Journal of Big Data

RESEARCH Open Access

An approach to automatic classification of hate speech in sports domain on social media

Staša Vujičić Stanković[1*] and Miljana Mladenović[2]

*Correspondence:
stasa.vujicic.stankovic@matf.bg.ac.rs

[1] Faculty of Mathematics, Department of Informatics, University of Belgrade, Belgrade, Serbia
[2] Faculty of Pedagogy, University of Niš, Niš, Serbia

Abstract
Hate Speech encompasses different forms of trolling, bullying, harassment, and threats directed against specific individuals or groups. This phenomena is mainly expressed on Social Networks. For sports players, Social Media is a means of communication with the widest part of their fans and a way to face different cyber-aggression forms. These virtual attacks can harm players, distress them, cause them to feel bad for a long time, or even escalate into physical violence. To date, athletes were not observed as a vulnerable group, so they were not a subject of automatic Hate Speech detection and recognition from content published on Social Media. This paper explores whether a model trained on the dataset from one Social Media and not related to any specific domain can be efficient for the Hate Speech binary classification of test sets regarding the sports domain. The experiments deal with Hate Speech detection in Serbian. BiLSTM deep neural network was learned with different parameters, and the results showed high Precision of detecting Hate Speech in sports domain (96% and 97%) and pretty low Recall.

Keywords: Hate speech, Sport, Automatic hate speech recognition, Social networks, Social media

Introduction

Nowadays, Social Media (SM) are an essential part of human life. They are used for business, entertainment, communications with friends and fellow workers, representing skills, knowledge, and abilities, and acquiring new ones. SM are closely related to all human activities. They represent a means of communication between sports players with the broadest part of their fans. Athletes can get extra incentives for further effort and better results through contact with their fans. However, on the other hand, SM can be used by unwell-meaning people to destabilize and frustrate athletes in their efforts. It is crucial to point out that, by using SM, sports players can face up: to different kinds of aggression, like flaming, harassment, hate, and trolling, as well as other kinds of hate speech like an insult, quarrels, swearing, and invective, obscene, obscure, offensive, profanity, toxic speech, up to threats. This kind of harmful communication can negatively impact players, upset them, and lead to negative feelings or even real-life violence.

Objectionable Content (OC) is a term introduced in the USA in 1996 by Communications Decency Act [1]. It denotes "sexual, homicide, and violent text, pornography content, drugs, weapons, gambling, violence, hatred, and bullying and hate speech" [2, 3]. Facebook [4] also uses the term to designate different aggressor-victim relationships that appear via SM and social networks (SN). Most SN, like Facebook, Twitter, Instagram, and YouTube, have a strictly defined code of prevention and mechanisms for removing all kinds of OC. However, those phenomena are linguistically diverse and geographically widespread. One kind of OC is Hate Speech (HS). According to [5], HS is considered "*a broad umbrella term for numerous kinds of insulting user-created content, as the most frequently used expression for this phenomenon, and is even a legal term in several countries.*"

Therefore, building, using, and continually improving methods for automatically monitoring the content on SN, detecting, predicting, and mitigating OC effects can intensify the community's fight against them. It is essential to address this problem for languages in which no one has thoroughly dealt with it so far, such as Serbian.

The paper is organized as follows. After the introduction, the problem definition is outlined in Sect. "Problem definition". The related work is considered in Sect. "Related work". Section "Data preparation" describes the datasets used in the study, including the sources of the data and the steps taken to preprocess the data—how to gather HS examples from SM and how to make features for a method of automatic recognition of HS in the sports domain. In Sect. "Experimental setup—automatic recognition of hatespeech in the sports domain", our method for the automatic recognition of HS in the sports domain is presented. Section "Results and discussion" presents the experimental results and discusses the model's performance on different datasets. Section "Conclusion and future work" gives conclusions, summarizes the main contributions to the field of study, and highlights directions for future research.

Problem definition

Fortuna [6] highlighted definitions of HS adopted by major international regulatory bodies, institutions, and significant SM and descriptions adopted by the scientific community. According to the European Union (EU) Code of conduct [7], HS is "All conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic."

There are more definitions of HS in the field of study which deals with automatic detection methods. According to [2], "The automatic identification of hate speech has been mostly formulated as a natural language processing problem." So far, the scientific community has been using automatic detection methods to identify HS on online social platforms such as Facebook and MySpace [8–11], Twitter, Tumblr [8, 12–15], YouTube, Instagram, Whisper [16–27], Reddit, Slashdot [11, 28–31], or Pinterest [16]. This paper focuses specifically on HS expressed through text on SM platforms. In computational linguistics, it is known as online hate [32], cyber hate [33], or HS [34].

The primary objective of this study is to examine a particular form of HS that pertains to the sports domain and is expressed in the Serbian language. Namely, sports players and their fans are connected in many ways. According to Wasserman [35], "fans become participants, seeking to help their teams win through their cheering rituals and songs and cheers." But Wasserman also asks questions, "how far does the right to engage in this expression go?", "will fans be able to cheer and jeer using profanity?", "Can cheering rely on sexual innuendo?". US law grants freedom of speech and allows HS. At the same time, legislation in Europe tends to protect decency in communication and suppress violence, hatred, and aggression toward persons or groups determined by race, religion, ethnicity, nationality, sexual orientation, intelligence, disability, and other types of differences among people. The conflict between the protection of freedom of speech and the safety of a person from abuse, harassment, or threat makes detecting these phenomena challenging.

As SM platforms expand and the phenomenon becomes more varied, detecting and addressing this issue has become more complex. Therefore, building a generalized valuable method in different domains and SM is the aim of the field of study.

The issue of expressing hatred related to sports has been studied for a long time [36–38], but there are few studies dealing with the recognition of HS in SM that is directed against athletes [39–43] and an insignificant number that deals with automatic recognition of HS in that domain [44–48]. They are primarily in English; therefore, insults, hatred, and even threats to athletes written in different languages cannot be straightforwardly recognized and removed or recognized from SM. We believe it is vital for sports science to be linked to other sciences (like linguistics and computer science), which can help to successfully detect insults, hatred, and threats against all people in sports, regardless of language and cultural belonging.

There are many techniques and methods to automatically detect different kinds of HS in other languages [2, 32, 49]. However, they are all directed to vulnerable groups (race-related, ethnic, gender-related, refugees, groups of people with disabilities, supersized persons, and other vulnerable groups) [6, 50–55]. Different studies [36, 38, 56–58] conclude that athletes are the vulnerable group, too. However, as far as we know, these groups still require comprehensive automated HS detection like previously mentioned ones. The aim is to explore whether broadly used, generalized HS recognition methods can be adjusted for particular (sports) cases like those initially studied, for example, by Toraman et al. [48]. Transferring a domain is successfully explored in the HS detection and recognition field [48, 59, 60].

