Hate speech on Russian forums: SNA analysis

Ekaterina Gryaznova [*]
Saman Hallajian [†]

Abstract

Abstract: This paper is dedicated to hate speech analysis on Russian media forums. The research is performed by using SNA methods, such as network visualization and cluster analysis for EDA, as well as SVD graph embeddings for link prediction with a goal to classify target groups of hateful messages. The results of EDA helped grasp a better understanding of vulnerable groups on russian forums and distribution of hateful words used. The basic classifier based on SVD graph embeddings performed poorly, however, it leaves a lot of space for future research on this topic.

Keywords-hate speech, cyberbullying, SNA, SVD, link prediction

Contents

[*] esgryaznova@edu.hse.ru
[†] skhalladzhian@edu.hse.ru

1 Introduction

Cyberbullying is one of the most widely researched topics connected to social media, as it can have a huge impact on an individual and possess a real threat to one's mental health and well-being, while being hard to trace and control due to its often anonymous nature. Social media platforms have "passive reporting mechanisms", and user guides for taking cyberbullying under control, however there is no system to automatically recognize hate speech [6].

The goal of this paper is to contribute to understanding hate speech on Russian web forums using social network analysis (SNA). Firstly, it can help build automatic systems of hate speech detection on removal. For instance, it can spot a rising violent behavior towards one of the minority groups, and early detection of the public's mood can help prevent real life danger. Secondly, more for sociologists, it can help with deeper understanding of human behavior on the internet.

2 Related Work

There are some different methods to analyze cyberbullies. For instance, in [6], authors used a *Graph Convolutional Network classifier* for predicting the target group of hateful tweets: gender, age, ethnicity, religion or not hateful. The classifier is based on cosine similarities between tweet embeddings, and the results prove that such approach matches or exceeds the results of traditional classifiers, such as Logistic Regression, Naive Bayes and others. Authors of this article argue that cyberbullying detection and classification is an especially important topic during Covid-19 times due to increased screen time.

In another research [4], authors focused on analysing cyberbullying in a form of a Momo Challenge social network analysis (SNA) is performed using NodeXL, where graph metrics are calculated and users are grouped based on their relationship to one another.

Other research focuses less on content and more on authors of the 'malicious content' in cyberspace (cyberbullies) to prevent hate speech online, for instance [1].

An older research, but insightful, SNA was applied to examine the networks of social interactions and cyberbullying among thirteen- to fourteen-year- old pupils [2]. Big focus of this research was on identifying what kind of people are more susceptible to being a cyberbully or a victim.

3 Data

Data for this project, RuHateBe was collected as a part of a research seminar on the Computational Linguistics programme at HSE It is a benchmark dataset for hate speech in Russian. RuHateBe consists of comments collected from ProDota-russian largest gaming forum for Dota game, which has a lot of obscene content, and Dvach-largest russian anonymous imageboard, like 4chan. Among hate datasets there are also some non-hate data from other resources, but they were not used for this research. The dataset contains mainly information about the model's evaluations and predictions of toxicity of each comment. The comments were scraped in the following pattern - a toxic response to someone's comment and the initial comment. For this research we only used the toxic response and the assigned target group. Target groups are as follows: women, men, LGBT+, place of birth (skin color), migrants, children and others.

Minor feature engineering was performed, as we were curious as to in which target group the toxic response tended to follow a non-toxic comment, but it was not used in the SNA analysis.

However it is worth exploring in the future, as an “unprovoked” hateful response can be considered higher on the toxicity scale than a “provoked” one. We attach the results of this in Figure 3 in the appendixA. It can be suggested that the “unprovoked” hate is usually aimed at women’s comments, compared to other groups, however the classes are quite imbalanced and this claim needs to be researched further.

4 Methods and Results

4.1 EDA-building a network of bigrams

For this step, data was prepressed in a classic manner. Toxic comments were lemmatized, tokenized, stop words and symbols were deleted (as this research focuses on language part of web communication only). This network (Figure 1) is a directed graph, where the nodes are words, and the edges are the frequency of two words being together. As the network is too big to analyze, the better quality picture of it can be found in this project’s repo on github[1].

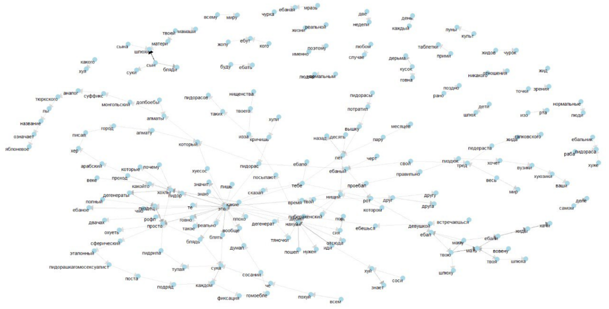

Figure 1: The directed graph of this network

While the graph may be very simple in its core idea, it gives a lot of information and there’s a clear cluster division in some cases. The main conclusions that can be made from the graph are as follows:

- There is almost no cluster for LGBTQ+ target group, as words like “пидорас”, “педрила” are often times used not solely to insult a gay person, but any man.
- Following the first observation, the biggest cluster seems to contain 2 topics at once - women and gay people, connected to words describing rape and sexual acts.
- There is a clear distinction between migrants’ target group and place of birth, as those 2 clusters are almost polar opposite from each other. In the cluster connected to the place of birth target group there are words like “ukrainians”, “mongolians”, “arabs”, but no words describing people from Middle Asian countries. As a vast majority of migrants in Russia are from Middle Asian countries and face a different level of hatred and discrimination, they are clustered separately.

