Reducing Unintended Identity Bias in Russian Hate Speech Detection

Nadezhda Zueva[1]**, Madina Kabirova**[1]**, Pavel Kalaidin**[1,2]
[1]VK, [2]VK Lab
`{firstname.lastname}@vk.com`

Abstract

Toxicity has become a grave problem for many online communities and has been growing across many languages, including Russian. Hate speech creates an environment of intimidation, discrimination, and may even incite some real-world violence. Both researchers and social platforms have been focused on developing models to detect toxicity in online communication for a while now. A common problem of these models is the presence of bias towards some words (e.g. woman, black, jew or женщина, черный, еврей) that are not toxic, but serve as triggers for the classifier due to model caveats. In this paper, we describe our efforts towards classifying hate speech in Russian, and propose simple techniques of reducing unintended bias, such as generating training data with language models using terms and words related to protected identities as context and applying word dropout to such words.

1 Introduction

With the ever-growing popularity of social media, there is an immense amount of user-generated online content (e.g. as of May 2019, approximately 30,000 hours worth of videos are uploaded to YouTube every hour[1]). In particular, there has been an exponential increase in user-generated texts such as comments, blog posts, status updates, messages, forum threads, etc. The low entry threshold and relative anonymity of the Internet have resulted not only in the exchange of information and content but also in the rise of trolling, hate speech, and overall toxicity [2].

Harassment is a pervasive issue for most online communities. A Pew survey conducted in 2014[3] found that 73% of Internet users have witnessed online harassment, and 40% have personally experienced it.

Explicit policies against hate speech can be considered an industry standard[4] across social platforms, including platforms popular among Russian-speaking users (e.g. VK, the largest social network in Russia and the CIS[5]).

The study of hate speech, in online communication in particular, has been gaining traction in Russia for a while now due to it being a prevalent issue long before the Internet (Lokshina, 2003). The number of competitions and workshops (e.g. HASOC at FIRE-2019; TRAC 2020; HatEval and OffensEval at SemEval-2019) on the topic of hate speech and toxic language detection reflect the scale of the situation.

Social platforms utilize a wide variety of models to detect or classify hate speech. However, the majority of existing models operate with a bias in their predictions. They tend to classify comments mentioning certain commonly harassed identities (e.g. containing words such as woman, black, jew or женщина, черный, еврей) as toxic, while the comment itself may lack any actual toxicity. Identity terms of frequently targeted social groups have higher toxicity scores since they are found more often in abusive and toxic comments than terms related to other social groups. If the data used to train a machine learning model is skewed towards these words, the resulting model is likely to adopt this bias[6].

Inappropriately high toxicity scores of terms related to specific social groups can potentially negate the benefits of using machine learning models to fight the spread of hate speech. This motivated us to work towards reducing these biases. In

---

[1]https://vk.cc/aANMR4
[2]https://vk.cc/aANMZn
[3]https://vk.cc/aANN6p

[4]https://vk.cc/aANNbQ
[5]https://vk.cc/ayxecu
[6]https://vk.cc/aANNqT

this paper, our main goal is to reduce the false toxicity scores of non-toxic comments that include identity terms empirically known to introduce model bias.

2 Related Work

2.1 Hate Speech Detection in Russian

Little research has been done on the automatic detection of toxicity and hate speech in the Russian language. Potapova and Gordeev (2016) used convolutional neural networks to detect aggression in user messages on anonymous message boards. Andrusyak et al. (2018) proposed an unsupervised technique for extending the vocabulary of abusive and obscene words in Russian and Ukrainian. More recently, Smetanin (2020) utilized pre-trained BERT (Devlin et al., 2019) and Universal Sentence Encoder (Yang et al., 2019) architectures to classify toxic Russian-language content.

2.2 Reducing Unintended Bias

Dixon et al. (2018) introduced Pinned AUC to control for unintended bias. In this paper, we adopt Generalized Mean of Bias AUCs (GMB-AUC) introduced by (Borkan et al., 2019b), following a study by (Borkan et al., 2019a) showing the limitations of Pinned AUC.

Vaidya et al. (2020) proposed a model that learns to predict the toxicity of a comment, as well as the protected identities present, in order to reduce unintended bias as shown by an increase in Generalized Mean of Bias AUCs. Nozza et al. (2019) focused on misogyny detection, providing a synthetic test for evaluating bias and some mitigation strategies for it.

To our knowledge, there is no published research on reducing text classification bias in Russian.