This paper is among the first studies to explore the automatic recognition of HS in the sports domain at SM in languages other than English. To the author's knowledge, this may be the first study for Serbian.

Related work

Many studies have been conducted on HS in different languages, with a particular emphasis on English. To our best knowledge, a few studies have collected datasets from SM related to the sports domain to deal with automated HS detection problem.

Pavlopoulos et al. [61] created a dataset from 1.6 million user comments from the Greek sports site Gazzetta. The dataset is publicly available, and the authors used it in Deep Learning (DL) methods to classify sports comments into accepted (not hate speech) or rejected (hate speech). The best result, measured by AUC (Area Under the ROC curve), has been achieved by RNN (Recurrent Neural Network), and it raised AUC = 0.80 and AUC = 0.84 for two different datasets produced from the original one.

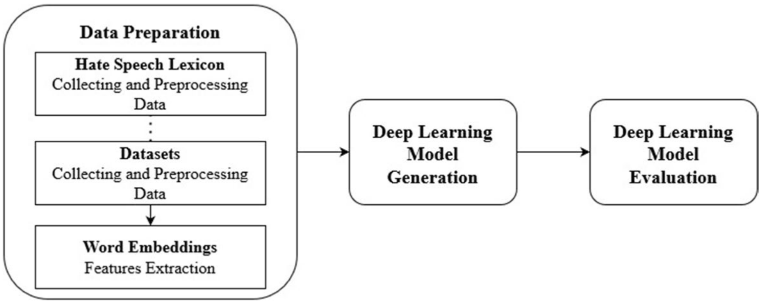

Data Preparation
Hate Speech Lexicon
Collecting and Preprocessing Data
Datasets
Collecting and Preprocessing Data
Word Embeddings
Features Extraction
Deep Learning Model Generation
Deep Learning Model Evaluation

Fig. 1 An overview diagram illustrating the approach adopted in this study

De Pelle and Moreira [62] collected a dataset with 1,250 randomly selected comments from the Globo news site on politics and sports news in Portuguese. Three annotators reviewed and marked each comment for the presence of categories such as 'racism,' 'sexism,' 'homophobia,' 'xenophobia,' 'religious intolerance,' and 'cursing.' A binary classifier into offensive or not offensive comments achieved the best F1 = 0.80.

Toraman et al. [48] retrieved more than 200 thousand top-level English and Turkish tweets published in 2020 and 2021 from five hate domains—religion, gender, racism, politics, and sports, where each tweet can belong to a single domain. Twenty thousand tweets in English and Turkish were related to the sports domain.

Kapil and Ekbal [63] also considered this problem in English. They discussed how the internet and SM platforms had created numerous opportunities for people to voice their opinions and how these platforms have facilitated the dissemination of hate speech. They proposed a model trained on a large dataset collected from diverse sources, including online forums, blogs, and SM platforms. It achieved high accuracy on all tasks. The authors also comprehensively analyse the model's performance and show that it outperforms several baseline models regarding macro-F1 and weighted-F1. Their findings suggest that distinct datasets classified into multiple subclasses help one another in the classification process. However, rather than generating a new dataset and labelling it with additional classes (which may overlap with pre-existing ones), authors recommend focusing on data classified into two primary classes—Offensive and Non-Offensive. Furthermore, Non-Offensive posts should be considered as non-hate speech, while Offensive posts can be further studied and classified into additional subclasses according to their sentiment. We also employed this approach in our research.

As can be seen from the related work presented above, the HS detection problem related to the sports domain is still an active area of research that has not been fully explored or given the attention it deserves, especially in cases of languages other than English. In this paper, we focused on Serbian. We explored whether a DL method learned on the dataset created by gathering text from different domains can be successfully applied to detect HS in the sports domain. That is, whether a generalized model can be applied to a specific case.

We achieved the following contributions, as presented in Fig. 1:

- We constructed a digital lexicon of HS terms and phrases because there was no publicly available resource of this type for Serbian.

- We crawled, refined, and formatted five datasets containing 180,785 comments. Three of them are manually annotated by 33 students, and the annotations are evaluated. The comments have been published over two years as reactions to the news and sports news on web pages on portals and YouTube channels.
- Two datasets are labelled automatically using a HS lexicon and a keyword-based approach. The datasets are used to learn domain-agnostic and domain-specific word embeddings. Word embeddings are used as features for generating DL models. We explore if models trained based on domain-agnostic features can be used for HS classification in the specific (sports) domain.

Data preparation

Hate speech lexicon

According to Mladenović et al. [64], "to generate valuable features for automatic Cyber aggression classifiers, it is necessary to include HS lexicons, psycho-linguistic resources, semantic networks, sentiment analysis lexicons, and tools." Therefore, one of the first steps in creating an application for automatic HS recognition is to make a HS lexicon of terms and phrases commonly used in a natural language which is a subject of the study. HS lexicons are important resources in automatic HS detection tasks. According to [65], "a lexicon-based approach is effective in cross-domain classification."

To induce a contemporary HS lexicon in Serbian, we retrieved scientific papers in linguistics [66–68], scientific conference proceedings [69, 70], conference papers [71–73], and lexicons published in books [74]. In the proceeding edited by Marković [69]—vulgarisms in the discourse of telephone conversations were analysed in [72], obtaining obscene words as the products of suffixation was broadly explored in [73], and generating derivatives from obscene words was presented by Bogdanović [71]. Aleksić [66] explored obscene words in a novel written by one of the contemporary writers for youth in Serbian. The author extracted vulgar and slang speech terms and collocations related to obscene meaning, swearing, and cursing. Particularly significant research [67] was conducted on the collection containing 2,130 nouns regarding pejorative, contemptuous, mocking, or ironic contexts. The collection was created from five dictionaries (The Dictionary of Serbo-Croatian Literary and Vernacular Language of the Serbian Academy of Sciences and Arts, The Matica srpska six-volume Dictionary, The Matica srpska one-volume Dictionary, Two-Way Dictionary of Serbian Slang by Dragoslav Andrić, Contemporary Belgrade Slang Dictionary by Borivoje and Nataša Gerzić). Another source of our HS lexicon is *Rečnik opscenih reči*—The dictionary of obscene words [74]. It is a comprehensive dictionary in the field of study in Serbian. We manually selected 1,209 items from this dictionary. In the proceeding edited by Marković [69]—vulgarisms in the discourse of telephone conversations were analysed in [72], obtaining obscene words as the products of suffixation was broadly explored in [73], and generating derivatives from obscene words was presented by Bogdanović [71].

