Legal Framework, Dataset and Annotation Schema for Socially Unacceptable Online Discourse Practices in Slovene

Darja Fišer
Faculty of Arts
University of Ljubljana
Aškerčeva cesta 2
1000 Ljubljana, Slovenia
darja.fiser@ff.uni-lj.si

Nikola Ljubešić
Dept. of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39
1000 Ljubljana, Slovenia
nikola.ljubesic@ijs.si

Tomaž Erjavec
Dept. of Knowledge Technologies
Jožef Stefan Institute
Jamova cesta 39
1000 Ljubljana, Slovenia
tomaz.erjavec@ijs.si

Abstract

In this paper we present the legal framework, dataset and annotation schema of socially unacceptable discourse practices on social networking platforms in Slovenia. On this basis we aim to train an automatic identification and classification system with which we wish contribute towards an improved methodology, understanding and treatment of such practices in the contemporary, increasingly multicultural information society.

1 Introduction

In Slovenia, Socially Unacceptable Discourse (SUD) practices, such as prosecutable hate speech, threats, abuse and defamations, but also not prosecutable but still indecent and immoral insults and obscenities, are heavily researched by sociologists (Dragoš, 2007; Leskošek, 2004). They receive regular coverage in the media, public debates are held about it in the parliament, several national and international initiatives and activities address them (Motl and Bajt, 2016), all with the aim to raise awareness and propose efficient prevention strategies.

Despite all these efforts, their success has been limited as was clearly indicated in the second half of 2015 when extreme forms of SUD flooded social media as a response to the migrant crisis in the Balkans. This trend is confirmed by the records of the Spletno Oko (Web Eye) national hotline service for reporting online hate speech, which forwarded 75% more of the applications received to the police in 2015 than the year before (Vehovar and Motl, 2015). Even when criminal or civil cases are filed, very few of them make it as far as a court hearing, let alone a conviction. Here, the biggest bottleneck is not the definition of legally unacceptable forms of speech in the Penal Code (public promotion of hatred, violence or intolerance) but in the syllogism process, i.e. the application of the general legal norm to the facts of a particular case (Rovšek, 2011; Šalamon, 2015).

This shows that new interdisciplinary theoretical and analytical methods and approaches are needed to improve our understanding as well as to enable efficient and comprehensive identification and classification of SUD in the contemporary, increasingly multicultural information society. As of yet, there are no publications reporting on successful attempts to automate the identification of SUD for Slovene, which is hardly surprising as most work has so far been limited to English, with a few exceptions for Dutch (van Halteren and Oostdijk, 2013) and German (Ross et al., 2017). State-of-the-art approaches tackle this task through supervised machine learning (Sood et al., 2012; Dadvar et al., 2013). For this, of course, manually annotated datasets are needed.

A major limitation of most existing work in this area is that it is based on an ad-hoc treatment of SUD classification in natural language processing and a lack of detailed guidelines that are necessary for reliable annotation (Ross et al., 2017). Annotated datasets have started to emerge only recently (Nobata et al., 2016; Waseem and Hovy, 2016), but nevertheless they lack precise documentation on data annotation and make use of only very basic

annotation schemas. The community could benefit from input by experts from the area of SUD, which is the goal of this paper, in which we present the legal framework, the database, and the annotation schema of Slovene socially unacceptable online discourse practices that was developed in collaboration by sociologists and legal experts who specialize in SUD. Since Slovene legislation is in line with all the relevant EU directives, the proposed schema and annotation principles could be applied to other languages as well.

2 The FRENK Project

The work presented in this paper serves as the foundation for FRENK, a new 3-year interdisciplinary national basic research project funded by the Slovenian national research agency from May 2017 to May 2020. For the first time, the project combines researchers from the fields of NLP, sociolinguistics, sociology and law. Its goal is the development of resources, methods and tools for the understanding, identification and classification of various forms of SUD in the information society. The project aims to combine state-of-the-art quantitative and qualitative multidisciplinary approaches which will be employed to investigate the use of socially unacceptable discourse in its sociocultural context.

In the scope of the project we will use social media data to construct a large corpus of SUD that will be highly structured and their (often non-standard) texts linguistically processed as well as enriched with various metadata with the help of our toolchain for the processing of noisy user-generated content (Fišer et al., 2017). Using the typology of socially unacceptable discourse and its targets and the manually annotated representative sample of texts presented in this paper we will apply machine learning techniques to flag and categorise SUD texts and their targets.