3 Experiments

3.1 Datasets

For our experiments, we manually collected a corpus[7] of comments posted on a major Russian social network. The mean length of each sample is 26 characters; samples over 50 characters (5% of the total number of samples) were shortened. The corpus consists of 100,000 samples that we randomly split into training, validation and test sets in the ratio 8:1:1. Each comment was assigned a label

---

[7]The corpus is available on request to authors upon submitting a license agreement.

based on whether or not it contained various forms of hate speech or abuse, including threats, harassment, insults, mentions of family members, as well as language used to promote lookism, sexism, homophobia, nationalism, etc.

As benchmarks, we also used a small corpus of 2,000 samples in mixed Russian and Ukrainian collected by (Andrusyak et al., 2018), and a corpus in Russian used by (Smetanin, 2020) (around 14,000 samples).

3.2 Task & Evaluation

We considered the prediction of labels related to hate speech as a task and validated performance using introduced Generalized Mean of Bias AUCs (Borkan et al., 2019b) to analyze whether or not the proposed methods help reduce text classification bias.

3.3 Protected Identities

We manually compiled a list of Russian words related to protected identities. The words were split, based on the type of hate speech used, into the following classes: lookism, sexism, nationalism, threats, harassment, homophobia, and other. Extracts from the full list are provided in Table 1. Total number of words in the list is 214. The full list of protected identities and related words is available here: `https://vk.cc/aAS3TQ`.

3.4 Models

We used a model based on the self-attentive encoder (Lin et al., 2017). We directly feed the token embeddings matrix to the attention layer instead of the bi-LSTM encoder, making it a pure self-attention model similar to the one used in Transformer (Vaswani et al., 2017). An advantage of this architecture is that the individual attention weights for each input token can be interpretable (Lin et al., 2017). This makes it possible to visualize what triggers the classifier, giving us an opportunity to explore the data and extend our list of protected identities. To overcome the problem of out-of-vocabulary words, we trained byte pair encoding (Sennrich et al., 2015) on a corpora of Russian subtitles taken from a large dataset collected by (Shavrina and Shapovalova), and used it for input tokenization.

We also evaluated a CNN-based text classifier (as in (Potapova and Gordeev, 2016)) to use as a baseline for comparison.

lookism
корова korova "cow"
пышка pishka "donut (meaning "plump")"

sexism
женщина zhenshchina "woman"
баба baba "woman (derogatory)"

nationalism
чех chekh ""Chechen" (derogatory) lit. "Czech""
еврей evrei "Jew"

threats
выезжать vyezhat "to come (after somebody)"
айпи aipi "ip"

harassment
киска kiska "pussy"
секси seksi "sexy"

homophobia
гей gay "gay"
лгбт LGBT "LGBT"

other
мамка mamka "mother"
админ admin "admin"

Table 1: Extracts from the full list of protected identities and related words.

3.5 Data Generation with Language Models

To reduce model bias, we propose to extend the dataset with the output of pre-trained language models. We used the pre-trained Transformer language model[8] trained on the Taiga dataset (Shavrina and Shapovalova). As Taiga contains 8 sources of normative Russian text (news, fairy tales, classic literature, etc.), we assumed that the model would be able to generate non-toxic comments even with one word from protected identities given as context. We took a random word from a list of protected identities and related words as a single word prefix for language generation, and generated samples up to 20 words long or until an end token was generated. An additional 25,000 samples were generated using the described approach and added to the existing training set.

3.6 Identity Dropout

Random word dropout (Dai and Le, 2015) was shown to improve text classification. We utilized this technique to randomly (with 0.5 probability)

replace protected identities in input sequences with the <UNK> token during training.

3.7 Multi-Task Learning

Following (Vaidya et al., 2020), we evaluated a multi-task learning framework, where we extended a base model by predicting a protected identity class from an input sequence. In our setup, the loss from an extra classifier head is weighted equal to the loss from the toxicity classifier.

3.8 Training Details

We trained our models for 100,000 iterations with a batch size of 128, the Adam optimizer (Kingma and Ba, 2014), and a learning rate of 1e-5 with betas (0.9, 0.999) on a single NVIDIA Tesla T4 GPU. Each experiment took approximately 1 hour to run. We used embeddings pre-trained on the corpora of Russian subtitles (Shavrina and Shapovalova). We experimented with 2 different architectures (self-ATTN, CNN) in several scenarios by applying Data Generation with Language Model, Identity Dropout, and Multi-Task learning, as well as combining these approaches. We used binary cross-entropy loss as the loss function for the single-task approach. As the loss function for Multi-Task learning, we used the average loss score between two tasks: predicting the toxicity score, and predicting the protected identity class. We trained our model on the training set, controlled the training process using the validation set, and evaluated metrics on the test set. We repeated each experiment 3 times and showed the mean and standard deviation values of the measurements. We applied an early stopping approach with patience level 50. The code is available on Google Drive[9].