Recent research in [70] has shown that a dialect's specificity must be taken into account for a better understanding of HS and to get a more generalized lexicon of obscene, vulgar, and hate words and phrases. At this stage of our research, we do not include language dialects. It is the lack of research, but this is the first version of our

Table 1 Datasets statistics

No	Purpose	Source	Number of comments	non-HS labels	HS labels	Labels
1	Training	YouTube—entertainment channels	109,676 after refining 47,884	38,789	9,095	Automatically labelled by HS lexicon
2	Testing	YouTube—entertainment channels	5,317 after refining 5,200	1,542	3,658	Manually labelled
3	Testing	YouTube—sports channels	270	11	259	Manually labelled
4	Training	News portals *blic.rs* and *b92.net*—sports news	65,155	56,316	8,839	Automatically labelled by HS lexicon
5	Testing	News portals *blic.rs* and *b92.net*—sports news	367	229	138	Manually labelled

HS lexicon. Finally, our HS lexicon has 4,705 entries representing lexemes, collocations, MWEs, and sentences[1].

We used the HS lexicon and a keyword-based approach for automatic labelling training datasets. A dataset entry is automatically labelled as a hater if a HS lexicon entry is found in the dataset entry.

Datasets

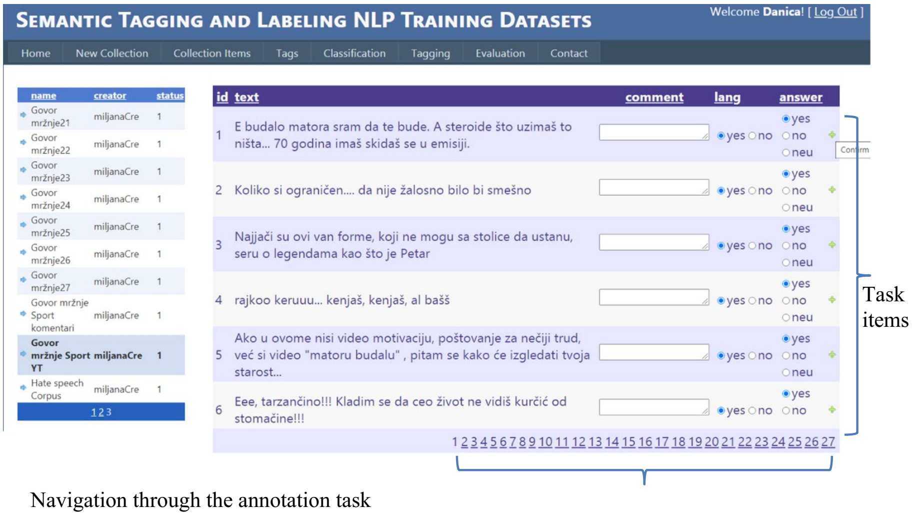
Nowadays, SM are making great efforts to suppress hate speech. Still, there are YouTube channels in Serbian where one can get hateful comments. Furthermore, such comments persist even on the two most prominent Serbian news portals, namely *blic.rs*, and *b92.net.* Therefore, we decided to use them to prepare five datasets for modelling a binary HS classifier and exploring the efficiency of transferring a model from the general domain to the source-specific domain (sports domain). Datasets are composed of comments published over a two-year period, encompassing two main sources: (1) comments from popular entertainment and sports channels on YouTube and (2) comments related to news and sports news articles on the portals *blic.rs* and *b92.net* [75].

Two datasets (one from YouTube and another from *blic.rs* and *b92.net*) are prepared to be used as training sets. The first is created of comments not specific to any particular subject or domain. These comments are considered domain-agnostic, meaning they cover a wide range of topics and are not limited to a specific subject area. The other is created of comments regarding sports (domain-specific). Three additional datasets are constructed in a similar manner, consisting of comments published as reactions to news articles and sports news on the portals *blic.rs* and *b92.net*. These datasets capture the comments specifically related to news and sports topics on these platforms. Datasets statistics are shown in detail in Table 1.

We used STL4NLP[2] [76], the web application for manual semantic annotation of a corpus in Serbian, to manually annotate test datasets. They were divided into 29 parts containing approximately the same number of comments and automatically imported into

[1] Swears are inserted in the HS lexicon as whole sentences.

[2] Available at http://ankete.mmiljana.com/.


Semantic Tagging and Labeling NLP Training Datasets
Welcome Danica! [Log Out]
Home
New Collection
Collection Items
Tags
Classification
Tagging
Evaluation
Contact
Confirm
Task items
Navigation through the annotation task

name	creator	status
Govor mržnje21	miljanaCre	1
Govor mržnje22	miljanaCre	1
Govor mržnje23	miljanaCre	1
Govor mržnje24	miljanaCre	1
Govor mržnje25	miljanaCre	1
Govor mržnje26	miljanaCre	1
Govor mržnje27	miljanaCre	1
Govor mržnje Sport komentari	miljanaCre	1
Govor mržnje Sport YT	**miljanaCre**	**1**
Hate speech Corpus	miljanaCre	1

1 2 3

id	text	comment	lang	answer
1	E budalo matora sram da te bude. A steroide što uzimaš to ništa... 70 godina imaš skidaš se u emisiji.		●yes ○no	●yes ○no ○neu
2	Koliko si ograničen.... da nije žalosno bilo bi smešno		●yes ○no	●yes ○no ○neu
3	Najjači su ovi van forme, koji ne mogu sa stolice da ustanu, seru o legendama kao što je Petar		●yes ○no	●yes ○no ○neu
4	rajkoo keruuu... kenjaš, kenjaš, al bašš		●yes ○no	●yes ○no ○neu
5	Ako u ovome nisi video motivaciju, poštovanje za nečiji trud, već si video "matoru budalu" , pitam se kako će izgledati tvoja starost...		●yes ○no	●yes ○no ○neu
6	Eee, tarzančino!!! Kladim se da ceo život ne vidiš kurčić od stomačine!!!		●yes ○no	●yes ○no

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27

Fig. 2 Annotator *Danica* labelled the task named *Govor mržnje Sport YT (Hate speech Sport YT)*

STL4NLP. In that way, 29 semantic annotation tasks were created and annotated over one month. The semantic annotation task was assigned to 33 students, and each of them annotated from three to seven parts. They used three tags {'yes', 'no', 'neu'} (Fig. 2).

