Detection of Low-toxic Texts in Similar Sets Using a Modified XLM-RoBERTa Neural Network and Toxicity Confidence Parameters

Yaroslav A. Seliverstov
Solomenko Institute of Transport Problems of the Russian Academy of Sciences
Saint Petersburg, Russia
University 20.35
Moscow, Russia
silver8yr@gmail.com

Andrew A. Komissarov
University 20.35
Moscow, Russia
Andrew.Komissarov@gmail.com

Eleonora D. Poslovskaia
Saint Petersburg State University
Saint Petersburg, Russia
el.poslovskaya@gmail.com

Alina A. Lesovodskaya
University 20.35
Moscow, Russia
a.lesovodskaya@2035.university

Artur V. Podtikhov
University 20.35;
Higher School of Economics — National Research University
Moscow, Russia
a.podtikhov@2035.university

2021 XXIV International Conference on Soft Computing and Measurements (SCM) | 978-1-6654-3974-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/SCM52931.2021.9507117

***Abstract*—The article considers the problem of classifying low-toxic texts using a modified neural network of the XLM-RoBERTa transformer architecture, trained on highly toxic texts. Comments from the School of Pedagogical Design at the University of 20.35 were used as kits for identifying low-toxic texts. The network was not retrained on low-toxic texts. Instead, the classification of low-toxic texts was carried out by varying the toxicity confidence parameter. An approximation dependence of the number of low-toxic texts on the parameter of toxicity reliability was constructed and a threshold value of the toxicity reliability parameter was obtained, at which the quality of the classification of low-toxic texts is maximal. The hypothesis of the similarity of the toxicity of homogeneous information resources was also formulated and confirmed**

Keywords—classification, deep learning, internet content, toxicity, XLM-RoBERTa

I. Introduction

Communication is one of the most important processes of interaction between people in modern society. The rapid development of information technologies, communications and social networks over the past decade has transformed the environment of human communication, moving the process of everyday communication into the virtual Internet space [1]. The virtual world changes the specifics of interpersonal communication. Virtual communication is global in nature and differs from real interaction by anonymity, multilingualism, mediation, uncontrollability, lowering the limits of moral and social boundaries, which can lead to a wide class of hostile communicative actions - bullying, threats, belittling business reputation, ridicule, extortion, slander, humiliation honor and dignity, insults, manifestations of hostility and discrimination on the basis of social, religious, gender and national intolerance, drug propaganda, calls for terrorist and extremist activities, suicide, civil disobedience and various forms of deviant behavior. The dissemination of such toxic information can cause irreparable damage not only to an individual or company, but also to the country as a whole. That is why solving problems related to identifying and preventing the dissemination of information that leads to negative consequences is a step that is highly relevant and timely [2].

II. Related Work

Currently, on the basis of modern deep learning methods [3, 4, 5], new software models are being actively developed and tested, capable of recognizing and filtering illegal and malicious content.

Since 2020, pretrained neural networks of transformer architecture have been actively used to identify toxic content, such as mBERT [6], XLM [7], ruBert [8], M-USE [9], which are mainly used to solve problems of interlanguage understanding [10].

The article [11] reveals aggression in user messages using a binary classifier based on a random forest algorithm and a convolutional neural network.

In [12], it was proposed to use an uncontrolled probabilistic method with a source dictionary to identify offensive comments on social networks in Russian and Ukrainian.

In [13], based on neural network models of transformer architecture, a new method of multiclass classification of threats in Russian is developed, which allows to reduce the number of false-positive predictions.

To develop high-quality multiclass classifiers for the content of malicious content on web pages, it is necessary to have corpuses marked with the appropriate classes. For the Russian language, there are currently only a few open sets, namely:

1) a set of offensive comments in Russian and Ukrainian, about 2000 in size [12];
2) an open set of Russian-language toxic comments published on the Kaggle resource [14];
3) the same set [14], additionally marked and verified by assessors from Toloka [8];

4) a marked set of threats collected on the Vkontakte social network [13].

Among the latest scientific studies in the field of toxicity carried out by international research teams, the following works should be noted.

In [15], a model is developed for detecting and classifying hostile posts and their further classification into fakes, insults, hate and slander using a convolutional relational network graph. The model proposed by the authors works at the XLM-RoBERTa level from Google on this dataset.

In [16], a team of authors presented a new dataset for detecting hostility in the Hindi language, consisting of 8,200 online messages. The annotated dataset covers four aspects of hostility: fake news, hate speech, and offensive and libelous messages.

