspaper articles. It should be noted that the focus of this research was different from hate speech detection today, as today's systems mostly deal with heath-speech directly (as found in user-generated content) and not with reports about it.

1.3 Paper outline

We describe in this paper the process of building the first data set of abusive language in Serbian. As the data source, we used tweets from the Twitter social network. Tweets from user timelines of 111 Twitter accounts were gathered and annotated by ten independent annotators working in pairs so that each tweet was annotated by two independent annotators while one supervising annotator resolved inconsistent annotations. Related work is given in Section 1.2, with a short overview of different approaches for developing an abusive speech data set. Our work in acquiring and annotating the data set, including a description of the annotation manual, is presented in Section 2. In Section 2.1, we describe the process of building the data-set of abusive language in Serbian. Further details about manual data annotation of corpus data are given in Section 2.2 while the extraction of abusive triggers is explained in Section 2.3. The Section 3 presents results of our research: Twitter data analysis (Subsection 3.1), the outcome of the annotation (Subsection 3.2) and the structure of the lexicon of abusive words (Subsection 3.3). We summarize the results of our research and indicate further research In discussion and conclusion Section 4.

2 Data Acquisition and Annotation

2.1 Data collection

When deciding which approach should we take when building the corpus of abusive language, several future implications were considered: 1) To the best of our knowledge, AbCoSER (Abusive Corpus for SERBIAN) is the first corpus tackling abusive language phenomenon in the Serbian language; 2) This corpus is to be used to enrich our lexicon of hate speech as described in [42]; 3) Classifiers trained on corpora containing general abusive speech, can be used to classify a domain hate speech corpus, while domain-specific classifiers perform poorly on the general data set and corpora from other hate speech domains ([46, 29]); therefore, instead of investigating domain-specific abusive speech, the phenomenon should be considered in a broader sense. Here we investigate abusive speech that covers vulgar speech, hate speech, and derogatory speech.

It is estimated that 2 to 3% of user-generated content contains abusive language. This means that the number of offensive messages is much smaller than non-offensive ones [38], which would impose practical problems on data annotators as well as automatic detection systems. To overcome this problem, researchers resort to searching by keywords or hashtags ([45, 50, 39, 36]), collecting comments directed to standard targets of abusive speech, or collecting comments from users who are notorious for using offensive language in their writing [47]. However, these approaches introduce bias into the data. While the keyword approach seems to be biased towards explicit expressions of abuse words used in the search [5], with the user timeline collection approach one must be careful when training the classifier so that it does not learn the writing style of a user instead of learning to detect the abusive messages [47]. In this work, the combined methodology is used with an iterative approach aiming at gathering as much abusive messages as possible. Twitter was used as a source for our data collection since it contains a much higher proportion of offensive language than other social networks [47]. Although a random sample of tweets would probably represent the unbiased set of data, we opted to sample tweets from the timeline of numerous Twitter user accounts. When sampling tweets from Twitter, we also imposed certain formal restrictions on the tweets to be extracted similar to those listed in [47]. They are as follows: 1) Each tweet had to be written in Serbian, 2) No tweet was allowed to contain any URLs, 3) No tweet was allowed to be a retweet.

At first, we started with the list of 80 user accounts gathered via crowd-sourcing. To this list, we added various users accounts whose tweets were reported as hate speech on the h8index,[1] an online platform for reporting hate speech, verbal violence, bullying, and discrimination on the Web. Initially, we gathered 450,000 tweets from the timeline of 120 user accounts via Twitter API[2] that were further cleaned by removing tweets that were retweeted from other users timeline and tweets containing URLs, leaving 150,000 tweets in the list. Although each tweet has a language column, in the majority of cases language of Serbian tweets was marked as *und* – unidentified – since Twitter cannot reliably recognize the Serbian language. For example, out of 150,000 tweets, only 8,000 were marked as tweets in Serbian, while 120,000 were marked as *und*. Therefore, we could not use this feature to filter tweets written in Serbian and have to rely on manual annotation.

In order to check how representative our data set is we sampled 200 tweets from it. Still, the ratio of tweets with abusive speech was just 12%. Therefore, the users' list was manually checked for the type of users and users were removed that are less likely to generate abusive speech such as:

1. Public users, like telecommunication and similar companies as well as newspapers and news portals were removed from the list of users since one cannot expect that this type of users will generate abusive speech, as proved in [11].
2. Fan pages and official pages of public persons, including politicians and sportsmen, or political parties were removed from the list for the same reason.
3. Users that tweet in a foreign language.
4. Users that do not generate abusive speech were detected by inspecting their timeline.

Thereafter, an initial list of a few seed words was identified, and Twitter was searched for occurrences of those words. We did not just add those tweets to our data set, we rather identified users that created those tweets and added them to the user list. The reason for

[1] https://h8index.org/

[2] https://developer.twitter.com/en/docs/twitter-api

such an approach was to retain the variety of offensive terms occurring in the collected tweets ([47, 6]). Finally, their followers and those who reply to abusive tweets were added to the list as well. At the end of this step, we extracted the timeline of 111 users, and we come up with 320,440 tweets. The next step was to remove duplicates, empty tweets, retweets, tweets with URLs, and tweets that contained just mentions. The list was reduced to 194,348 tweets. From this corpus, we randomly sampled 6,500 tweets. In the next step, we identified tweets written potentially in English by filtering out tweets whose language was marked with “en” (112 tweets). This set was manually checked and 64 English tweets were removed. The remaining 48 tweets were wrongly marked as written in English, while actually written in Serbian. The resulting data set had 6,436 tweets and this set was used for annotation.