After annotation, we estimated the Inter-Annotation Agreement (IAA) to evaluate the quality of students' annotations. For that purpose, we used *Krippendorff's α (Kalpha)* [77] statistical measure because there were more than two annotators on each task, and some students missed annotating some comments

The value of the *Kalpha* statistical measure can be in the interval [0, 1] where *Kalpha* = 1 represents the degree of complete agreement, and *Kalpha* = 0 the degree of complete disagreement. The average IAA *Kalpha* for all 29 annotating tasks is *Kalpha* = 0.58. This value is under acceptable value ($\alpha < 0.67$), but *Kalpha* is more rigid than other statistical measures. Therefore, we have adopted all three datasets.

Datasets cleaning

The initial stage of data cleaning and preprocessing involved eliminating irrelevant characters, such as special characters, symbols, and emoticons, which were removed from all comments. We utilized the Natural Language Toolkit (NLTK) [78] for that task. Then we utilized our srNLP Python library, developed for Serbian, to split texts into sentences, tokenization, stop word removal, and transliteration from Cyrillic to Latin. Namely, the Serbian language has two official scripts, Cyrillic and Latin. Therefore, one of the vital preprocessing steps is the transliteration in one of these scripts – in our work, the transliteration of texts written in Cyrillic to Latin script.

Given the particularities of the Serbian language, we also encountered challenges concerning the use of diacritics in written texts. Several letters in Serbian Latin script include letters with diacritics (letters ć, č, đ, š, and ž). Notwithstanding, one of the problems presented in contemporary written Serbian on SM is the conspicuous omission of diacritics. Due to the lack of evaluated code available for diacritics restoration, we removed all diacritics during preprocessing.

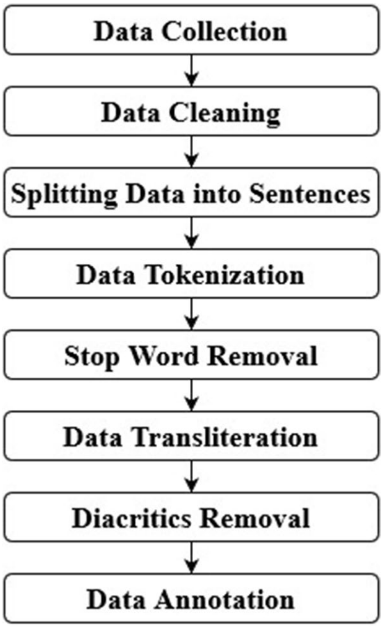


Fig. 3 Dataset preparation process

For the further process, Serbian stop words list was prepared. It originated from [79] and contained 1,267 words.

After performing a thorough preprocessing step, we generated two vocabularies. The vocabulary for the "News portals *blic.rs* and *b92.net*" training dataset initially consisted of 167,114 tokens; however, we retained only the tokens with a minimum occurrence of 2, resulting in a vocabulary of 57,535 tokens. Similarly, we generated the vocabulary for the "YouTube" training dataset. Initially, it contained 104,221 tokens, subsequently reduced to 36,319 tokens upon the same removal rule. The preparation process of the datasets is depicted in Fig. 3.

Word embeddings

Building a HS classifier with big data techniques is more manageable in English than other languages. There are powerful resources—embeddings (powered by Word2Vec [80], GloVe [81], and fastText [82]), datasets (for example, ClueWeb09 [83] and ClueWeb12 [84] corpora), and tools (NLTK, LIWC [85]), that help fast and efficient development in this field in English. Therefore, every new resource, tool, or dataset created in some other languages can be valuable for further research in the field.

Recent research on HS [48, 59, 60] suggests that big data techniques can be effectively applied using word embeddings, which involve learning a representation of words in a corpus such that semantically similar words have similar representations. Last few years, embeddings are pushing the boundaries of text classifiers. In DL techniques, they are successfully used as text-derived features.

For English and a few other languages, there are pre-trained embeddings. The good thing is that they shorten development time. However, according to Pamungkas and Pati [86], who experimented with pre-trained models (GloVe, Word2Vec, and FastText), "the result is lower compared to a self-trained model based on the training set." Also, Saleh

Table 2 Word embedding statistics

Word embedding	Dimension	Min word count	Context window length	Vocabulary length
Domain- agnostic	300	3	7	26,974

et al. [87] found that "domain-specific word embeddings outperform domain-agnostic word embedding models because it is more knowledgeable about the hate domain, while domain-agnostic are trained on books and Wikipedia, which rarely have hate community context."

For these reasons, we decided to create a word embedding representation using by the domain-agnostic dataset, i.e., the word embedding derived from dataset 1 (Table 1).

We used the Continuous Bag-of-Words (CBOW) model in Gensim [88]. The corpus contained over a million tokens without stop words. Embedding parameters are shown in Table 2.

Experimental setup—automatic recognition of hate speech in the sports domain

This section describes the experimental setup for evaluating the performance of the HS detection models, including the evaluation metrics and the training/testing procedures.

However, there are different approaches to cross-domain HS classification. Using a HS detection model trained by a specific dataset on another dataset (domain) with the same class labels is called Transfer Learning in HS detection. However, cross-domain classification is not used in the sports domain, although athletes are threatened by HS on SN. Recently, Toraman et al. [48] studied different domains cross-domain classification. They explored seven transformer-based (BERT-based) language models and two neural (CNN and LSTM) models in Turkish. They found that transformer-based language models outperform conventional ones in large-scale HS detection. However, their results have shown that "while sports can be recovered by other domains," it "cannot generalise to other domains."

The basic idea of this study is to explore if a model trained on the dataset from one SM and not related to any specific domain can be efficient for the binary classification on HS and non-HS of test sets regarding the sports domain. Therefore, we compared the results of two models trained on domain-agnostic and domain-specific datasets. The other study fact is that HS datasets usually have a high or medium level of imbalance because HS is not so frequently occurring on most SM in actual situations. For example, in the research of Davidson et al. [89], a dataset comprehending 25,000 tweets was manually annotated by the crowdsourcing technique. The annotation showed, with a very high IAA, that only 5% of tweets contained HS. Other studies showed similar HS distributions [90, 91]. Zhang et al. [92] created a 300,000 tweets training dataset and found HS is under 1% ("extremely rare"). However, the effort to find such rare data is reasonable if we remember how significant the negative influence on targeted people/groups in the real

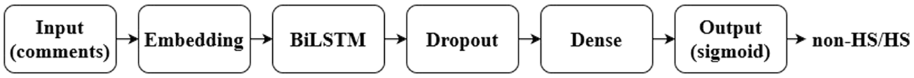


Fig. 4 Bi-LSTM learning architecture

world they have. For training/testing our networks, we used the Google Colab platform [93] with TensorFlow [94] and Keras [95] library.[3] Models are trained by Bi-LSTM.

