No room for hate: What research about socially unacceptable discourse taught us about collaboration?

Ajda Šulc[1] and Kristina Pahor de Maiti[2]

[1] Faculty of Social Sciences, University of Ljubljana, Slovenia
[2] Faculty of Arts, University of Ljubljana, Slovenia
ajda.sulc@gmail.com
kristina.pahordemaiti@ff.uni-lj.si

Abstract. This paper offers insights into the collaboration process of a research team that brought together social scientists, humanists and computer scientists on the topic of socially unacceptable discourse online. What seemed as a straightforward problem, proved to be a complex phenomenon that required intense discussions and several iterations of solutions development in order to arrive at a result that would satisfy the individual needs of the disciplines involved. More specifically, we present the challenges faced before and during the creation of a corpus of socially unacceptable Facebook comments. From a collaboration point of view, we learned that it is crucial to set aside enough time for regular brainstorming sessions and feedback throughout the project since this prevents possibly fatal detours due to misunderstanding with regard to terminology or the scope of research. Moreover, we saw how a lack of a common system for taking scrupulous notes on all interventions into common data resource can lead into multiple iterations of simple tasks. Finally, the collaboration thought us that listening is crucial in order to optimally combine and exchange knowledge and analytical approaches among the disciplines, but also to rationally simplify tasks whenever possible.

Keywords: Socially unacceptable discourse, Hate speech, Social media, Annotation schema

1 Introduction

In the last two decades, rapid development and raising popularity of social media considerably changed our communication habits. This applies especially to written communication which is now predominantly digital. We can find a large portion of our everyday exchanges on social media, but despite all the positive aspects that this can have, many of these exchanges now reflect intolerant ideas and even encouragements to violent acts. Such utterances are frequently found in comments to posts from news media outlets that use social media platforms, such as Facebook, to disseminate their content. It has been shown that intolerant and abusive speech harms the targets as well as the society as a whole (Nielsen, 2002). To prevent these negative consequences, ef-

forts have been made to develop automated detection of intolerant utterances, and researchers from various disciplines (e.g., media studies, law, psychology, computational linguistics, sociology) are studying the phenomenon of socially unacceptable discourse with the aim to gain better understanding of its dynamics and curb its proliferation.

In this paper we use the developments and outcomes of our research on socially unacceptable discourse as an example on which we base our report on the collaboration experience in an interdisciplinary team of researchers. In Section 2, we explain our research problem from three scientific perspectives and state collaboration opportunities. In Sections 3 and 4, we present the collaboration challenges and solutions that lead to a creation of a language resource that meets the needs of social scientists, humanists and computer scientists. Section 5 concludes the paper with the main takeaways from our collaborative experience.

2 Research problem and collaboration opportunities

Socially unacceptable discourse (SUD), as we named it, is an umbrella term for communication practices that are openly or covertly harassing, provocative or insulting, incite to violence or express negative generalizations, stereotypical judgements, obscenities or incivilities (Vehovar et al., 2020). In addition to this broad spectrum of its possible manifestations, SUD is influenced by many contextual factors, such as the identity of the author/target, cultural setting, medium, language and so on (Schmidt & Wiegand, 2017). Due to its proliferation on social media in the last decade, SUD has become a trending topic in various scientific fields, but due to its complexity, the scientific community still struggles with a comprehensive description of the phenomenon. In order to contribute to the pool of insights into the nature of SUD and improve the understanding of SUD on social media, we joined forces of three scientific fields: Sociology, Linguistics and Computational linguistics. Each of the three had its own research interests in the project.

SUD is primarily a concept of Social Sciences. For this reason, the research on it in these disciplines is rich and varied. Several studies have covered SUD or some of its forms in the fields of sociology (Dragoš, 2007), communication studies (Bajt, 2018), media studies (Vehovar et al., 2012) or journalism (Milosavljevič, 2012). In our project, broadly speaking, the sociologists were mainly interested in the impacts of SUD on ideological stances of users and public communication. Therefore, they wanted to study the scope and forms of SUD in the comments, the influence of contextual factors on the formation of SUD (e.g., the media post topic, media type, target, etc.) and network interconnectivity.