Tweeter 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 such data. Some of them are:

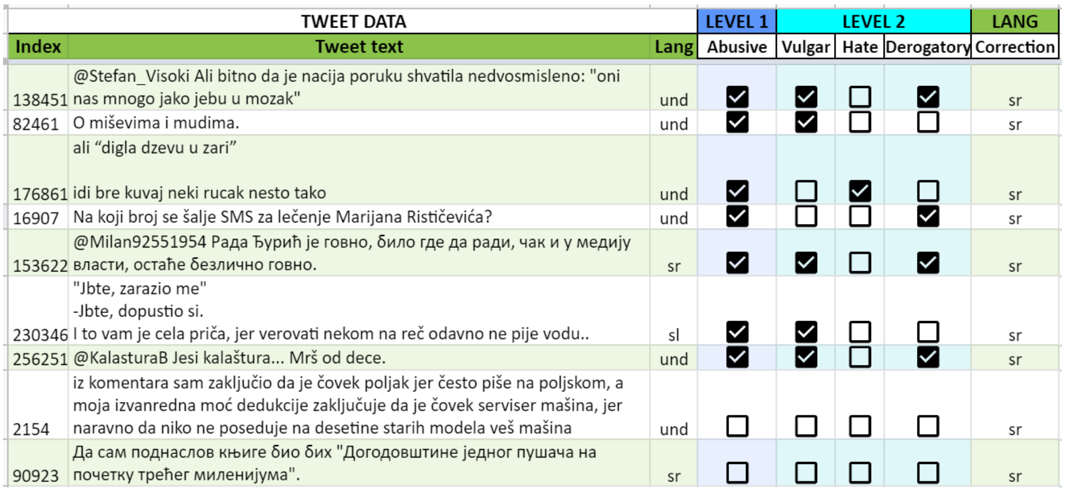
1. Spelling, grammar and typing errors and regional variations are more frequent;
2. Frequent use of out of vocabulary words or intentionally misspelled words (e.g. fejv, lajna, QURAZ);
3. Excessive use of abbreviations, e.g. *nznm* (eng. I don’t know), *mupm* (eng. go fuck your mum), *np* (eng. no problem), *jbt* (eng. fuck you), *jbg* (eng. fuck it) etc.;
4. Equal use of Cyrillic and Latin script, omission of diacritics, and different Unicode characters;
5. Use of foreign language words and emoticons (e.g. `:’-),:-P, :@)`);
6. Twitter-specific text: mentions, retweets and URLs as well as hashtags (e.g. #TLZP, #Utisak, #u6reci).

2.2 General Corpus annotation for classification of tweets

The reliability of the corpus annotation is key to the successful training of an automatic hate speech detection system [36]. Since a set of data annotated by only one person can be biased because it reflects his (or her) personal opinion [3], we decided each tweet be annotated by two independent annotators. Ten annotators participated in the annotation and therefore the data set was split into five parts and each part was annotated by an annotator pair. All annotators were native Serbian speakers.

To obtain a successful annotation scheme, it is important to satisfy the following criteria ([15, 11]): 1) The annotation scheme enables coarse and fine-grained classification; 2) The annotation scheme is accompanied by a detailed annotation guide; 3) The choice of classes corresponds to the expected use of data. Since our primary goal was to investigate abusive speech and its sub-categories, as well as to explore the possibility of distinguishing them, we decided to define a hierarchical two-layer annotation scheme and categories, which is similar to one developed by Nobata et al. [25] and in adherence to the annotation guidelines resulted from the research of Fortuna et al. [16]. The annotation was performed considering the content of a whole tweet, as was the case for the majority of data sets ([47, 23, 50, 49, 36]). At the first level annotators marked a tweet as abusive (TRUE) or non-abusive (FALSE). At the second level annotators determined the category of abusive speech in tweets marked as abusive:

1. Profanity (PROF), the tweet contains simplicity and vulgarity (e.g. “lakše se kenja i preti iz anonimnosti...”/“it’s easier to talk shit and threaten out of anonymity ...”);
2. Hate speech (HS), if a tweet contains an attack, disparagement, or promotion of hatred towards a group of people or members of that group on the basis of race, ethnicity, nationality, gender, religion, political orientation, sexual orientation (e.g. “@USER Što se mene tiče ne trebate nam. Iz Crne Gore dolaze mafijaški klanovi. Nismo mi vama poslali



TWEET DATA			LEVEL 1	LEVEL 2			LANG
Index	Tweet text	Lang	Abusive	Vulgar	Hate	Derogatory	Correction
138451	@Stefan_Visoki Ali bitno da je nacija poruku shvatila nedvosmisleno: "oni nas mnogo jako jebu u mozak"	und	☑	☑	□	☑	sr
82461	O miševima i mudima.	und	☑	☑	□	□	sr
176861	ali “digla dzevu u zari” idi bre kuvaj neki rucak nesto tako	und	☑	□	☑	□	sr
16907	Na koji broj se šalje SMS za lečenje Marijana Rističevića?	und	☑	□	□	☑	sr
153622	@Milan92551954 Рада Ђурић је говно, било где да ради, чак и у медију власти, остаће безлично говно.	sr	☑	☑	□	☑	sr
230346	"Jbte, zarazio me" -Jbte, dopustio si. I to vam je cela priča, jer verovati nekom na reč odavno ne pije vodu..	sl	☑	☑	□	□	sr
256251	@KalasturaB Jesi kalaštura... Mrš od dece.	und	☑	☑	□	☑	sr
2154	iz komentara sam zaključio da je čovek poljak jer često piše na poljskom, a moja izvanredna moć dedukcije zaključuje da je čovek serviser mašina, jer naravno da niko ne poseduje na desetine starih modela veš mašina	und	□	□	□	□	sr
90923	Да сам поднаслов књиге био бих "Догодовштине једног пушача на почетку трећег миленијума".	sr	□	□	□	□	sr

Figure 1 The annotation interface with examples.

mafijaše. Da nije Srba u Crnoj Gori ta država bi mi bila draga koliko i Hrvatska.” / “@USER As far as I'm concerned we don't need you. Mafia clans come from Montenegro. We didn't send you mobsters. If there were no Serbs in Montenegro, that country would be as dear to me as Croatia.”);

3. Derogatory speech (DS), a tweet is used to attack or humiliate an individual or group in a general sense, not in a way hate speech does (e.g. “@USER To je jedna budala, ne veruj mu ništa.”/“@USER He's a fool, don't believe him.”)
4. Other (OTH), abusive speech that does not belong to the above-mentioned categories e.g. ironic or sarcastic tweets.