Long Short-Term Memory Network (LSTM) is a RNN that has a repeating module. The special type of LSTM is Bidirectional Long-Short Term Memory (Bi-LSTM). These networks are used in NLP tasks like language translation, text classification, and speech recognition. RNN learns sequence patterns and uses them to make predictions of sequential data. LSTM learns order dependence and uses it to predict also sequential data. It includes a repeating module with a more complex structure than RNN repeating module. Bi-LSTM predicts a sequence by learning sequence information in both directions from future (forward) and past (backward).

We trained two models with the same DL architecture (Fig. 4) and parameters. The training set "YouTube entertainment channels" (dataset 1 from Table 1) is used to get the first model. The training parameters are as follows: trainable parameters 12,898,945, vocabulary 57,531 tokens, epochs 5, dimension 200. LSTM output size is set to 64. The dropout rate is 50%. The model gained a training accuracy of 91%. The "News portals training set" (dataset 4 from Table 1) was trained with the same parameters and achieved an accuracy of 93%. The first training set was also used for training with 20 epochs and the same rest parameters. It reached a training accuracy of 97%. For embedding we used two types of representations: BoW model with count values vectors and one-hot encoded vector for each word.

The model is compiled with the Adam optimizer, and the loss parameter is set to binary-crossentropy value which is the recommendation for binary classification models. The output layer takes one unit with Sigmoid activation function.

After that, both datasets were used to train models with one-hot embedding. Because of the platform limits, we changed training parameters, so for this case vocabulary is 5000 tokens, and dimension is 50. The rest od parameters were the same as for the BoW. Training accuracy reached 98.93%.

The performance evaluation measures are Accuracy, Precision, Recall, and F1. This study evaluates specific measures for each class (HS and non-HS) because a high accuracy value does not necessarily indicate good performance on other evaluation measures in highly imbalanced datasets. In that case, a more reliable measure is F1, and Precision and Recall can also provide valuable conclusions.

Results and discussion

The test results are presented in Table 3. We should take into account two facts. Training datasets are highly unbalanced toward non-HS. Both training datasets are automatically labelled using by HS lexicon and a technique that detects HS lexicon entry in a training set's entry. Both models, trained on the "YouTube entertainment channels" and "News

[3] All datasets and notebooks are published on GitHub (https://github.com/mmiljana/hs).

Table 3 Testing results based on the accuracy, precision, recall, and F1 on non-HS and HS classes

Test dataset	Epochs	Precision non-HS / HS	Recall non-HS / HS	F1 non-HS/HS	Weighted avg Acc
Training dataset: YouTube entertainment channels—dataset 1 in Table 1					
YouTube entertainment channels (2)	5	0.29/0.68	0.80/0.18	0.43/0.28	0.36
News portals – sports news (5)	5	0.64/0.40	0.62/0.42	0.63/0.41	0.54
YouTube sports channels (3)	5	0.04/**0.97**	0.91/0.13	0.08/0.23	0.23
Training dataset: News portals sports comments—dataset 4 in Table 1					
YouTube entertainment channels (2)	5	0.29/0.68	0.78/0.20	0.42/0.31	0.34
News portals – sports news (5)	5	0.60/0.23	0.87/0.07	0.71/0.17	0.48
YouTube sports channels (3)	5	0.04/**0.96**	0.91/0.10	0.08/0.18	0.18

Table 4 Testing Results on non-HS and HS classes using the YouTube training model (dataset 1) trained with different epochs and tested on the YouTube sports comments test dataset (dataset 3)

Test dataset	Embedding	Epochs	Precision non-HS/HS	Recall non-HS/HS	F1 non-HS/HS	Weighted avg Acc
YouTube sports channels (3)	BoW	5	0.04/**0.96**	0.91/0.10	0.08/0.18	0.18
YouTube sports channels (3)	BoW	20	0.03/**0.92**	0.73/0.14	0.07/0.24	0.23
YouTube sports channels (3)	one_hot	5	0.04/**0.95**	0.64/0.32	0.07/0.48	0.46

portals sports comments" datasets, achieved high Precision in HS classification on the test "YouTube sports channels" dataset (emphasized values in Tables 3 and 4). Unlike non-HS class, which achieved high Recall values on all test datasets, the HS class has low Recall values. Nevertheless, the promising results stem from the notably high Precision values of the HS class in the sports domain, considering the highly unbalanced nature of the training datasets, their automatic annotation, and the fact that one of the trained models was not in the domain of the test dataset.

Overall results are pretty weak, but we did not expect better ones considering the mentioned facts, the small number of epochs, and not very deep network. This study investigates, inter alia, whether to continue the research of domain transferring under the given conditions of having unbalanced datasets that are automatically annotated. The results show that the Precision of predicting HS is better when YouTube is used as a source for the training data. The results indicate what needs to be improved. The Recall has to be notably higher for both sports domain test datasets. A large number of HS comments remain unfound. We can conclude that embedding has to be changed, and the network architecture and the automatic annotation has to be improved.

Table 4 shows whether training with more epochs can improve overall BoW results. Although the accuracy was slightly enhanced, HS detection was not improved. However, the embedding change improved all evaluation measures regarding the HS class.

Conclusion and future work

Considering the popularity of SM and the accompanying opportunity to express an opinion on any subject freely, HS emerges consequently in different domains. As this topic has been examined thoroughly from many points of view in general, in this paper, we have discussed the importance of the development of datasets, HS lexicon, and appropriate machine learning models, to effectively apply automatic HS recognition methods from content published in SM related to one specific domain, the sports domain. Since most research deals with English, we focused on developing resources for Serbian. We constructed a digital lexicon of HS terms and phrases. We designed a dataset composed of comments to the sports news on portals and YouTube sports channels and manually annotated for training and test purposes in our DL model. Then we trained two-word embeddings, domain-agnostic and domain-specific, regarding sports. Word embeddings are known as valuable features for generating DL models. This paper explores if models trained based on domain-agnostic features can be used for HS classification in the specific domain. We pointed out that players were not seen as a vulnerable group regarding hate speech. However, the fact is that HS on SM can have a significant impact on players and their lives. Therefore, they must also be treated as a hate speech-targeted group. In future work, we will work on the refinement of the classifier results, extending of presented datasets and resources, as well as its usage through other models.

Abbreviations

AUC	Area under the ROC curve
Bi-LSTM	Bidirectionallong short-term memory network
DL	Deeplearning
HS	Hatespeech
IAA	Inter-annotationagreement
LSTM	Longshort-term memory network
OC	Objectionablecontent
RNN	Recurrentneural network
SM	Socialmedia
SN	Socialnetwork

Acknowledgements
Not applicable.