In [17], the authors presented the ALONE[1] multimodal dataset on toxic social media interactions among high school students, along with descriptive explanations. Each interaction includes tweets, images, emoji's, and associated metadata.

In [18], an approach based on transfer learning of pretrained neural networks is presented for classifying messages on social networks (such as Twitter, Facebook, etc.) in Hindi Devanagari as hostile or unfriendly.

The analysis of the subject area indicates the relevance of research on the detection of highly toxic content in social networks and web resources using pretrained neural networks of transformer architecture.

III. Task & Evaluation

The general trend of negative research in the Internet space is focused on the identification of toxic and highly toxic content, which is usually distributed on social networks and frequently visited web resources. Meanwhile, there is a large number of educational and specialized web resources, which are characterized by a different type of user. As a rule, these are educated people with higher education, involved in various types of labor activities. Such users are characterized by good manners, restraint in statements and expressions of emotion. Despite this fact, heated discussions also arise on these web resources, characterized not by highly toxic, but by low-toxic statements – ridicule, sharp jokes, provocative statements and hidden injections. Unfortunately, at the moment there are no marked-up corpuses and sets of similar low-toxic texts necessary for constructing classifiers in the public domain.

The main goal of the work is to study the possibility of detecting low-toxic texts using a binary classifier model based on a modified neural network of the XLM-RoBERTa transformer architecture, trained only on toxic and highly toxic texts and regulating the toxicity reliability parameter, taking into account the principle of similarity of homogeneous Internet resources.

The comments of the School of Pedagogical Design at 20.35 University is used as low-toxic texts.

[1] ALONE - AdoLescents ON twittEr

IV. Datasets & Methods & Models

The analysis of the subject area showed that it is advisable to develop the classifier model on the basis of the multilingual XLM-RoBERTa[2] transformer [12].

The following architectural changes were made to the XLM-RoBERTa-m[3] model - the masked-Im layer was replaced with a fully connected layer. The open platform Kaggle was used as a computing environment.

The algorithm for detecting low toxicity responses based on the XLM-RoBERTa-m model and the toxicity confidence parameter is shown in Fig. 1.

The XLM-RoBERTa-m model was trained on data from the Jigsaw Multilingual Toxic Comment Classification competition[4] held on the Kaggle platform. The data consisted of training (jigsaw toxic comment train.csv), validation (validation.csv), and test (test.csv) sets. The training sample contained 223,549 annotated user comments taken from the discussion pages on Wikipedia. This set is the largest publicly available corpus on Kaggle.

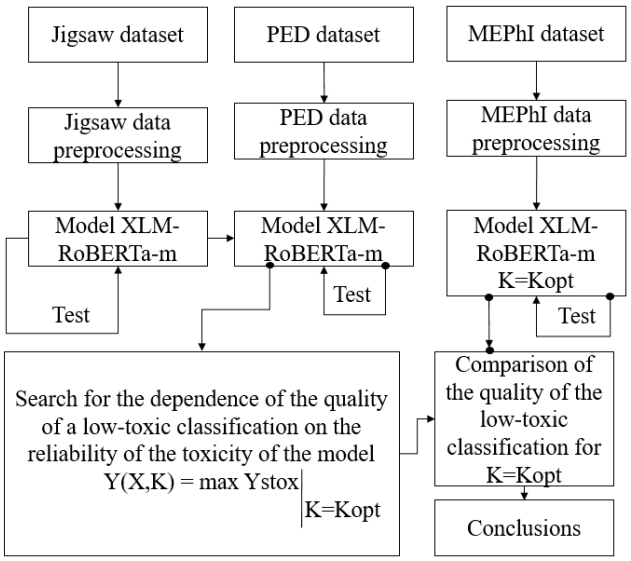

Fig. 1. Algorithm for detecting low-toxic reviews based on the principle of similarity of homogeneous Internet resources

The comments presented were categorized by experts into six classes: Toxic, Very Toxic, Insult, Threat, Obscene, and Identity Hate. Comments can be associated with several classes at the same time, which forms a multicomponent classification.

Before processing text data, the following text preprocessing methods are used: punctuation removal, lemmatization, and stop word removal.

The quality of the binary classification of the trained model on the test data was AUC ROC = 0.9459.

The confidence distribution of toxicity using XLMRoberta-m for the test dataset is shown in the histogram in Fig. 2.