An abusive tweet belongs to at least one of the categories from the second annotation level. An example of a tweet that belongs to both PROF and DR category is “@USER NAME je govno, bilo gde da radi, čak i u mediju vlasti, ostaće bezlično govno” (eng. “NAME is shit, wherever he/she works, even in the government media, he/she will remain an impersonal shit”).

All annotators were provided with training and annotation guidelines containing examples similar to ([33, 25]). For each of the category, annotators obtained its definition, some examples and an indicative list of trigger words characteristic to it, as described in [42]. Besides annotation guidelines, annotators received the decision list for abusive speech identification similar to the one used in [45], but upgraded and adopted for the general abusive speech. Since Twitter does not identify Serbian as a language successfully, and thus the language column of a tweet could not be relied upon, the annotators were given one more task – to check the language of a tweet and whether it could be interpreted. They needed to mark tweets with meaningless content, tweets written in a foreign language or multilingual tweets. After annotating the initial set of 200 tweets, an additional workshop with annotators was conducted to comment on the first annotation results and discuss discovered problems. Annotation was done online using Google sheets,as presented in Figure 1.

As annotation guidelines in the form of the decision tree are proven to be good a practice ([45, 6, 23]), we prepared the guidelines for annotators in the same format (shown in Figure 2). One can see from the decision tree that a tweet marked as abusive has to be tested for each subcategory, since, as mentioned earlier, one tweet can belong to one or more subcategories. Annotators had a tab in their annotation interface with examples of all possible annotation combinations.

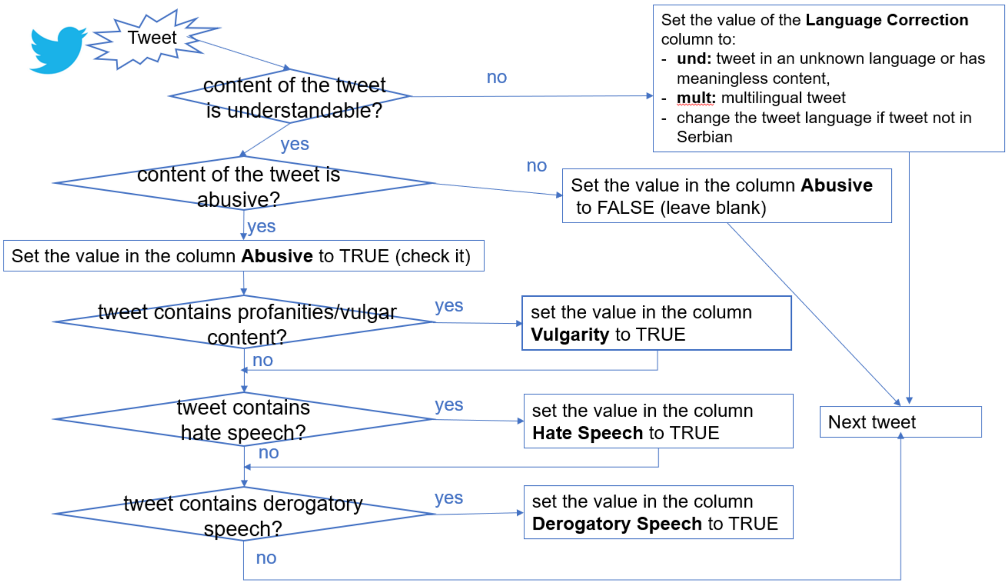
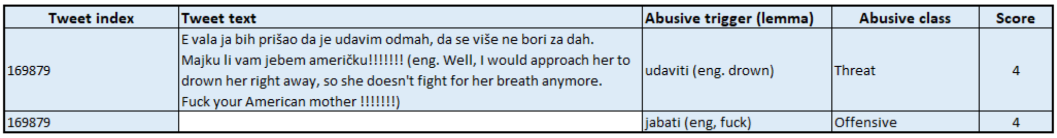

Tweet
content of the tweet is understandable?
no
Set the value of the **Language Correction** column to:
- **und:** tweet in an unknown language or has meaningless content,
- **mult:** multilingual tweet
- change the tweet language if tweet not in Serbian
yes
content of the tweet is abusive?
no
Set the value in the column **Abusive** to FALSE (leave blank)
yes
Set the value in the column **Abusive** to TRUE (check it)
tweet contains profanities/vulgar content?
yes
set the value in the column **Vulgarity** to TRUE
no
tweet contains hate speech?
yes
set the value in the column **Hate Speech** to TRUE
no
Next tweet
tweet contains derogatory speech?
yes
set the value in the column **Derogatory Speech** to TRUE
no

Figure 2 Annotation guidelines in the form of the decision tree.



Tweet index	Tweet text	Abusive trigger (lemma)	Abusive class	Score
169879	E vala ja bih prišao da je udavim odmah, da se više ne bori za dah. Majku li vam jebem američku!!!!!!! (eng. Well, I would approach her to drown her right away, so she doesn't fight for her breath anymore. Fuck your American mother !!!!!!!)	udaviti (eng. drown)	Threat	4
169879		jabati (eng, fuck)	Offensive	4

Figure 3 Abusive span annotation entries for one tweet example.