Author contributions

MM conceived, designed the study, and collected the data. SVS contributed the data, analysis tools, and interpretation. Both authors performed the analysis and manuscript preparation. Both authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials

All results presented in this paper are reproducible. All accompanying resources regarding this paper can be found in the GitHub repository: https://github.com/mmiljana/hs. Further information is available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 28 February 2023 Accepted: 17 May 2023
Published online: 22 June 2023

References

1. EFF—Electronic Frontier Foundation. https://www.eff.org/issues/cda230. Accessed 20 Feb 2023.
2. Yang F, Peng X, Ghosh G, Shilon R, Ma H, Moore E et al. Exploring deep multimodal fusion of text and photo for hate speech classification. In: Proceedings of the third workshop on abusive language online. 2019. p. 11–8.
3. Altarturi HHM, Saadoon M, Anuar NB. Cyber parental control: a bibliometric study. Child Youth Serv Rev. 2020;116:105134.
4. Facebook Community Standards. https://www.facebook.com/communitystandards/objectionable_content. Accessed 20 Feb 2023.
5. Schmidt A, Wiegand M. A survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Valencia, Spain: Association for Computational Linguistics; 2017. p. 1–10. http://aclweb.org/anthology/W17-1101. Accessed 28 Feb 2023.
6. Fortuna P. Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes. Master's thesis. Faculdade de Engenharia da Universidade do Porto; 2017.
7. European Commission. Countering illegal hate speech online—Commission initiative shows continued improvement, further platforms join. An official website of the European Union. 2018. https://ec.europa.eu/commission/presscorner/detail/en/IP_18_261. Accessed 20 Apr 2023.
8. Badjatiya P, Gupta S, Gupta M, Varma V. Deep Learning for Hate Speech Detection in Tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion—WWW '17 Companion. Perth, Australia: ACM Press; 2017. p. 759–60. http://dl.acm.org/citation.cfm?doid=3041021.3054223. Accessed 28 Feb 2023.
9. Maisto A, Pelosi S, Vietri S, Vitale P. Mining Offensive Language on Social Media. In: Basili R, Nissim M, Satta G, editors. In: Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017. Accademia University Press; 2017. p. 252–6. http://books.openedition.org/aaccademia/2441. Accessed 28 Feb 2023.
10. Mossie Z, Wang JH. Social Network Hate Speech Detection for Amharic Language. In: Computer Science and Information Technology. Academy and Industry Research Collaboration Center (AIRCC); 2018. p. 41–55. https://airccj.org/CSCP/vol8/csit88604.pdf. Accessed 28 Feb 2023.
11. Yin D, Xue Z, Hong L, Davison B, Kontostathis A, Edwards L. Detection of harassment on web 2.0. In: Proceedings of the Content Analysis in the WEB. 2009. p. 1–7.
12. Aggrawal N. Detection of Offensive Tweets: A Comparative Study. Computer Reviews Journal. 2018;1(1):75–89.
13. Huang Q, Inkpen D, Zhang J, Van Bruwaene D. Cyberbullying intervention interface based on convolutional neural networks. In: Proc First Workshop Trolling Aggress Cyberbullying. 2018. p. 42.
14. Alfina I, Mulia R, Fanany MI, Ekanata Y. Hate speech detection in the Indonesian language: A dataset and preliminary study. In: 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS). Bali: IEEE; 2017. p. 233–8. https://ieeexplore.ieee.org/document/8355039/. Accessed 4 Mar 2023.
15. Almeida TG, Souza B, Nakamura FG, Nakamura EF. Detecting hate, offensive, and regular speech in short comments. In: Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web. Gramado RS Brazil: ACM; 2017. p. 225–8. https://doi.org/10.1145/3126858.3131576. Accessed 4 Mar 2023.
16. Bourgonje P, Moreno-Schneider J, Srivastava A, Rehm G. Automatic classification of abusive language and personal attacks in various forms of online communication. In: Rehm G, Declerck T, editors. Language technologies for the challenges of the digital age, vol. 10713. Cham: Springer International Publishing; 2018. p. 180–91. https://doi.org/10.1007/978-3-319-73706-5_15.
17. Chen Y, Zhou Y, Zhu S, Xu H. 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. Amsterdam, Netherlands: IEEE; 2012. p. 71–80. http://ieeexplore.ieee.org/document/6406271/. Accessed 28 Feb 2023.
18. Dadvar M, Jong FD, Ordelman R, Trieschnigg D. Improved cyberbullying detection using gender information. In: Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012). University of Ghent; 2012.
19. Dadvar M, Trieschnigg RB, De Jong FM. Expert knowledge for automatic detection of bullies in social networks. In: 25th Benelux Conference on Artificial Intelligence, BNAIC 2013. Delft University of Technology; 2013. p. 57–64.
20. Dinakar K, Jones B, Havasi C, Lieberman H, Picard R. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. ACM Trans Interact Intell Syst. 2012;2(3):1–30.
21. Dinakar K, Reichart R, Lieberman H. Modeling the detection of textual cyberbullying. Proc Int AAAI Conf Web Soc Media. 2021;5(3):11–7.
22. Hosseinmardi H, Mattson SA, Rafiq RI, Han R, Lv Q, Mishr S. Prediction of cyberbullying incidents on the instagram social network. arXiv. 2015. https://doi.org/10.48550/arXiv.1508.06257.
23. Liu P, Guberman J, Hemphill L, Culotta A. Forecasting the presence and intensity of hostility on Instagram using linguistic and social features. arXiv. 2018. https://doi.org/10.48550/arXiv.1804.06759.
24. Mondal M, Silva LA, Benevenuto F. A Measurement study of hate speech in social media. In: Proceedings of the 28th ACM Conference on Hypertext and Social Media. Prague Czech Republic: ACM; 2017. p. 85–94. https://doi.org/10.1145/3078714.3078723. Accessed 28 Feb 2023.
25. Silva LA, Mondal M, Correa D, Benevenuto F, Weber I. Analyzing the targets of hate in online social media. Proceedings of the International AAAI Conference on Web and SocialMedia (ICWSM'16). 2016;4(1):687–690.
26. Xu Z, Zhu S. Filtering offensive language in online communities using grammatical relations. In: Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference. 2010. p. 1–10.
27. Zhong H, Li H, Squicciarini AC, Rajtmajer SM. Content-driven detection of cyberbullying on the Instagram Social Network. In: IJCAI. 2016. p. 3952–8.