[2] fairseq/examples/xlmr at master pytorch/fairseq GitHub
[3] Index "m" means modification of the XLM-RoBERTa neural network model
[4] https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification/data

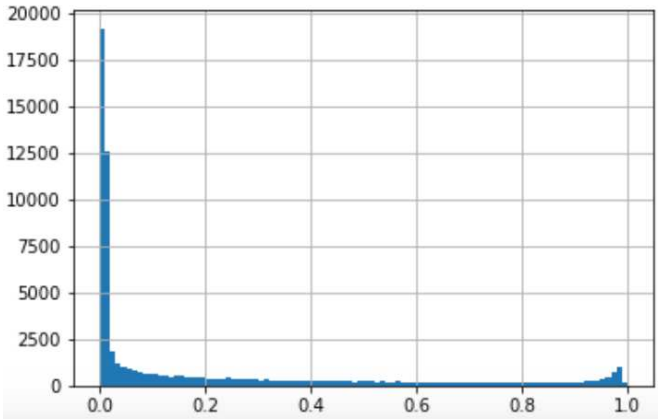

Fig. 2. Distribution of toxicity confidence on the XLMRoberta-m model for the Jigsaw test dataset

An example of toxic comments filtered by the XLM-RoBERTa-m model is shown in Fig. 3.

id			
216	216	ru	I don't understand ... what an idiot you have to be to ...
219	219	ru	You fucked up, you'll still be spanking me ...
412	412	ru	Frequently absolutely delusional articles, there are no ...
428	428	ru	You are wrong because you are a fool (s). Oh lord, shut up
...	...	ru	....
63620	63620	ru	What idiot wrote the article? Counter-Strike is resting ..
63672	63672	ru	Dart is an Internet fool. Where in the rules ...
63677	63677	ru	I was never interested in either homophobes or ...ни...
63730	63730	ru	and preferably those who have completed the service ...
64769	64769	ru	During your work on Wikipedia, you have shown ...

Fig. 3. Example of toxic comments from the Jigsaw dataset

Further, the corpus of the School of Pedagogical Design of the University 20.35 with low-toxic statements was submitted to the input of the model as a test set. The corpus consisted of 54352 records.

The distribution of toxicity confidence using XLMRoberta-m for the School of Pedagogical Design test dataset is shown in the histogram in Fig. 4.

Figure 4 shows that in the range of values from 0.4 to 0.6 of the model toxicity reliability parameter, there is an increase in comments. Manual review of the test kit confirmed that some of the comments were of low toxicity.

Let us formulate a hypothesis of the similarity of the toxicity of homogeneous information resources in relation to educational Internet resources.

Hypothesis. User comments on similar Internet resources (for example, educational) have a similar level of toxicity (low- toxicity[5]).

To test the hypothesis, we will perform the following steps:

1) We will construct an approximate dependence of the number of reviews on the value of the model's reliability based on the control points selected by the experts, focusing on Fig. 4.

Control points: 0.03; 0.52; 055; 0.6; 0.75.

Quadratic, exponential, exponential, cubic and logarithmic regressions were considered as approximation functions.

The analysis was carried out using the PLANETCLAC [6].

[5] By a similar level of low toxicity, we mean a similar number of low toxicity comments.

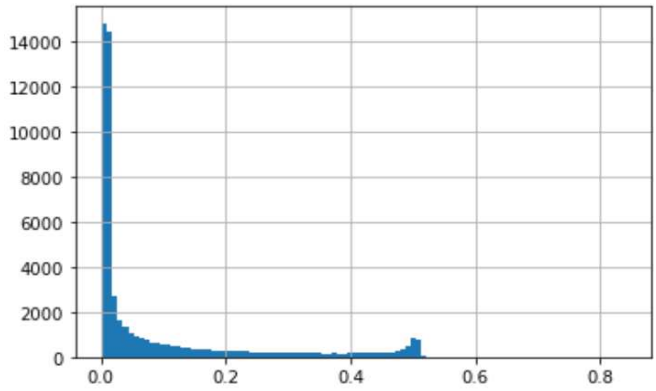

Fig. 4. Distribution of toxicity confidence on XLMRoberta-m for the Pedagogical Design school dataset

As a result of the analysis, the best approximation accuracy was shown by cubic regression (Table 1).