2.3 Annotation of abusive token spans for lexicon building

Several systems for the detection of abusive language as well as data sets were developed for English ([50, 13, 49, 45]), German [46], Slovene [23], Danish [40] and Turkish [11] that classify whole documents or comments/tweets as abusive without identifying which token spans were abusive. It would be very useful to have those abusive triggers highlighted so that human moderators can react timely to the abusive content. Following the Toxic Spans Detection task on Semeval 2021, and in line with our goal to enrich our lexicon of abusive speech with new entries and the usage examples [42], the additional manual annotation was conducted on 1,564 tweets that at least one of the annotators marked as abusive. This set of tweets was divided and shared among the already trained annotators with a task to detect triggers in each of the tweets. The annotators were given oral and written instructions together with examples of abusive speech categories and respective triggers as discussed in [42]. Figure 3 presents an abusive tweet in which two triggers were identified, classified into different categories of abusiveness and assigned abusiveness score. The purpose of identifying abusive triggers is to use them to enrich the lexicon of abusive words whose structure is presented in Section 3.3.

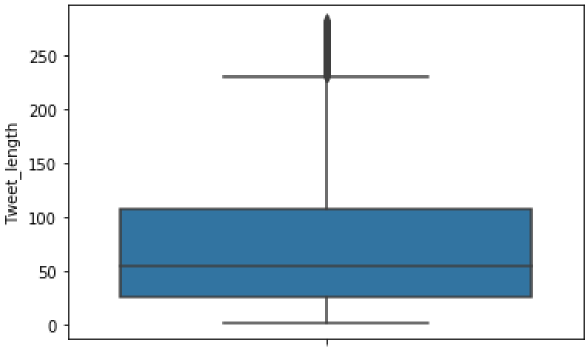


Figure 4 The distribution of tweets length.

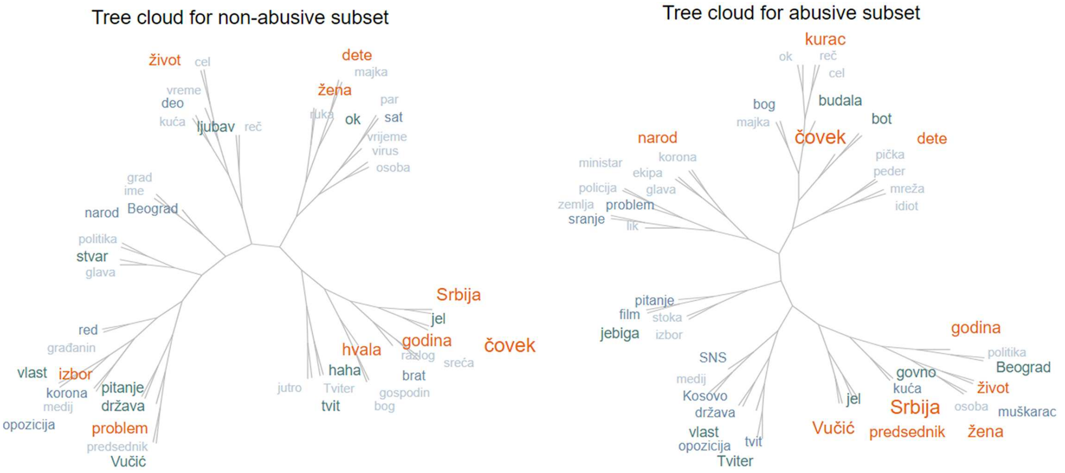


Figure 5 Tree clouds of abusive language corpus: the non-abusive subset (left) and abusive subset (right).

3 A Corpus and Lexicon for Abusive Speech

3.1 Analysis of the twitter corpus

The distribution of our corpus tweets length after removal of mentions is shown in Figure 4: the median value is 54 characters and the mean value is 78.56. As explained in Subsection 2.1, the corpus of tweets needs further pre-processing. First, all Cyrillic characters were replaced with corresponding Latin script characters[3], then punctuation, special and non-printable Unicode characters were also removed, and hash sign # deleted from hashtags. In the end, Tweet tokenizer from Python nltk[4] tokenizes the tweets removing at the same time mentions and an excessive number of repeated characters in tokens.

After data pre-processing, the data visualization technique is applied to cleaned tweets corpus to gain insights into data content. The tree cloud model [17] was employed for data visualization of non-abusive and abusive subsets of data (Figure 5). Besides depicting more frequent words in a larger font, words are also arranged on a tree in a way that reflects their

[3] cyrtransli Python library is available at `https://pypi.org/project/cyrtranslit/`.
[4] `https://www.nltk.org` Python Natural Language Toolkit

Table 1 The inter-annotator agreement per categories of abusive speech.

Category/Subcategory of hate speech	**The inter-annotator agreement, accuracy**
Offensive/Non-offensive	$\kappa = 0.513$, $accuracy = 0.860$
Profanities	$\kappa = 0.612$, $accuracy = 0.956$
Hate speech	$\kappa = 0.263$, $accuracy = 0.949$
Derogatory speech	$\kappa = 0.370$, $accuracy = 0.895$

semantic proximity in the text. We are interested in the most common words, as well as hashtags, in the data sets for both labels (abusive and non-abusive). We can notice that the most frequently used words in the non-abusive dataset are rather neutral words such as *hvala* (eng. thank you), *ljudi* (eng. people), *godina* (eng. year), *problem* (eng. problem), *pitanje* (eng. question), etc. Word *korona* appears among the top 15 words, which is due to the current Covid-19 pandemic. We assumed that our data set might be biased considering the period of data capturing and that proved to be true. In the non-abusive set, many high-frequency words referring to Serbian authority and state politics occur. No abusive word was identified among the top 50 words of this subset.