28. Bastidas A, Dixon E, Loo C, Ryan J. Harassment detection: a benchmark on the #HackHarassment dataset. arXiv. 2016. https://doi.org/10.48550/arXiv.1609.02809.
29. Kennedy G, McCollough A, Dixon E, Bastidas A, Ryan J, Loo C et al. Technology Solutions to Combat Online Harassment. In: Proceedings of the First Workshop on Abusive Language Online. Vancouver, BC, Canada: Association for Computational Linguistics; 2017. p. 73–7. http://aclweb.org/anthology/W17-3011. Accessed 28 Feb 2023.
30. dos Santos CN, Melnyk I, Padhi I. Fighting offensive language on social media with unsupervised text style transfer. arXiv. 2018. https://doi.org/10.48550/arXiv.1805.07685.
31. Saleem HM, Dillon KP, Benesch S, Ruths D. A web of hate: tackling hateful speech in online social spaces. arXiv. 2017. https://doi.org/10.48550/arXiv.1709.10159.
32. Gao L, Huang R. Detecting online hate speech using context aware models. arXiv. 2018. https://doi.org/10.48550/arXiv.1710.07395.
33. Blaya C. Cyberhate: a review and content analysis of intervention strategies. Aggress Violent Behav. 2019;45:163–72.
34. Fortuna P, Nunes S. A survey on automatic detection of hate speech in text. ACM Comput Surv. 2019;51(4):1–30.
35. Wasserman HM. Fans, free expression, and the wide world of sports. U Pitt Rev. 2005;67:525.
36. Burnap P, Rana OF, Avis N, Williams M, Housley W, Edwards A, et al. Detecting tension in online communities with computational Twitter analysis. Technol Forecast Soc Change. 2015;95:96–108.
37. Knežević A. Hate speech in sport: causes, forms, targets and consequences. combating hate speech in sport. In: A workshop bringing together youth and sport officials, researchers and policy-makers to deepen the understanding of hate speech in sport and identify appropriate responses. 2017.
38. McLean L, Griffiths MD. Moral disengagement strategies in videogame players and sports players. In: Management Association IR, editor. Research anthology on business strategies, health factors, and ethical implications in sports and eSports. Hershey: IGI Global; 2021. p. 958–78.
39. Sanderson J. From loving the hero to despising the villain: sports fans, Facebook, and social identity threats. Mass Commun Soc. 2013;16(4):487–509.
40. Cleland J. Racism, football fans, and online message boards: how social media has added a new dimension to racist discourse in English football. J Sport Soc Issues. 2014;38(5):415–31.
41. Kohno Y, Kitamura K. International Sporting events and racism on the web: a study on Japanese web comments regarding the 2014 FIFA World Cup in Brazil. J Jpn Soc Sports Ind. 2017;27(2):2149–62.
42. Dogar Y. Analyzing the cyberbullying behaviors of sports college students. Int Educ Stud. 2019;12(11):36.
43. MacPherson E, Kerr G. Online public shaming of professional athletes: gender matters. Psychol Sport Exerc. 2020;51:101782.
44. Sanderson J, Truax C. "I hate you man!": exploring maladaptive parasocial interaction expressions to college athletes via Twitter. J Issues Intercoll Athl. 2014;7(1):333–51.
45. Litchfield C, Kavanagh EJ, Osborne J, Jones I. Virtual maltreatment: sexualisation and social media abuse in sport. Psychol Women Sect Rev. 2016;18(2):36–47.
46. Sanderson J. Elite quarterbacks do not laugh when they are losing: exploring fan responses to athletes' emotional displays. Int J Sport Exerc Psychol. 2016;14(3):281–94.
47. MacPherson E, Kerr G. Sport fans' responses on social media to professional athletes' norm violations. Int J Sport Exerc Psychol. 2021;19(1):102–19.
48. Toraman C, Şahinuç F, Yilmaz EH. Large-scale hate speech detection with cross-domain transfer. arXiv. 2022. https://doi.org/10.48550/arXiv.2203.01111.
49. Fauzi MA, Yuniarti A. Ensemble method for Indonesian twitter hate speech detection. Indones J Electr Eng Comput Sci. 2018;11(1):294.
50. Tulkens S, Hilte L, Lodewyckx E, Verhoeven B, Daelemans W. The Automated Detection of Racist Discourse in Dutch Social Media. Computational Linguistics in the Netherlands Journal. 2016 Dec 1;6:3–20.
51. Sanguinetti M, Poletto F, Bosco C, Patti V, Stranisci M. An Italian Twitter Corpus of Hate Speech against Immigrants. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). 2018. p. 2798–2805. https://aclanthology.org/L18-1443.pdf. Accessed 20 Nov 2022.
52. Sharifirad S, Jafarpour B, Matwin S. Boosting Text Classification Performance on Sexist Tweets by Text Augmentation and Text Generation Using a Combination of Knowledge Graphs. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). Brussels, Belgium: Association for Computational Linguistics. 2018. p. 107–114. http://aclweb.org/anthology/W18-5114. Accessed 20 Nov 2022.
53. Malik JS, Pang G, Hengel A van den. Deep Learning for Hate Speech Detection: A Comparative Study. arXiv. 2022. http://arxiv.org/abs/2202.09517.
54. Davidson T, Bhattacharya D, Weber I. Racial Bias in Hate Speech and Abusive Language Detection Datasets. arXiv. 2019. https://doi.org/10.48550/arXiv.1905.12516.
55. Watanabe H, Bouazizi M, Ohtsuki T. Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection. IEEE Access. 2018;6:13825–13835.
56. Ntoumanis N, Biddle S. The relationship between competitive anxiety, achievement goals, and motivational climates. Res Q Exerc Sport. 1998;69(2):176–87.
57. Bordman EY. Freedom of Speech and expression in Sports the Balance between the Rights of the individual and the best interests of Sport. Mich BAR J. 2007;86(9):36.
58. Conroy DE, Coatsworth JD, Kaye MP. Consistency of fear of failure score meanings among 8- to 18-Year-old female athletes. Educ Psychol Meas. 2007;67(2):300–10.
59. Bashar MA, Nayak R, Luong K, Balasubramaniam T. Progressive domain adaptation for detecting hate speech on social media with small training set and its application to COVID-19 concerned posts. Soc Netw Anal Min. 2021;11(1):69.
60. Yang C, Zhu F, Liu G, Han J, Hu S. Multimodal hate speech detection via cross-domain knowledge transfer. In: Proceedings of the 30th ACM International Conference on Multimedia. 2022. p. 4505–14.
61. Pavlopoulos J, Malakasiotis P, Androutsopoulos I. Deep learning for user comment moderation. arXiv. 2017. https://doi.org/10.48550/arXiv.1705.09993.