In Linguistics, SUD has not been so thoroughly researched, but since it is primarily realized through linguistic means, there exists a certain volume of research on SUD from different theoretical and analytical perspectives (e.g., sociolinguistics (Gorjanc, 2005; McEnery, 2004), psycholinguistics (Kapoor, 2016; Pinker, 2008), pragmatics (Jay & Janschewitz, 2008; Pahor de Maiti & Fišer, 2020), foreign language learning (Horan, 2013), critical discourse analysis (Methven, 2017), etc.). The central linguistic research question was whether SUD is characterized by specific linguistic features, and

if so, what are they. To this end, the analysis of SUD needed to be conducted on different levels of linguistic description. The researchers wanted to look at orthographic, grammatical and lexical dimensions of SUD as well as investigate the power relations that are being constructed or maintained through language use.

Given the negative influence of SUD on communication level and society as a whole, efforts have been dedicated in the last decade to the development of tools that would enable automatic detection and removal of online SUD. But due to the complexity of SUD, accurate and timely detection has yet to be achieved (ElSherief et al., 2018; Vidgen & Yasseri, 2019; Zhang & Luo, 2019). The problem is usually regarded as a machine learning classification task in which researchers develop algorithms or produce descriptive statistics (Fortuna & Nunes, 2018). But related work shows that many challenges remain unsolved. They are mainly related to the lack of a common definition of the phenomenon, the absence of a commonly accepted benchmark corpus and a predominant focus on English data (Schmidt & Wiegand, 2017). Furthermore, researchers usually develop their datasets based on project-specific annotation schemas and use various sets of features for detection purposes (ibid.). All this hinders comparative analysis and consequently the generalization of findings. In our project, the computational linguistics group of researchers had two main research interests. The first was related to the development of a robust annotation schema that would be applicable across languages and cultures, and the second was linked to the creation of a set of features that would prove most useful for detection tasks of Slovene SUD.

Following the research interests outlined above, we saw two main collaboration opportunities: (1) annotation schema and dataset creation, and (2) the exchange of theoretical knowledge and analytical approaches. In order to be able to address all the individual needs of the three disciplines, we needed a dataset that would be enriched with extensive metadata and several annotation layers. In this step, the main collaborative efforts were therefore put into defining the necessary categories of metadata and linguistic annotations while balancing these requirements with limitations imposed by privacy regulation and computational possibilities. During the dataset creation, as well as in the following analytical phases of the project, the collaboration focused on the exchange of theoretical knowledge and methodological approaches. This collaboration was crucial due to the complex nature of the studied phenomenon. Since almost all the aspects of SUD surpass single scientific domain, we understood that in order to provide a comprehensive and reliable interpretation of the results, we will need close interdisciplinary collaboration.

3 The solution

The main idea was to extract a suitable volume of online communication to be manually annotated and thus categorized according to previously designed annotation schema. We needed a clean dataset with enough relevant comments that could be used for quantitative analyses, but at the same time manually annotated. Since this was our common goal, the solution seemed simple. Sociology and Linguistics knowledge contributed to

the content selection, while Computational linguistics experts took care of technical aspects – mainly accessing and extracting the material.

3.1 Defining and balancing the research goals

For the purposes of all the three disciplines, we agreed that we want to analyze authentic communication, i.e. real-world discourse, written spontaneously by users on the web. We needed a public source since we did not want to (and were not allowed to) invade the privacy of individuals, but also a source that would provide us with a coherent and extensive discussion. Consequently, we decided to use user comments under public news posts on Facebook that were published by the country's most read media outlets. We found that most of them are using Facebook to regularly share their own articles, and a number of followers are regularly commenting the content shared thus forming a connected string of discourse.