On the other hand, when we looked at the top 50 words in the abusive speech subset, we noticed that it contained a number of derogatory and vulgar words such as different forms of *jebati* (to fuck), *peder* (gay) *kurac* (dick), *budala* (fool), *govna* (shit), etc. High on the list of the most frequent words are also *Srbija* (Serbia), *sns* (the acronym of *Srpska Napredna Stranka*, Serbian ruling political party), and *Vučić* (the president of Serbia). Among the top 15 words on the list is also *žena* (woman).

We also made a hypothesis that hashtags can be indicators of abusive tweets. Therefore we looked at hashtags used in non-abusive and abusive tweets to check whether there is some pattern of their usage. Two lists of hashtags were created from the raw tweet data and compared with the frequency of hashtag appearance. The distributions of hashtags of non-abusive and abusive tweets were analysed, but we could not confirm our hypothesis because the number of tweets with hashtags was rather low – only 110 non-abusive and 24 abusive tweets contained hashtags.

3.2 Statistics and availability of the corpus

As a measure of inter-annotator agreement, we used Cohen's Kappa coefficient. The results for each of the categories are presented in Table 1. Cohen's Kappa score for the binary annotation Offensive/Non-offensive speech equals 0.513. When further analyzing the results, we noted that the best agreement was achieved with the annotation of profanities ($kappa = 0.612$), while the worst results were for the hate speech ($kappa = 0.263$). Since this level of agreement was not satisfactory, one of the authors of this paper acted as the 3rd supervising annotator, whose task was to resolve annotations on which the first round annotators disagreed. In total, 2,185 differences were identified and harmonized at both annotation levels and the decision was made for all of them.

The resulting data set has in total 1,416 tweets labelled as abusive, out of a total of 6,436 tweets in the data set. 472 tweets are marked as PROF, 273 as HS, 843 as DS, and 169 as OTH. 637 tweets are assigned to more than one abusive category. We are currently working on expanding the data set with additional tweets after which it will be made publicly available.

3.3 The lexicon of abusive speech

The lexicon of abusive speech, consisting of words that could be used as triggers for the recognition of abusive language is being built, with the idea that the Serbian system for the recognition and normalization of abusive expressions will also take into consideration phrases and figurative speech as indicators.

In addition to the improved version of Hurtlex [2], resources that can be useful for the creation of a lexicon of offensive words are lists of swear words, curses, abusive expressions, existing general dictionaries, slang dictionaries, surveys and contributions through crowdsourcing, translation of dictionaries and lexicons from other languages, lexicons of sentiment words and expressions, rhetorical figures, etc. To expand the dictionary, synsets from the Serbian WordNet and the dictionary of synonyms will be used for linking with Twitter examples.

Regarding the categorization of terms in the lexicon, the Hatebase scheme[5] was used as a guideline because it is already a kind of a standard in this area, and then supplemented with additional categories according to hate targets as presented in [41], namely *Category* can be Race, Behavior, Physical, Sexual orientation, Class, Gender, Ethnicity, Disability, Religion, Other. A certain term in the lexicon can be assigned several categories, in case it appears in the context of several types of hate speech. The *Severity* attribute has values within a range 1–5 that represent the degree of insult that can be assigned or automatically calculated from the annotated data set presented in Section 2.3. The value *OffensLevel* is assigned as a measure of a chance that the word is used in the offensive meaning and it is calculated based on the number of different meanings in the comprehensive explanatory dictionary of Serbian, and need to match neither corpus nor probability of use. An excerpt from the dictionary for the word *lopov* (thief) is presented in Listing 1. It can be seen that this word can be used to refer to immoral or criminal activities or as a derogatory word to insult someone.

Information integration beyond the level of dictionaries and across the language resource community has become an important concern. The most promising technology for information integration is the Linked (Open) Data (LOD) paradigm that is used for publishing lexical resources by using URIs to unambiguously identify lexical entries, their components and their relations in the web of data. Moreover, it is used to make lexical data sets accessible via http(s), to publish them in accordance with W3C-standards such as RDF and SPARQL, and to provide links between lexical data sets and other LOD resources [8].

The goal of our research is to make its results compatible with the Linked Data approach, using its set of design principles for sharing machine-readable interlinked data on the Web. This vision of globally accessible and linked data on the internet is based on RDF standards for semantic web, using RDF serialisation for data representation. To that end, our approach envisages export of trigger words as lexical data in RDF that is compliant with the *The OntoLex Lemon Lexicography Module*[6], lexicog [5], as an extension of Lexicon Model for Ontologies (lemon)[7] [24]. This is also in line with activities within NexusLinguarum COST action[8], which promotes synergies across Europe between linguists, computer scientists, terminologists, language professionals, and other stakeholders in industry and society in

[5] https://hatebase.org/ The world's largest structured repository of regionalized, multilingual hate speech

[6] https://www.w3.org/2019/09/lexicog/

[7] https://www.w3.org/2016/05/ontolex/

[8] https://nexuslinguarum.eu/

Listing 1 An excerpt from the XML version of the dictionary.

```
<LexicalEntry id="SR0001" lng="sr" pos="n" Probability="0.8">
  <lemma>lopov</lemma>
  <OffensCategories>
    <OffensCategory Severity="4" OffensLevel="0.7">
      <Examples type="Immoral or criminal activities">
        <Example beginIndex="0" endIndex="6" form="lopovu">
                  lopovu je mesto u zatvoru</Example>
        <Example beginIndex="19" endIndex="25" form="lopovi">
                  svi političari su lopovi</Example>
      </Examples>
    </OffensCategory>
    <OffensCategory Severity="3" OffensLevel="0.4">
      <Examples type="Derogatory words and insults">
        <Example beginIndex="12" endIndex="17" form="lopov"
              type="MWU">ružan kao lopov</Example>
      </Examples>
    </OffensCategory>
  </OffensCategories>
</LexicalEntry>
```

Listing 2 An excerpt from RDF version of dictionary.