62. De Pelle RP, Moreira VP. Offensive comments in the Brazilian web: a dataset and baseline results. In: Anais do Brazilian Workshop on Social Network Analysis and Mining (BraSNAM). Sociedade Brasileira de Computação—SBC; 2017 https://sol.sbc.org.br/index.php/brasnam/article/view/3260. Accessed 28 Feb 2023.
63. Kapil P, Ekbal A. A deep neural network based multi-task learning approach to hate speech detection. Knowl-Based Syst. 2020;210:106458.
64. Mladenović M, Ošmjanski V, Vujičić Stanković S. Cyber-aggression, cyberbullying, and cyber-grooming: a survey and research challenges. ACM Comput Surv. 2022;54(1):1–42.
65. Wiegand M, Ruppenhofer J, Schmidt A, Greenberg C. 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. 2018. p. 1046–56.
66. Aleksić K. Lexicon in the novel Claws by Marko Vidojković —unpublished master's thesis. University in Novi Sad, Faculty of Philosophy, Novi Sad; 2019.
67. Jovanović J. Lexicon of derogatory meaning in naming a person in the Serbian language—unpublished doctoral thesis. University of Belgrade, Faculty of Philology, Belgrade; 2018.
68. Mandić M, Đurić L. Serbian Swearwords as a Folklore Genre: the case of jebem ti sunce ('I fuck your sunshine'). Contemporary serbian folkloristics. University Library "Svetozar Marković", Belgrade; 2015. (II).
69. Marković J. Opscena leksika u srpskom jeziku—zbornik radova sa istoimenog naučnog skupa. Niš: University of Niš, Faculty of Philosophy; 2017.
70. Marković J, Jović N. Opscena leksika i druga kolokvijalna leksika u srpskom i makedonskom jeziku—zbornik radova sa istoimenog naučnog skupa. Niš: University of Niš, Faculty of Philosophy; 2019.
71. Bogdanović N. Opsceno i vulgarno kao determinacija. In: Zbornik radova sa naučnog skupa "Opsceno i vulgarno kao leksika i druga kolokvijalna leksika u srpskom i makedonskom jeziku." University of Niš, Faculty of Philosophy, Niš; 2017. p. 15-20.
72. Jovanović I. Vulgarizmi u diskursu telefonskih razgovora: jedan primer iz ruralne sredine. In: Zbornik radova sa naučnog skupa "Opsceno i vulgarno kao leksika i druga kolokvijalna leksika u srpskom i makedonskom jeziku." Niš: University of Niš, Faculty of Philosophy, Niš; 2017. p. 75–96.
73. Lilić D. Tvorba opscene leksike i vulgarizama. In: Zbornik radova sa naučnog skupa "Opsceno i vulgarno kao leksika i druga kolokvijalna leksika u srpskom i makedonskom jeziku." University of Niš, Faculty of Philosophy, Niš; 2017. p. 165-171.
74. Šipka D. Rečnik opscenih reči i izraza. Buffalo: Prometej; 2011.
75. Mladenović M, Momčilović V, Prskalo I. Stl4nlp–web tool for manual semantic annotation of digital corpora. In: The strategic directions of the development and improvement of higher education quality: challenges and dilemmas. 2020. p. 200–212.
76. Mladenović M. STL4NLP—Semantic tagging and labeling NLP training datasets. 2023. http://ankete.mmiljana.com/. Accessed 22 Nov 2022.
77. Krippendorff K. Content analysis: an introduction to its methodology. 2nd ed. Thousand Oaks: Sage; 2004. p. 413.
78. NLTK—Natural Language Toolkit. 2023. https://www.nltk.org. Accessed 02 Feb 2023.
79. Mladenović M, Mitrović J, Krstev C, Vitas D. Hybrid sentiment analysis framework for a morphologically rich language. Journal of Intelligent Information Systems. 2016;46(3):599–620.
80. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv. 2013. https://doi.org/10.48550/arXiv.1301.3781.
81. Pennington J, Socher R, Manning C. Glove. Global Vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics; 2014. p. 1532–43. http://aclweb.org/anthology/D14-1162. Accessed 28 Apr 2023.
82. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics. 2017;5:135–46.
83. ClueWeb09—The. ClueWeb09 Dataset. http://lemurproject.org/clueweb09. Accessed 02 Feb 2023.
84. ClueWeb12—The ClueWeb12 Dataset. 2023. http://lemurproject.org/clueweb12. Accessed 02 Feb 2023.
85. LIWC. 2023. https://www.liwc.app. Accessed 02 Feb 2023.
86. Pamungkas EW, Patti V. 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. Florence, Italy: Association for Computational Linguistics; 2019. p. 363–70. https://www.aclweb.org/anthology/P19-2051. Accessed 28 Feb 2023.
87. Saleh H, Alhothali A, Moria K. Detection of hate speech using BERT and hate speech word embedding with deep model. Appl Artif Intell. 2023;37(1):2166719.
88. Gensim Development Team. Gensim Word2Vec CBOW model (version 4.0.0). 2020. https://radimrehurek.com/gensim/models/word2vec.html. Accessed 02 Feb 2023.
89. Davidson T, Warmsley D, Macy M, Weber I. Automated hate speech detection and the problem of offensive language. Proc Int AAAI Conf Web Soc Media. 2017;3(1):512–5.
90. Fišer D, Erjavec T, Ljubešić N. Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in slovene. In: Proceedings of the First Workshop on Abusive Language Online. Vancouver, BC, Canada: Association for Computational Linguistics; 2017. p. 46–51. http://aclweb.org/anthology/W17-3007. Accessed 13 May 2023.
91. Risch J, Krestel R. Delete or not delete? Semi-automatic comment moderation for the newsroom. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). Santa Fe, New Mexico, USA: Association for Computational Linguistics; 2018. p. 166–76. https://aclanthology.org/W18-4420. Accessed 10 May 2023.
92. Zhang Z, Robinson D, Tepper J, et al. Detecting hate speech on twitter using a convolution-GRU based deep neural network. In: Gangemi A, Navigli R, Vidal ME, Hitzler P, Troncy R, Hollink L, editors., et al., The semantic web, vol. 10843. Cham: Springer International Publishing; 2018. p. 745–60. https://doi.org/10.1007/978-3-319-93417-4_48.
93. Google Colaboratory. 2023. https://colab.research.google.com/. Accessed 03 Feb 2023.

94. TensorFlow. A machine learning library for numerical computation. 2023. https://www.tensorflow.org/. Accessed 03 Feb 2023.

95. Keras. 2023. https://keras.io/. Accessed 03 Feb 2023.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.