With the methodologies and instruments of corpus linguistics, critical discourse analysis and inferential statistics, interdisciplinary (socio)linguistic analyses will be performed on the collected and processed resources, focusing on migrants and Islamophobia, and homophobia and gay rights. These approaches will be supplemented with a corpus analysis of legal aspects of socially unacceptable discourse and sociological surveys on its the perception in the Slovene society.

3 Legal framework

The term *hate speech*, the strongest form of socially unacceptable discourse practices, is not explicitly used in the Slovene legislation. Instead, criminally prosecutable acts due to public promotion of hatred, violence or intolerance that can be understood as hate speech are included in Article 297 of the Penal Code. However, (Šalamon, 2015) warns that with the most recent amendment of the Code in 2012, the definition became much more precise and narrow, perhaps even too narrow, as it excludes acts of verbal outrage that do not include elements of a threat or abuse and cannot endanger law and order.

(Motl and Bajt, 2016) reach a similar conclusion in their overview of the legal framework and legal practice in Slovenia where they show that hate speech is becoming commonplace and still very rarely sanctioned. What is more, the issue of (non criminal) intolerant speech is more often than not underestimated and treated as occasional excess by the key stakeholders.

According to its treatment in the Slovene legal framework, (Vehovar et al., 2012) defined three levels of SUD found online. The largest share is represented by *Inappropriate Speech* with which they signify various forms of socially undesired, indecent and immoral discourse practices, such as swearwords, insults, vulgar or obscene language and profanities. While there are no legal grounds for the prosecution of such types of discourse practices as they are protected by the free speech provisions, they are typically regulated with codes of conduct by owners of online portals.

The second level are instances of *Inadmissible Speech*, which comprise discourse practices that contain false statements that harm the reputation of an individual, group of people or organization or those that threaten someone's life or security. Both are punishable by the Penal Code and, depending on whether they are directed towards a *social group* due to race, ethnicity, religion, sexual orientation of their members, or towards a *specific individual*, prosecuted ex officio or by the party concerned.

Finally, the highly restrictive account of *Hate Speech* is specifically reserved for discourse practices that are directed towards, promote intolerance and call to violence against a social subgroup based on their racial or ethnic profile, religion, sexual orientation or political affiliation.

4 Database of Slovene SUD

The biggest and most authoritative database of socially unacceptable online discourse practices in Slovene is being collected through the Spletno Oko[1] (Web Eye) hotline service that enables Internet users anonymous reporting of hate speech and child sexual abuse content they come across online. The hotline was established in 2006 within the international Safer Internet Program[2] and is financed by the European Commission (INEA agency) and the Slovenian Ministry of Public Administration. Its main mission is to reduce the amount of child sexual abuse content and hate speech online in cooperation with the police, internet service providers, and other governmental and non-governmental organizations. Apart from awareness raising campaigns and exchange of best practices with other hotlines in the network, the Safer Internet Centre performs a fast analysis of the submissions and reports the potentially criminal cases to the authorities.

The most recent version of the Spletno Oko database contains reported SUD instances from online networking and social media sites from 2010 onwards, comprising 13,000 text instances or about 900,000 tokens. All the reported text instances were examined and classified into one of the categories according to the legal framework by a professional analyst with a degree in sociology, criminology or law and specialised training for the job at the hotline service. In the first years of the hotline's operations, most of the reported text instances were news comments from online news portals. This is why the hotline drafted the "Code for the regulation of hate speech on online portals"[3] in 2013 which has since been signed by most major online news portals in the country. As a result, the amount of reported instances from online news portals has declined substantially. In the past few years, the prevailing, and increasing, source of reports to the hotline are Facebook groups and pages.

5 Annotation of Slovene SUD

A prerequisite of any automatic approaches to the detection and classification of SUD is the compilation of a manually annotated dataset. In the FRENK project we will build upon the invaluable Spletno Oko database but since the annotation at the hotline service was not set up in a way that would directly enable a successful transfer to the machine learning environment, a number of steps are needed to harmonise both initiatives, which we describe in this section.

First and foremost, the flat annotation schema needs to be redesigned in such a way that it allows for both coarse- and fine-grained SUD classification (see Section 5.1) and complemented by detailed annotation guidelines, which ensure consistent annotation as well as serve for documentation purposes and for potential future annotation campaigns to improve comparability of the results. To overcome low annotation agreement, instead of the existing single annotations multiple annotations need to be obtained for each data point in the early phases of the annotation campaign, followed by a post-hoc adjudication procedure.