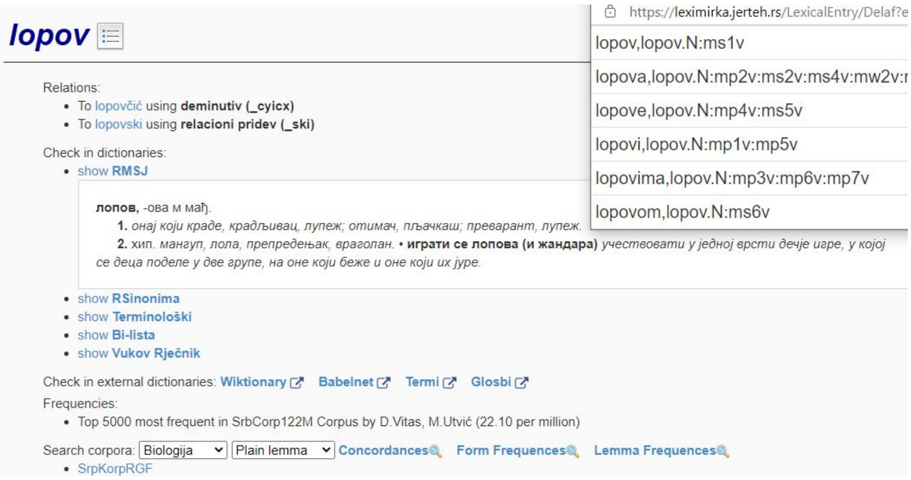
```
:lopov a ontolex:lopov;
   dct:language <http://id.loc.gov/vocabulary/iso639-1/sr> ;
   lexinfo:partOfSpeech lexinfo:noun;
   ontolex:lexicalForm :lopov-form;
   ontolex:sense :lopov-sense.
:lopov-form a ontolex:Form;
   ontolex:writtenRep "lopov"@sr.
:lopov-sense skos:definition "onaj koji krade, kradljivac,
   lupež; otimač, pljačkaš; prevarant, lupež"@sr;
   ontolex:reference <https://www.wikidata.org/wiki/Q3562775>.
```

order to investigate and extend the area of linguistic data science. As an illustration, the RDF model in Turtle syntax[9] is presented in Listing 2, using the same word *lopov* (thief) as an example.

In addition, the usage of the novel module for frequency, attestation and corpus information for Ontolex Lemon (FrAC) [9] is developed. Our intention is to select trigger words that can be found in the corpus AbCoSER and to link usage samples to actual tweets. Lexical variants of trigger words were also included, which is especially important in this case because Twitter users tend to use many irregular forms. Since Serbian is a highly inflective and morphologically rich language that uses a lot of different word suffixes to express different grammatical, syntactic, or semantic features, we also established the relation with the Serbian electronic dictionaries and the management platform Leximirka (Figure 6) [22], which enables the recognition of all inflected forms of trigger words.

For the ranking and selection of illustrative tweets (or its parts) as a kind of dictionary usage examples, we have used a weighted score derived from lexical, word-based and other features (e.g., sentence length, number of all no space chars, digits, weird chars, commas, full

[9] https://www.w3.org/TR/turtle/



https://leximirka.jerteh.rs/LexicalEntry/Delaf?e

lopov,lopov.N:ms1v
lopova,lopov.N:mp2v:ms2v:ms4v:mw2v:r
lopove,lopov.N:mp4v:ms5v
lopovi,lopov.N:mp1v:mp5v
lopovima,lopov.N:mp3v:mp6v:mp7v
lopovom,lopov.N:ms6v

lopov

Relations:
- To lopovčić using **deminutiv (_cyicx)**
- To lopovski using **relacioni pridev (_ski)**

Check in dictionaries:
- show **RMSJ**

лопов, -ова м мађ.
1. *онај који краде, крадљивац, лупеж; отимач, пљачкаш; преварант, лупеж.*
2. хип. *мангуп, лола, препредењак, враголан.* • **играти се лопова (и жандара)** *учествовати у једној врсти дечје игре, у којој се деца поделе у две групе, на оне који беже и оне који их јуре.*

- show **RSinonima**
- show **Terminološki**
- show **Bi-lista**
- show **Vukov Rječnik**

Check in external dictionaries: **Wiktionary** **Babelnet** **Termi** **Glosbi**

Frequencies:
- Top 5000 most frequent in SrbCorp122M Corpus by D.Vitas, M.Utvić (22.10 per million)

Search corpora: Biologija Plain lemma **Concordances** **Form Frequences** **Lemma Frequences**
- SrpKorpRGF

Figure 6 The Leximirka application for lexical database management and use.

stops, punctuation, number of all tokens, average token length, max token length). We use this score to rank examples, but system allows a different number of examples for a different purpose. For example, for dict2wec [43] a larger number of examples will be provided.

The relative frequency (normalized per million) was assigned to lexical entries both for the abusive language (derived from the abusive tweet corpus) and for neutral language (derived from the corpus of non-abusive tweets), which enables calculation of the so-called keyness score, which should represent the extent of the frequency difference. These frequencies can also be compared with the corpus of standard Serbian (as reference). Since frequency information is a crucial component in human language technology, the FrAC module facilitates sharing and utilising this valued information [9], as presented in Listing 3.