To be able to extract a sufficient amount of posts and associated comments, we chose the top three media outlets by their popularity according to the Alexa service[1] (i.e., 24ur.com, SIOL.net and Nova24TV), and extracted the news posts they shared on their official Facebook profile. At the time of the extraction, RTV Slovenia was also among the most popular media outlets in Slovenia, but their Facebook shares did not have enough comments to be used for the analysis so we did not include it (Ljubešič, Fišer and Erjavec, 2019). In the next step, we agreed that we need a relevant sample of comments. Since we planned to manually annotate the harvested comments, preferably each comment by several annotators to reduce the possibility of error and subjectivity, we could not afford to use random discourse, since we assumed that most of the discourse would be neutral and thus not relevant for our analysis. To ensure time and cost-efficient annotation process, we therefore choose to filter our data. Following Social Science experts' experiences on typical hateful discourse triggers, we chose the news posts on then controversial topics on two minority groups: the LGBT community and migrants/refugees. Comments under these posts were recognized as the most relevant and therefore chosen to be extracted separately for annotation.

A combination of manual and automated classifying based on key words was performed in order to filter out the posts about LGBT and migrants (Ljubešič, Fišer and Erjavec, 2019). We extracted all of the posts that were published on these two topics on the official page of the media outlet from the time their Facebook profile was activated until the time of the data collection (the end of 2017). For the Slovene data, the algorithm identified 93 posts and 4.571 comments about LGBT and 967 posts and 43.000 comments about migrants. The latter were reduced to 30 most relevant posts with 6.545 comments for the annotation process in order to have similar and manually doable amount of comments for both minorities (Vehovar et al., 2020).

[1] https://www.alexa.com/topsites/countries

4 The collaboration experiences

Following the agreement on what data was to be annotated, an annotation schema had to be designed, tested and used. It first seemed like a simple task of choosing the relevant categories of discourse, but we found that there were quite some dilemmas resulting from different understandings of the main concept and different needs of the three disciplines.

4.1 Annotation schema

What are investigating?

The main question we had to answer was '*What are we researching?*' Harmonizing the concepts between different disciplines required detailed discussion on our understanding, definitions and possibilities to adjust to others' needs. First, the idea was to research hate speech, but noticeable divergence occurred at this stage. From Sociologists' point of view, hate speech term is closely related to the social power concept and is taking into account the social position of the speaker and targets of such speech. European Commission against Racism and Intolerance (ECRI) defines hate speech as a speech that: "entails the advocacy, promotion or incitement of the denigration, hatred or vilification of a person or group of persons, as well any harassment, insult, negative stereotyping, stigmatization or threat of such person or persons and any justification of all these forms of expression – that is based on a non-exhaustive list of personal characteristics or status that includes "race", color, language, religion or belief, nationality or national or ethnic origin, as well as descent, age, disability, sex, gender, gender identity and sexual orientation" (European Commission against Racism and Intolerance, 2016). The focus in Sociological sense is therefore on the background of the person or group that is a target of hate speech. Additionally, Social Sciences' research of hate speech is usually in relation to its legal aspects considering the current legal practice in this field as an important criterion for categorization of hate speech. In Slovenia, Public incitement to hatred, violence or intolerance is a criminal offense under the Article 297 of Criminal Code (KZ-1, 2008), but the conditions for prosecution are more specific than just general incitement, taking into account also how radical the speech is, how likely it will encourage a concrete hostile act, and previously mentioned social position of the target. According to the Supreme State Prosecutor's Office's "Position on the prosecution of the criminal offense of Public Incitement to Hatred, Violence or Intolerance under Article 297 of Criminal Code" (2013), public incitement to hatred, violence or intolerance should generally be expressed towards disprivileged, vulnerable social groups, or minorities, that are deprived of political and social power in a certain society, and whose inequality is further deepened by such speech.

Accordingly, the categorization of hate speech from the Sociologists' point of view is a very complex task that surpasses the sole content analysis. On the other hand, Linguistics and Computational linguistics experts needed a categorization that would separate hateful speech from non-hateful one, using broader definition without a relation to social groups belonging and social relationship between the speaker and the target.