The cubic regression equation is:

$$Y = -103642.5563x3 + 195274.8384x2 - - 121590.6997x + 25089.7540$$

TABLE I. INITIAL (Y) AND APPROXIMATING (YREG) VALUES OF THE NUMBER OF REVIEWS FROM THE MODEL RELIABILITY VALUE (X)

	Approximating values				
X	0.03	0.52	0.55	0.6	0.75
Y	21615	85	54	42	15
Yreg	21614.98	91.93	41.97	47.48	14.62

2) with the help of experts, we will determine the threshold value of the toxicity reliability parameter at which the number of low-toxic responses is maximum, quantitatively and qualitatively comparing the sets of low-toxic responses identified at different values of the toxicity reliability with the reference set.

As a result of comparison, the threshold value of the toxicity reliability parameter was determined as K = 0.52.

The graph of the approximating function with the marked threshold value of the toxicity confidence parameter is shown in Fig. 5.

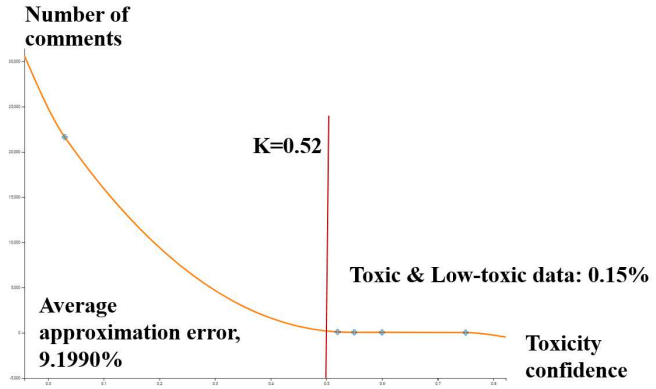

Fig. 5. Graph of the approximating function

The number of filtered reviews was 85 (see Fig. 6).

[6] https://planetcalc.ru/5992/?xstring=0.03%200.52%200.55%200.60%200.75&ystring=21615%2085%2054%2042%2015&dolinear=0&doquadratic=1&dopower=1&docubic=1&doexponential=1&dologarithmic=1&dohyperbolic=0&doeexponential=1

id	lang	comment_text	id
12735	ru	Shut up!	12735
18488	ru	Losers!	18488
37176	ru	Bullshit - sophomores are being hunted for.....	37176
45188	ru	Get lost!	45188
48358	ru	Why fly up	48358
49353	ru	Well, so many defenders - someone has to shoot	49353
49664	ru	Yes, I answer all the stupid losses there...	49664
50034	ru	Now I will put your twitching eye in place..	50034
50459	ru	Freak, puts pluses to everyone..	50459
51417	ru	Damn, well then remove me from this shitty site	51417
51441	ru	Sinister!	51441
52551	ru	Turn it on here, you dumb ones!	52551
52902	ru	As a rule, losers in reflection write about their failures and..	52902
53784	ru	Where did so many slow-witted and brakes come from?	53784

Fig. 6. Low-toxic comments found on the Pedagogical Design school dataset

Thus, in the set of 54352 comments from the Pedagogical School of the University 20.35, there are low-toxic comments in the amount of 0.156 %

$$\text{Tox (Ped)} = (85 * 100) / 54352 = 0.156\ \%$$

3) evaluate the performance of our model on a new test dataset of similar educational topics and the threshold value of the toxicity reliability obtained at step 2 equal to K = 0.52.

As a new test set, we will use the comments of the MEPhI school. The corpus includes 9929 entries.

The number of filtered reviews for the given values was 14 (see Fig. 7).

Thus, in the set of 9929 comments from the MEPhI school, there are low-toxic comments in the amount of 0.141 %

$$\text{Tox (MEPhI)} = (14 * 100) / 9929 = 0.141\ \%$$

id	lang	comment_text	id
12	ru	everything is fucked up!	12
2097	ru	Managers and HRs have a lousy knowledge of math	2097
3272	ru	Faith and everything else is complete nonsense.	3272
3531	ru	In general, the abundance of tools is both good and disgusting.	3531
4461	ru	How can! Everything is the same! How much you can endure this boring!	4461
4984	ru	Friends! Help register for the event for free.	5297
5297	ru	Bore! Aren't you tired of talking?	5518

Fig. 7. An example of revealed weakly toxic comments in the set of the MEPhI school

The magnitude of the discrepancy in the number of detected low-toxic responses from two different educational schools is less than 1 %.

V. Conclusion

As a result of research work on the analysis of comments and reviews of online participants of the School of Pedagogical Design at the University of 20.35, the problem of identifying and classifying low-toxic texts was solved using a modified neural network of the XLM-RoBERTa-m transformer architecture trained highly toxic texts and regulating the value of the toxicity reliability parameter.