This will help us arrive at gold-standard annotations as well as work out possible issues either in the annotation schema or the annotation guidelines. We will adopt the MATTER annotation framework (Pustejovsky and Stubbs, 2012), i.e. Modelling the phenomenon, Annotating it, Training and Testing the ML methods, Evaluating their fitness of purpose, and possibly Revising the procedure on the basis of the evaluation. The annotation process should not be linear but proceed in several cycles accompanied by the refinement of the annotation schema and the guidelines and resulting in a high-quality dataset that can at the same time be used also for linguistic, sociological as well as legal investigations of SUD. By following these principles we believe we can advance the state-of-the-art in computational linguistic SUD investigations, where such datasets have so far been annotated in a rather cursory fashion.

5.1 Annotation schema

For the annotation campaign within the FRENK project the typology developed by the "Spletno Oko" hotline experts (Vehovar et al., 2012) has been modified to better facilitate automatic identification and classification of SUD, our ultimate goal. The originally flat typology was reorganized into a two-level schema which allows for both coarse- (2-class: SUD, not SUD), medium- (4-class: category level 1 in Figure 1) and fine-

[1] http://www.spletno-oko.si/english/

[2] https://www.betterinternetforkids.eu/web/portal/policy/insafe-inhope

[3] http://www.spletno-oko.si/sovrazni-govor/za-urednike-spletnih-mest

Typology of SUD
1. No Elements of Problematic Speech
1.1 Reports are false, void
1.2 Texts contain no unacceptable speech
2. Inappropriate Speech
2.1 Texts contain insulting, offensive speech
2.2 Text contain obscenity, profanity, vulgarity
3. Inadmissible Speech
3.1 Texts contain defamatory speech
3.2 Texts contain abusive, threatening speech
4. Hate Speech
4.1 Socially unacceptable hate speech
4.2 Potentially legally punishable hate speech
Target of SUD
1. Ethnicity
2. Race
3. Sexual orientation
4. Political affiliation
5. Religion
Metadata
1. Date of submission
2. URL of the reported SUD
3. Text of the reported SUD

Figure 1: SUD Annotation schema used in the Spletno Oko database.

grained (8-class: category levels 1 and 2 in Figure 1) treatment of SUD. It will be interesting to explore which of those yield better results for each of the stakeholders (NLP researchers, sociologists, lawyers, moderators of online portals).

As can be seen in Figure 1, the underlying legal principles described in Section 2 serve as the basis of a hierarchical two-level SUD annotation schema that is applied to classify the reports submitted through the helpline, which yield the 4 top categories: *1. No Elements of Problematic Speech*, *2. Inappropriate Speech*, *3. Inadmissible Speech*, and *4. Hate Speech*.

Each of the top categories has two subcategories, all of which have a legal basis with one exception, namely the subcategory *4.1 Socially Unacceptable Hate Speech*. This additional subcategory was introduced in the final typology because real-life cases of highly volatile online discourse practices contain some but not all elements of hate speech as required by the Penal Code. While as opposed to potentially *Legally Punishable Hate Speech*, is not a legal category according to Slovene legislation, it is of high social and sociological relevance and therefore deserves special attention.

Reports that meet the criteria of *4.1 Socially Unacceptable* and *4.2 Potentially Legally Punishable Hate Speech* are further annotated with who is the *target* of SUD: ethnicity (e.g. Roma), race (e.g. African Americans), sexual orientation (e.g. gays), political affiliation (e.g. the United Left) and religion (e.g. Islam). The target information will be an interesting feature to examine in machine learning as as well as socio-linguistic and legal analyses.

In addition to SUD type and target, annotators also record when the report was submitted, where the disputed communication was observed, as well as the entire disputed text.

5.2 Analysis of annotations

Nearly half of all the reports in the Spletno Oko database contain no elements of problematic speech (no unacceptable content 23% or false report 20%). This shows that many users of the hotline often report content which they find generally upsetting or because they feel personally insulted or attacked.

Almost a quarter of the reports contain inappropriate speech (15% insulting or offensive content, 9% obscene or vulgar language, profanities, cursing, swearwords) and as such cannot be subject to prosecution but are restricted by most online content providers and removed by moderators. These results suggest that some online content providers either do not enforce their internal rules or cannot do it quickly enough to prevent exposure to SUD among their users.

Next, 16% of the reports contain inadmissible speech (15% defamatory content, 1% threats) which are prosecutable through public prosecution or by private lawsuit. As much as 13% of the reports meet some but not all of the criteria of Article 297 of the Penal Code (e.g. spread intolerance but do not promote violence). Even though it cannot be legally prosecuted as hate speech, such content is nevertheless perceived as socially unacceptable and therefore requires special attention and proper treatment by researchers, lawmakers and content providers alike.