For them, the focus was on a discourse that generally expresses discriminatory attitudes and hatred (Baider et al., 2017). Considering different approaches and definitions, we did not want to use the term 'hate speech', since no matter which discourse exactly we were about to cover with this term, it would not be accurate enough for at least one of the disciplines. This led us to introduce a new umbrella term – socially unacceptable discourse (SUD) which covers the broad definition of hateful discourse that we wanted to analyze.

Who is the target?

The question which targets are we interested in was closely related to the definition the individual disciplines used. Within that, Sociologists needed a distinction between the targets attacked because of their background and the other targets, either individuals or groups that are not socially protected or potentially disprivileged. They especially wanted to focus on chosen minorities (LGBT and migrants), so those had to be specifically labeled. Given the more general definition of SUD that the other two disciplines used, a distinction between other several target groups was also desired, but again had to be relevant according to the expected targets of hateful discourse online. The agreement on that was reached with a common expectation that the most usual targets, besides the subjects of the main article posted (in our case LGBT or migrants), were the media outlet or journalists and other commenters. As Hammod and Abdu-Rassul (2017) noticed, many commenters responding to other commenters' comments are indeed using some kind of aggression towards each other.

Should we consider the context?

Different understandings of the concept of discourse produced a dilemma of how much of the context of the individual comment we should consider during the manual annotation. For the Linguistic and Sociological analysis, the social, cultural and historical context are a crucial part of each text, assuming that the content often cannot be properly understood without knowing the background of what is expressed. Even though for the machine learning process this was not preferable, given the importance of the context for the message delivered, we choose to consider it.

Our dataset enables looking into the textual context as well, since the annotators were able to read the title of the main article as well as other previous comments, giving them an insight into what the conversation was about. In the end, all three disciplines agreed that the context should be included due to its importance as influencing factor.

Do we include borderline cases?

As much as sociological definition of hate speech is narrowing down the concept regarding the targets, it is, on the other hand, quite broad when it comes to the interpretation of message that the text is delivering. Researching hatred, Sociologists are also interested in indirect hateful messages, oblique allegations, and negative stereotyping that are reflected as everyday discrimination or remarks directed towards a person solely based on his or her belonging to a specific social group. For a cooperation with experts from Linguistics, though, this was not entirely desirable since they wanted a clear distinction between different levels of hatred expressed in the comments. The

agreement was that indirect messages can be considered unacceptable, but not when this would be too oblique to understand it as hateful. We also choose not to include the cases where the commenter only agreed with a hateful message, but did not (re)produce SUD in any form.

The solution

Considering all the needs and divergences described above, a complex two-level schema was designed that allowed grouping the annotated comments in a way to cater to the research needs of all the domains involved. On the first level, it distinguishes six types of speech according to the radicality of the content and according to why was the target assaulted:

- Acceptable speech
- Inappropriate speech
- Background – offensive speech
- Background – violence
- Other – offensive speech
- Other – violence

On the second level, one of the five different target groups needs to be chosen:

- Migrants/LGBT
- Related to migrants/LGBT (their supporters or alleged supporters)
- Journalist or media
- Commenter
- Other

4.2 Annotation process

Following the described annotation schema, 32 annotators, trained specially for the given task by our experts, started the annotation process for Slovene comments. They were working via online crowdsourcing tool PyBossa, which has its drawbacks, but is recognized as a useful tool for working with a large group of annotators. The main post text, published by the media outlet, and all the comments bellow it were displayed and annotators individually chose a type and a potential target for each of the comments. Their work was monitored by technical team, regularly extracting the information on their progress and agreement ratio, while a Social Sciences expert was analyzing the cases where the agreement was the lowest and giving the annotators advices and directions on how to improve their work.