4 Discussion and conclusion

In this paper, we presented AbCoSER 1.0, the first corpus of abusive language in Serbian which consists of tweets. We explained the process of data acquisition, annotation, and corpus construction. All tweets were annotated by two independent annotators, but as explained in Section 3.2, the inter-annotator agreement was moderate. Possible causes might be: 1) Lack of the generally accepted definitions of abusive speech ([14, 38]), it is often necessary to consider tweets on a case-by-case basis, 2) Individual bias of annotators due to cultural differences, personal sensibility and/or knowledge of the phenomenon, 3) Vague or incomplete annotation instructions, 4) Overlapping of abusive speech sub-categories. In general, our results are in alignment with the findings of other researchers who reported low inter-annotator agreement scores ([21, 33, 23]) As Ross et al. [36] noted, hate speech is a very vague concept that requires better definitions and guidelines. One of the characteristics of our annotation scheme is that tweets containing swear words corresponding to the category of Profanity in our data set can also be used in non-abusive informal speech. Moreover, they are often used to emphasize a positive phenomenon as in the example *Tajson je i sa 54g jebena mašina...* (Tyson is at 54 still a fucking machine...) and not just in the context of insults. The instruction to mark even those tweets as abusive might cause cognitive dissonance with annotators since they would in a regular case mark it as non-abusive ([27, 6]). This annotation approach was chosen to facilitate automatic detection of abusive speech by a system based on machine learning techniques. There are also cases when negation

Listing 3 An excerpt from RDF version with frequency and attestations.

```
# subproperty definition for frequency in twitter corpus
:atvitkoFrequency rdfs:subClassOf frac:CorpusFrequency .
:atvitkoFrequency rdfs:subClassOf [
   a owl:Restriction ;
   owl:onProperty frac:corpus ;
   owl:hasValue <https://app.sketchengine.eu/#
      dashboard?corpname=user%2Franka%2Fatwitco>] .
# frequency assessment (in twitter corpus)
:lopov frac:frequency [
   a :atvitkoFrequency;
   rdf:value "17"^^xsd:int].
# usage examples as attestations
:lopov frac:attestation attestation_1324567;
attestation_1324567 a frac:Attestation ;
   cito:hasCitedEntity  <https://app.sketchengine.eu/#
      dashboard?corpname=user%2Franka%2Fatwitco> ;
   rdfs:comment "Immoral or criminal activities" ;
   frac:locus :locus_2415677;
   frac:quotation "svi političari su lopovi." .
:locus_2415677 a :Occurrence ;
   nif:beginIndex 19 ;
   nif:endIndex 25.
```

and emoticons change the meaning, usage of irony and sarcasm that is hard to detect in written language, as well as the necessity to possess knowledge about the world and current circumstances to understand and annotate the message.

In the next phase, we plan to extend the AbCoSER corpus with new tweets and with texts from other sources e.g. online news comments. Meanwhile, we started developing models for the automatic classification of abusive tweets and the first results are comparable with the results on similar data sets for other languages ([25, 44, 47]). The focus of our current research is the usage of a hybrid approach that combines machine learning and lexical resources. Finally, a user-friendly interface that will enable the use of these resources on the Web is under development. As for the development of the lexical resources, we plan to prepare an ontology for the classification of abusive data, including tweets, to tackle ambiguity in hate speech detection [20]. The development of the lexicon of abusive words and the ontology using VocBench[10] will continue. We also plan to enrich the lexicon with triggers identified during the annotation of abusive token spans as described in Section 2.3 and use it to upgrade the AbCoSER corpus.

References

1 Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, 2019.

2 Elisa Bassignana, Valerio Basile, and Viviana Patti. Hurtlex: A multilingual lexicon of words to hurt. In *5th Italian Conference on Computational Linguistics, CLiC-it 2018*, volume 2253, pages 1–6. CEUR-WS, 2018.

[10] http://vocbench.uniroma2.it

3 Bastian Birkeneder, Jelena Mitrovic, Julia Niemeier, Leon Teubert, and Siegfried Handschuh. upInf-Offensive Language Detection in German Tweets. In *Proceedings of the GermEval 2018 Workshop*, pages 71–78, 2018.

4 Andrej Blagojević et al. The normative framework of hate speech in Serbia and Serbian media. *FACTA UNIVERSITATIS-Law and Politics*, 14(1):81–95, 2016.

5 Julia Bosque-Gil, Jorge Gracia, and Elena Montiel-Ponsoda. Towards a Module for Lexicography in OntoLex. In *Proceedings of the LDK workshops: OntoLex, LDK 2017, Galway, Ireland*, volume 1899, pages 74–84, 2017.

6 Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. I Feel Offended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language. In Calzolari et al., editor, *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France, May 11–16 2020. European Language Resources Association (ELRA).

7 Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80. IEEE, 2012.

8 Christian Chiarcos, Christian Fäth, and Maxim Ionov. The ACoLi dictionary graph. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 3281–3290, Marseille, France, 2020. European Language Resources Association. URL: `https://www.aclweb.org/anthology/2020.lrec-1.401.pdf`.

9 Christian Chiarcos, Maxim Ionov, Jesse de Does, Katrien Depuydt, Fahad Khan, Sander Stolk, Thierry Declerck, and John Philip McCrae. Modelling Frequency and Attestations for OntoLex-Lemon. In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, pages 1–9, Marseille, France, 2020. European Language Resources Association. URL: `https://www.aclweb.org/anthology/2020.globalex-1.1.pdf`.

10 Davide Colla, Caselli Tommaso, Valerio Basile, Jelena Mitrović, and Granitzer Michael. GruPaTo at SemEval-2020 Task 12: Retraining mBERT on Social Media and Fine-tuned Offensive Language Models. In *Proceedings of the 14th International Workshop on Semantic Evaluation(SemEvaleval)*, 2020.

11 Çağrı Çöltekin. A corpus of Turkish offensive language on social media. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6174–6184, 2020.

12 Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696. Springer, 2013.

13 Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11/1, 2017.

14 Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 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*, volume 12/1, 2018.