The hypothesis of the similarity of the toxicity of homogeneous information resources and the efficiency of the algorithm for detecting low-toxic responses based on the principle of similarity are formulated and, in a first approximation, confirmed. The last statement is especially relevant, since in the case of an analysis of similar Internet resources for low toxicity and the absence of marked low toxic sets, one can try to identify weakly toxic comments, relying only on the trained neural network of the transformer architecture and the model reliability parameter.

VI. Acknowledgements

The team of authors would like to thank the leading specialist of OCRV LLC, the natural language processing group Alexei Shonenkov for valuable recommendations and comments during the work on the XLM-RoBERTa neural network model.

References

[1] Thomas B., Thomas V.V. The impact of Internet communications on social interaction. Sociological Spectrum. 2005, vol. 25. No 3, pp. 335-348. DOI: 10.1080/02732170590925882.

[2] Naslund J.A., Bondre A., Torous J. et al. Social Media and Mental Health: Benefits, Risks, and Opportunities for Research and Practice. Journal of Technology in Behavioral Science. 2020, No 5, pp. 245–257 https://doi.org/10.1007/s41347-020-00134-x

[3] Seliverstov Y., Seliverstov S., Malygin I., Korolev O. Traffic safety evaluation in Northwestern Federal District using sentiment analysis of Internet users' reviews. Transportation Research Procedia. 2020, vol. 50, pp. 626–635. DOI: 10.1016/j.trpro.2020.10.074.

[4] Aken B. van et al.: Challenges for Toxic Comment Classification: An In-Depth Error Analysis. Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). 2018, pp. 33–42.

[5] Risch J., Krestel R. Toxic comment detection in online discussions. In book: Deep learning-based approaches for sentiment analysis. Springer, 2020, pp. 85–109. DOI:10.1007/978-981-15-1216-2_4.

[6] Devlin J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (NAACL-HLT). 2019. pp. 4171–4186. DOI:10.18653/v1/N19-1423.

[7] Lample G., Conneau A. Cross-lingual language model pretraining. 32rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada, arXiv preprint arXiv:1901.07291, 2019. https://arxiv.org/pdf/1901.07291.pdf.

[8] Smetanin S. Toxic Comments Detection in Russian. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference, Dialogue 2020, DOI:10.28995/NNNN-NNNN-2020-19-1-11.

[9] Yang Y et al. Multilingual Universal Sentence Encoder for Semantic Retrieval. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 87–94. In arXiv preprint arXiv:1907.04307.

[10] Conneau Alexis et al. Unsupervised Cross-lingual Representation Learning at Scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451. DOI:10.18653/v1/2020.acl-main.747

[11] Potapova R., Gordeev D. Detecting State of Aggression in Sentences Using CNN. International Conference on Speech and Computer, 2016, pp. 240-245. DOI: 10.1007/978-3-319-43958-7_28. arXiv:1604.06650v1 [cs.CL] 22 Apr 2016

[12] Andrusyak B, et al.: Detection of abusive speech for mixed sociolects of russian and ukrainian languages. The 12th workshop on recent advances in slavonic natural languages processing, RASLAN 2018. pp. 77-84.

[13] Zueva N., Kabirova M., Kalaidin P. Reducing Unintended Identity Bias in Russian Hate Speech Detection. Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 65–69, https://doi.org/10.18653/v1/P17, arXiv preprint arXiv:2010.11666

[14] Belchikov A. Russian language toxic comments, https://www.kaggle.com/blackmoon/russian-language-toxic-comments.

[15] Davidson T., Warmsley D., Macy M., Weber I. Automated Hate Speech Detection and the Problem of Offensive Language. 2017. https://arxiv.org/pdf/1703.04009.pdf

[16] Mohit Bhardwaj, et al. Hostility Detection Dataset in Hindi. 2020. https://arxiv.org/pdf/2011.03588v1.pdf

[17] Wijesiriwardene Thilini, Inan Hale, Kursuncu Ugur, Gaur Manas, Shalin Valerie, Thirunarayan Krishnaprasad, Sheth Amit, Arpinar Ismailcem. ALONE: A Dataset for Toxic Behavior among Adolescents on Twitter. 2020. https://arxiv.org/abs/2008.06465

[18] Gupta Ayush, Sukumaran Rohan, John Kevin, Teki Sundeep. Hostility Detection and Covid-19 Fake News Detection in Social Media. 2021. https://arxiv.org/abs/2101.05953.