Only after working with annotators as a fourth group of participants, and after the analysis of a significant amount of actual cases, some new dilemmas arose. We found that a certain amount of subjectivity will always be present when deciding on the degree of hatefulness of the text, so more annotations for one comment has proven to be a good solution, enabling the researchers to use the modal category when analyzing the data later. Authentic communication is also unpredictable – sometimes it is hard to understand, since the context might not be available or it can abuse several targets. Some of

the cases with the lowest agreement had to be additionally checked and annotated by experts.

5 Main takeaways on interdisciplinary collaboration

In this section, we discuss the main conclusions that will guide our collaboration efforts in our future projects. They are based on positive and negative experience from the project and are arranged into four categories which convey our main takeaways.

5.1 Take the time

Immerged into specific research questions and occasionally overwhelmed with administrative work, we saw project group meetings often as an unpleasant necessity, rather than a beneficial opportunity. Looking back, we see that cutting back on the time for discussions (of the whole project group or its parts) leads into misunderstanding that could otherwise be prevented. Consequently, what seemed at the beginning as time-saving measures, proved at the end as time-consuming ones. Moreover, our experience shows that not only the regularity of meetings, but their structure is of equal importance. We saw that our project group worked best on semi-structured meetings where the time was divided between the presentation of progress, pre-prepared Q&A time and ample time for open discussion. This last part proved especially beneficial in the initial phases of the project when we needed to negotiate the scope of the research and best approaches to dataset creation.

Being eager to start early, we immediately dived into work on annotation schema and started with a small sample of real-life data and some made-up examples. This is a perfectly suitable approach for certain phenomena, but it was soon clear that it is not the optimal approach for research on SUD. In our case, data collection and annotation has been a highly elaborated process since affective spontaneous discourse is highly unpredictable and often hard to understand even in the context. In the first phases, this process was even more complex since the guidelines accompanying the schema have been quite basic. In the later phases, we have added several special cases to the guidelines with expert explanations of the most appropriate tag. In our future projects, in order to lower the complexity of the annotation process, we will try to work on a considerable amount of real-life data from the beginning and reserve more time for testing the schema and for brainstorming sessions in order to improve the schema before the official launch of the annotation campaign.

It is inevitable that an interdisciplinary team of researchers will have different approaches to data management and different understanding of the importance of various interventions into the dataset. A rich dataset, such as ours, might not get properly used if its elements are not adequately recorded. When working on a common dataset, it is not only important to discuss any interventions beforehand, but it is also crucial to keep the notes on the interventions updated. We learned this by resolving the question how to deal with comments with two modal categories and how to mark them for later use. This question needed a lot of coordination between the individual research teams inside

the project since we did not share the common view on the usefulness of such comments. What was understood as an important detail in the sociological field, was perceived mainly as noise for (computational) linguists.

5.2 Listen to each other and stay open

Complex research problems, such as SUD, that surpass the domain of single scientific field require interdisciplinary approach. In fact, for a comprehensive description of such phenomena, it might not be enough to stick only to one own standard research techniques, but it might be beneficial to adopt and adapt techniques and approaches from other fields. In our project we thus first tried to share among ourselves the more general aspects that represent the strong points of each domain, such as the strictness in methodology from sociology, focus on qualitative interpretation from linguistics and goal-orientation from computer linguistics. In addition, we exchanged analytical techniques between the disciplines, for example corpus linguistic techniques were adopted by sociologists, while sociological survey and inferential statistical methods were adopted by linguists.

If special care is given to listening to the research needs and hesitations of all the researchers involved, the whole team can greatly benefit from this as was the case in our project. On the one hand, through careful listening and discussions we learned why certain compromises cannot be accepted by all stakeholders despite being reasonable to all the others (e.g., in order to respect the established concepts in sociology, we opted for new term – SUD – instead of sticking to the well know but nonunanimously defined term of hate speech). On the other hand, we observed that it is only possible to correctly interpret the findings and appropriately process the data if we are informed of as many aspect of the phenomenon as possible (e.g., Social Sciences experts helped the whole team understand what are the sensitive aspects of the data and raised awareness regarding the legal and ethical considerations that need to be taken into account when working with SUD data like the need for anonymization, the limitations regarding subsequent related data collection or the need for psychological support for annotators).