15 Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene. In *Proceedings of the first workshop on abusive language online*, pages 46–51, 2017.

16 Paula Fortuna, Juan Soler, and Leo Wanner. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6786–6794, 2020.

17 Philippe Gambette and Jean Véronis. Visualising a text with a tree cloud. In *Classification as a Tool for Research*, pages 561–569. Springer, 2010.

18 Cvetana Krstev, Sandra Gucul, Duško Vitas, and Vanja Radulović. Can we make the bell ring? In *Proceedings of the Workshop on a Common Natural Language Processing Paradigm for Balkan Languages*, pages 15–22, 2007.

19 Ivana Krstić. Report on the use of hate speech in Serbian media, 2020. URL: https://rm.coe.int/hf25-hate-speech-serbian-media-eng/1680a2278e.

20 K. Kumaresan and K. Vidanage. Hatesense: Tackling ambiguity in hate speech detection. In *2019 National Information Technology Conference (NITC)*, pages 20–26, 2019. doi:10.1109/NITC48475.2019.9114528.

21 Irene Kwok and Yuzhou Wang. Locate the hate: Detecting tweets against blacks. In *Proceedings of the twenty-seventh AAAI conference on artificial intelligence*, pages 1621–1622, 2013.

22 Biljana Lazić and Mihailo Škorić. From DELA based dictionary to Leximirka lexical database. *Infotheca – Journal for Digital Humanities*, 19(2):81–98, 2020. doi:10.18485/infotheca.2019.19.2.4.

23 Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. The FRENK datasets of socially unacceptable discourse in Slovene and English. In *International Conference on Text, Speech, and Dialogue*, pages 103–114. Springer, 2019.

24 John McCrae, Guadalupe Aguado-de Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asunción Gómez-Pérez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, et al. Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(4):701–719, 2012. doi:10.1007/s10579-012-9182-3.

25 Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153, 2016.

26 Government of the Republic of Serbia. Criminal code of the Republic of Serbia. *Službeni glasnik*, 35:1–104, 2019.

27 Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. Do you really want to hurt me? predicting abusive swearing in social media. In *The 12th Language Resources and Evaluation Conference*, pages 6237–6246. European Language Resources Association, 2020.

28 Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, 57(6):102360, 2020.

29 Endang Wahyu Pamungkas and Viviana Patti. 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*, pages 363–370, 2019.

30 Nikola Pantelić. CRIMINAL OFFENSES COMMITTED ON SOCIAL NETWORKS: Structure of the offense and position of the perpetrator, 2017.
URL: https://www.paragraf.rs/100pitanja/krivicno_pravo/krivicna-dela-izvrsena-na-drustvenim-mrezama- struktura-dela-i-polozaj-izvrsioca.html.

31 Ji Ho Park and Pascale Fung. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint*, 2017. arXiv:1706.01206.

32 Ted Pedersen. Duluth at SemEval-2019 task 6: Lexical approaches to identify and categorize offensive tweets. *arXiv preprint*, 2020. arXiv:2007.12949.

33 Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. Hate speech annotation: Analysis of an Italian twitter corpus. In *4th Italian Conference on Computational Linguistics, CLiC-it 2017*, volume 2006, pages 1–6. CEUR-WS, 2017.

34 Amir H Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. Offensive language detection using multi-level classification. In *Canadian Conference on Artificial Intelligence*, pages 16–27. Springer, 2010.

35 Mohammadreza Rezvan, Saeedeh Shekarpour, Lakshika Balasuriya, Krishnaprasad Thirunarayan, Valerie L Shalin, and Amit Sheth. A quality type-aware annotated corpus and lexicon for harassment research. In *Proceedings of the 10th ACM Conference on Web Science*, pages 33–36, 2018.

36 Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint*, 2017. `arXiv:1701.08118`.

37 Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths. A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint*, 2017. `arXiv:1709.10159`.

38 Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10, 2017.

39 Alessandro Seganti, Helena Sobol, Iryna Orlova, Hannam Kim, Jakub Staniszewski, Tymoteusz Krumholc, and Krystian Koziel. NLPR@ SRPOL at SemEval-2019 Task 6 and Task 5: Linguistically enhanced deep learning offensive sentence classifier. *arXiv preprint*, 2019. `arXiv:1904.05152`.

40 Gudbjartur Ingi Sigurbergsson and Leon Derczynski. Offensive language and hate speech detection for Danish. *arXiv preprint*, 2019. `arXiv:1908.04531`.

41 Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. Analyzing the targets of hate in online social media. *arXiv preprint*, 2016. `arXiv:1603.07709`.

42 Ranka Stanković, Jelena Mitrović, Danka Jokić, and Cvetana Krstev. Multi-word Expressions for Abusive Speech Detection in Serbian. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 74–84, 2020.

43 Julien Tissier, Christophe Gravier, and Amaury Habrard. Dict2vec: Learning word embeddings using lexical dictionaries. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, 2017.

44 William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26, 2012.

45 Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.

46 Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 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, June 1–June 6, 2018, New Orleans, Louisiana, Vol. 1*, 2018.

47 Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria – September 21, 2018. - Vienna, Austria*, pages 1–10, 2018.

48 Thilini Wijesiriwardene, Hale Inan, Ugur Kursuncu, Manas Gaur, Valerie L Shalin, Krishnaprasad Thirunarayan, Amit Sheth, and I Budak Arpinar. Alone: A dataset for toxic behavior among adolescents on twitter. In *International Conference on Social Informatics*, pages 427–439. Springer, 2020.

49 Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399, 2017.

50 Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the type and target of offensive posts in social media. *arXiv preprint*, 2019. `arXiv:1902.09666`.

51 Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 75–86. Association for Computational Linguistics, 2019. `doi:10.18653/v1/s19-2010`.

52 Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*, 2020.