Finally, 3% of the reports meet all the criteria for potentially legally punishable hate speech and were reported to the authorities.

6 Conclusions

This paper presented the legal framework, annotation schema and dataset of socially unacceptable online discourse practices for Slovene, which are the first important stepping stone towards the a comprehensive, interdisciplinary treatment of the linguistic, sociological, legal and technological dimensions of various forms of SUD in Slovenia. In our future work we will develop a tool for automatic identification and classification of SUD on social media. The research will result in a thorough examination of the characteristics of SUD as a linguistic phenomenon and the social context in which explicit or implicit forms of discriminatory language are manifested. These insights will facilitate an improved understanding of the differences between legally acceptable and unacceptable forms of communication.

For the first time in Slovenia, the FRENK project brings together computer science, linguistics, sociology and law, thereby contributing to the increasingly important new research directions of the Digital Humanities and Social Sciences (DHSS) and establishing infrastructure and knowledge transfer of approaches based on large amounts of textual, sociodemographic and behaviour data. The classifier we will develop within the project has a big potential to be integrated into the daily work of moderators of discussions on the most popular forums and administrators of readers' comments on the biggest online media sites who cannot cope with the volume of posts with manual methods and are finding simple, in-house-built lexicon methods insufficient.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful comments. The work described in this paper was funded by the Slovenian Research Agency within the national basic research project Resources, methods and tools for the understanding, identification and classification of various forms of socially unacceptable discourse in the information society (J7-8280, 2017-2020).

References

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving Cyberbullying Detection with User Context. In *Proceedings of the 35th European Conference on Advances in Information Retrieval*. Springer-Verlag, Berlin, Heidelberg, ECIR'13, pages 693–696. https://doi.org/10.1007/978-3-642-36973-5_62.

Srečo Dragoš. 2007. Sovražni govor (Hate speech). *Socialno delo* 46(3):135–144.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2017. The compilation, processing and analysis of the Janes corpus of Slovene user-generated content. In Ciara R. Wigham and Gudron Ledegen, editors, *Corpus de communication médiée par les réseaux: construction, structuration, analyse*, LHarmattan, Collection Humanités Numériques.

Vesna Leskošek. 2004. Sovražni govor kot dejanje nasilja (Hate speech as an act of violence). In Vesna Leskošek, editor, *Mi in oni: Nestrpnost na Slovenskem*, Mirovni inštitut.

Andrej Motl and Veronika Bajt. 2016. Sovražni govor v Republiki Sloveniji: pregled stanja (Hate speech in the Republic of Slovenia: an overview of the situation). Technical report, Mirovni inštitut.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, WWW '16, pages 145–153. https://doi.org/10.1145/2872427.2883062.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning*. O'Reilly Media.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. *CoRR* abs/1701.08118. http://arxiv.org/abs/1701.08118.

Jernej Rovšek. 2011. Ali je sovražni govor sploh mogoče omejiti (Is it possible to limit hate speech at all?). http://mediawatch.mirovni-institut.si/bilten/seznam/39/sovrazni.

Sara Owsley Sood, Judd Antin, and Elizabeth F. Churchill. 2012. Using Crowdsourcing to Improve Profanity Detection. In *AAAI Spring Symposium: Wisdom of the Crowd*.

Hans van Halteren and Nelleke Oostdijk. 2013. Variability in Dutch Tweets. An estimate of the proportion of deviant word tokens. *Journal for Language Technology and Computational Linguistics* 29(2):97–123.

Vasja Vehovar and Andrej Motl. 2015. Letno poročilo, Spletno oko 2015. (Annual Report Web Eye 2015). Technical report, Center za varnejš internet, prijavna toča Spletno oko. Fakulteta za družbene vede. http://www.spletno-oko.si/sites/default/files/spletno_oko_-_letno_porocilo_2015.pdf.

Vasja Vehovar, Andrej Motl, Lija Mihelič, Boštjan Berčič, and Andraž Petrovčič. 2012. Zaznava sovražnega govora na slovenskem spletu (Detecting hate speech on the Web). *Teorija in praksa* 49(1):95–111. http://dk.fdv.uni-lj.si/db/pdfs/TiP2012_1_Vehovar_idr.pdf.

Neža Kogovšek Šalamon. 2015. Sovražni govor kot uradno pregonljivo kaznivo dejanje (Hate speech as officially sanctioned criminal offence). In *1. dnevi prava zasebnosti in svobode izražanja*. IUS Software, GV založba, pages 23–27.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*. Association for Computational Linguistics, San Diego, California, pages 88–93. http://www.aclweb.org/anthology/N16-2013.