5.3 Make sure to have the terminology straight

Despite being aware from the beginning that SUD is first and foremost a concept from Social Sciences, we needed quite some time to really set the terminology and definitions to be used in our project. The main difficulty probably originated from the fact that SUD is a phenomenon that all of us frequently come across in our everyday life and thus we unconsciously felt that we know what our research problem really encompasses. However, experiencing something in everyday life is not the same as approaching it scientifically, and we can say that, at the beginning, we did not consider this aspect seriously enough. Initially, we wanted to stick to one of the existing terms in order not to introduce even more complexity into the already terminologically very varied field of research. But given the seeming familiarity with the studied phenomenon and the fact that scientific definition of hate speech does not correspond with its popular definition, we believe that coining a new umbrella term was a good choice in order to avoid confusion.

It is somewhat clear that in researching complex and not clear-cut phenomena, such as SUD, terminology and the scope of the research needs to be clearly defined in advance, and we even observed that it is welcome to regularly refresh this knowledge with the entire research team throughout the project. However, in an interdisciplinary project, the attention should not only be payed to such special cases as is the definition of the core phenomenon. Despite being tedious, we saw how important it is to avoid using too much discipline-specific jargon in order to ease the understanding of the discussion for the colleagues from other disciplines. Respecting this simple rule had a very positive impact on our work, since the discussions became more inclusive which led to several useful suggestions for future steps in the analysis from different members of the research group.

5.4 Simplify

The work we did on SUD was in many ways a great collaboration experience and an encouraging learning opportunity. One important conclusion is that compromises are inevitable, but that constant negotiation needs to be undertaken in order not to settle for simplistic solutions. This can be seen in the development process of our annotation schema. Even though we initially wanted a simple annotation schema, it was soon clear that a dataset based on such schema would not provide enough information to researchers. For this reason, we initially developed a highly complex schema that proved too complicated for efficient annotation process. This led us to a simplification phase in which we collected several rounds of feedback and use it to curb the schema. After many iterations, we can say that the final version of the annotation schema is simple enough to provide a solid framework for the annotators and a rich output in terms of metadata. It can be applied to different languages and cultures with slight modifications (e.g., with respect to the topic). Nonetheless, it can be further simplified and still remain useful. However, we believe that by better managing our expectations and dedicating more time to discussions and work on real data, we could arrive at such schema earlier.

Throughout the project we learned that simplifying is one of the keys to success, and especially so in interdisciplinary settings. We saw that the results of simplifying are nothing like the process that is needed to arrive to these results. Mainly, it takes a lot of time and we will try to consider this in our next project. In conclusion, we believe that interdisciplinary collaboration requires a step back in expectations of each individual discipline, and a step forward in looking for innovative research questions that intertwine knowledge of the disciplines, rather than just adding findings one beside the other.

Acknowledgement

The work described in this paper was funded by the Slovenian Research Agency within the national 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