[1] https://github.com/esgryaznova/analyzing-hatespeech-SNA-2022

- The bigram that is used the most often, more than 200 times, is “son of a whore”.
- Following the previous observation, the most interesting cluster in our opinion contains 3 target groups at once. Phrases like “мать твоя шлюха” (“your mom’s a whore”), “твою мать ебали жиды хачи”, are initially aimed to insult a man, but also touch upon women being whores, while using racial slurs.
- While it’s hard to pinpoint a target group for each cluster, as they are often mixed like in the example above, there is an interesting almost-a-cluster that seems to be focused on hate towards higher education, with words hinting on years spent wastefully on higher education.

Despite some informative observations that can be made from this network, aside from the class imbalance, there is also a bias in scraped threads. It is obvious that some clusters or just brunches are formed because of the thread topic, like probably the higher education one, or the discussion of suffixes in mongolian language.

4.2 SVD Embeddings

Graph Embeddings and graph-based neural networks, such as GAT, CNN etc. have recently gained popularity in text classification tasks. With SVD we try to classify hateful speech into target groups, by predicting the link between two entities in the network. The following figure (2) shows what the network looks like, where sources are hateful comments, and targets are target groups.

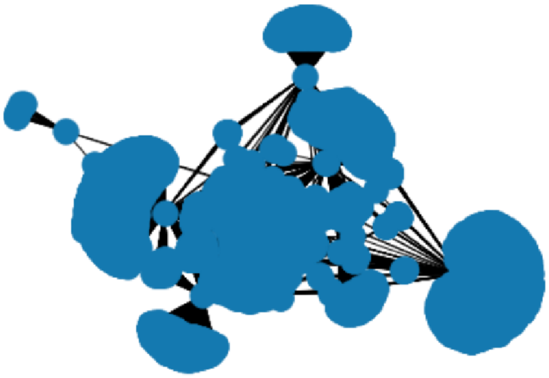

Figure 2: The shape of this network

The size of our data turned out to be quite too small for training. The problem occurred while splitting dataset into train, test and validation subsets, as there were not enough validation edges. Despite this, we still trained the model and ran ROC and AP (average precision) tests, and got results of 0.33 and 0.28 respectively. We then evaluated edge embeddings using Logistic Regression, and got ROC results of 0.50 and AP results of 0.35, which is basically worse than a random guess, however we do have 7 target groups, and as was shown in the EDA part of the research, often times target groups are mixed. Hence we have an idea, or a proposal for future research of this topic, to differentiate between hate aimed at an author of a comment, tweet, text, and hate towards some outsider(s) that do not take part in the discussion.

4.3 Other Results

Graph convolutional network (GCN) was also trained on this data, however, because of the class imbalance, as described above, it couldn't be evaluated and used for prediction.

5 Conclusion

This research demonstrated the complicated nature of hate speech online. While providing some useful insights into targets of cyberbullying on russian forums, it also exposed some biases in RuHateBe data and ambiguous side of hateful comments on toxic russian forums. As for future sociological research on this topic: on the one hand, anonymous forums like Dvach, 4chan seem perfect for cyberbullying analysis as they are incredibly toxic, on the other hand, this toxicity is an exception among other social media platforms, and perhaps does not provide an objective picture. In the future, we recommend research to take into consideration "target" when analyzing communication between online users. It also tapped into using graph embeddings based models for text classification, but we didn't achieve a high result just yet. With more balanced data, mindful approach to defining a target of hate speech and better understanding of new methods, we hope to make some progress in the future.

References

[1] Yoon-Jin Choi, Byeong-Jin Jeon, and Hee-Woong Kim. Identification of key cyberbullies: A text mining and social network analysis approach. *Telematics and Informatics*, 56:101504, 09 2020.

[2] Steven Eggermont, Heidi Vandebosch, and Denis Wegge. Who bullies whom online: A social network analysis of cyberbullying in a school context. *Communications*, 39(4):415–433, 2014.

[3] Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. Abusive language detection with graph convolutional networks. *CoRR*, abs/1904.04073, 2019.

[4] Swaranjit Singh, Vivek Thapar, and Sachin Bagga. Exploring the hidden patterns of cyberbullying on social media. *Procedia Computer Science*, 167:1636–1647, 2020. International Conference on Computational Intelligence and Data Science.

[5] I-Hsien Ting, Wun Sheng Liou, Dario Liberona, Shyue-Liang Wang, and Giovanny Mauricio Tarazona Bermudez. Towards the detection of cyberbullying based on social network mining techniques. In *2017 International Conference on Behavioral, Economic, Socio-cultural Computing (BESC)*, pages 1–2, 2017.

[6] Jason Wang, Kaiqun Fu, and Chang-Tien Lu. Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1699–1708. IEEE, 2020.

A Appendix

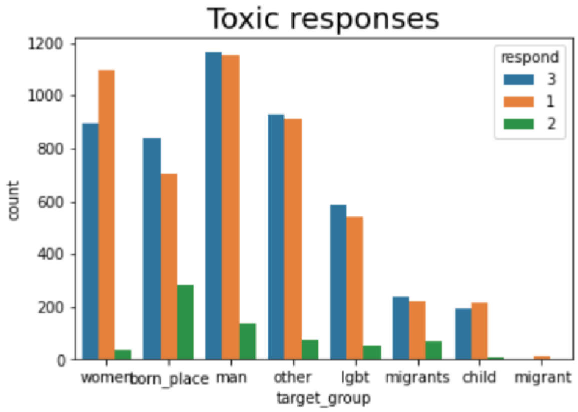

Figure 3: Vissualization of the data we used in 3