1. Bajt, V.: Online hate speech and the »refugee crisis« in Slovenia. In I. Žagar & et al. (Eds.), The disaster of European refugee policy: Perspectives from the »Balkan route«. pp. 133–155. Cambridge Scholars Publishing. (2018).
2. Dragoš, S.: Sovražni govor. Socialno Delo, 46(3), 135–144. (2007).
3. ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W. Y., & Belding, E.: Hate lingo: A target-based linguistic analysis of hate speech in social media. Twelfth International AAAI Conference on Web and Social Media. (2018).
4. Fortuna, P., & Nunes, S.: A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR), 51(4), pp. 1–30. (2018).
5. Gorjanc, V.: Diskurz slovenskih spletnih forumov–dokončen pokop strpnosti. Večkulturnost v Slovenskem Jeziku, Literaturi in Kulturi, 41, pp. 20–29. (2005).
6. Horan, G.: 'You taught me language; and my profit on't/Is, I know how to curse': Cursing and swearing in foreign language learning. Language and Intercultural Communication, 13(3), pp. 283–297. (2013).
7. Jay, T., & Janschewitz, K.: The pragmatics of swearing. Journal of Politeness Research, 4(2), pp. 267–288. (2008).
8. Kapoor, H.: Swears in context: The difference between casual and abusive swearing. Journal of Psycholinguistic Research, 45(2), pp. 259–274. (2016).
9. McEnery, T. Swearing in English: Bad language, purity and power from 1586 to the present. Routledge. (2004).
10. Methven, E. P.: Dirty talk: A critical discourse analysis of offensive language crimes. University of Technology Sydney. (2017).
11. Milosavljevič, M.: Regulacija in percepcija sovražnega govora: Analiza dokumentov in odnosa urednikov spletnih portalov. Teorija in Praksa, 49(1), pp. 112–130. (2012).
12. Nielsen, L. B.: Subtle, pervasive, harmful: Racist and sexist remarks in public as hate speech. Journal of Social Issues, 58(2), pp. 265–280. (2002).
13. Pahor de Maiti, K., & Fišer, D.: Analiza kazalnih zaimkov v družbeno nesprejemljivih spletnih komentarjih. In Slovenščina – diskurzi, zvrsti in jeziki med identiteto in funkcijo. Znanstvena založba Filozofske fakultete. (2020).
14. Pinker, S.: Freedom's curse. The Atlantic Monthly, 302, pp. 28–29. (2008).
15. Schmidt, A., & Wiegand, M.: A survey on hate speech detection using natural language processing. 1–10. (2017).
16. Vehovar, V., Motl, A., Mihelič, L., Berčič, B., & Petrovčič, A.: Zaznava sovražnega govora na slovenskem spletu. Teorija in Praksa, 49(1), pp. 171–189. (2012).
17. Vehovar, V., Povž, B., Fišer, D., Ljubešić, N., Šulc, A., & Jontes, D.: Družbeno nesprejemljivi diskurz na facebookovih straneh novičarskih portalov. Teorija in Praksa, 2, pp. 622–645. (2020).
18. Vidgen, B., & Yasseri, T.: Detecting weak and strong Islamophobic hate speech on social media. Journal of Information Technology & Politics, pp. 1–13. (2019).
19. Zhang, Z., & Luo, L.: Hate speech detection: A solved problem? The challenging case of long tail on twitter. Semantic Web, 10(5), pp. 925–945. (2019).
20. Ljubešić, N., Fišer, D. in Erjavec, T.: The FRENK Datasets of Socially Unacceptable Discourse in Slovene and English. (2019). Available at https://arxiv.org/pdf/1906.02045.pdf
21. European Commission against Racism and Intolerance: ECRI General Policy Recommendation no. 15 on Combating Hate Speech. (2019). Available at https://rm.coe.int/ecri-general-policy-recommendation-no-15-on-combating-hate-speech/16808b5b01

22. Kazenski Zakonik (KZ-1) [Criminal Code of the Republic of Slovenia]: *Uradni list RS*, 50/12. (2008).
23. Vrhovno državno tožilstvo Republike Slovenije [Supreme State Prosecutor's Office]: Pravno stališče o pregonu kaznivega dejanja Javnega spodbujanja sovraštva, nasilja ali nestrpnosti po 297. Členu KZ-1 [Position on the prosecution of the criminal offense of Public Incitement to Hatred, Violence or Intolerance under Article 297 of Criminal Code]. (2013).
24. Baider, F. H., Assimakopoulos, S., & Millar, S. L.: Hate speech in the EU and the CONTACT project. In Online Hate Speech in the European Union: A Discourse-Analytic Perspective (pp. 1–6). Springer. SpringerBriefs in Linguistics: https://doi.org/10.1007/978-3-319-72604- 5_1. (2017).
25. Hammod, N. M. in Abdul-Rassul, A.: Impoliteness Strategies in English and Arabic Facebook Comments. International Journal of Linguistics, 9(5), 97–112. (2017).