4 Results & Conclusion

The results are provided in Table 2.

We showed that, for our dataset and for the benchmark from (Smetanin, 2020), adding an extra task of predicting the class of a protected identity can indeed improve the quality of toxicity classification in terms of reducing unintended bias. Moreover, we observed that simple techniques such as regularizing the input and extending the training data with external language models can help reduce unintended model bias on protected identities even further.

---

[8] https://github.com/vlarine/ruGPT2

[9] https://vk.cc/aANO1g

Method	Our Dataset		(Andrusyak et al., 2018)		(Smetanin, 2020)	
	GMB-AUC	F1	GMB-AUC	F1	GMB-AUC	F1
CNN	.56±.005	.66±.003	.51±.005	.59±.001	.53±.003	.78±.002
CNN + multitask	.58±.001	.68±.008	.52±.002	.61±.002	.53±.010	.80±.002
Attn	.60±.002	.71±.010	.54±.001	.72±.003	.54±.005	.80±.010
Attn + multitask	.60±.004	.74±.012	.54±.009	.69±.009	.54±.007	.82±.004
Attn + LM data	.65±.003	.74±.002	.58±.003	.70±.001	.57±.006	.83±.009
Attn + LM data + multitask	.67±.002	.74±.016	.59±.003	.70±.010	.58±.003	.84±.008
Attn + identity d/o	.61±.001	.65±.003	.53±.004	.68±.001	.54±.007	.82±.011
Attn + identity d/o + multitask	.61±.005	.66±.007	.54±.004	.69±.008	.58±.009	.83±.007
Attn + identity d/o + LM data	.67±.004	.76±.005	.55±.003	.71±.002	**.59±.003**	**.86±.012**
Attn + identity d/o + LM data + multitask	**.68±.001**	**.78±.010**	.56±.004	**.73±.003**	**.60±.008**	**.86±.004**

Table 2: Generalized Mean of Bias AUCs (GMB-AUC) and F1 scores across datasets.

For the (Andrusyak et al., 2018) benchmark, we did not see much improvement in our metrics. This can be attributed to language differences, as the benchmark contains abusive words both in Russian and Ukrainian.

We also observed that the proposed models achieved competitive results across all three datasets when evaluated with F1 score. The best performing model (Attn + identity d/o + LM data + multitask setup) achieved an F1 score of 0.86 on the (Smetanin, 2020) benchmark, which is 93% of the reported SoTA performance of a much larger model fine-tuned from a BERT-like architecture.

5 Future Work

We are interested in automatically extending our compiled list of protected identities and related words. We also expect that fine-tuning a pre-trained BERT-like model would improve our results and plan to experiment with it.

6 Acknowledgements

References

Bohdan Andrusyak, Mykhailo Rimel, and Roman Kern. 2018. Detection of abusive speech for mixed soci-
olects of russian and ukrainian languages. In *Proceedings of Recent Advances in Slavonic Natural Language Processing*, page 77–84.

Daniel Borkan, Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019a. Limitations of pinned auc for measuring unintended bias. In *arXiv preprint arXiv:1903.02088*.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019b. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 491–500.

Andrew M. Dai and Quoc V. Le. 2015. Semi-supervised sequence learning. In *Advances in neural information processing systems*, page 3079–3087.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *arXiv preprint arXiv:1703.03130*, page 3079–3087.

Tanya Lokshina. 2003. Hate speech in russia: Overview of the problem and means for counteraction. In *Bulletin: Anthropology, Minorities, Multiculturalism*, volume 4.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*.

Rodmonga Potapova and Denis Gordeev. 2016. Detecting state of aggression in sentences using cnn. In *International Conference on Speech and Computer*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. In *arXiv preprint arXiv:1508.07909*.

Tatiana Shavrina and Olga Shapovalova. To the methodology of corpus construction for machine learning: "taiga" syntax tree corpus". In *Corpora-2017*.

Sergey Smetanin. 2020. Toxic comments detection in russian. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2020"*.

Ameya Vaidya, Feng Mai, and Yue Ning. 2020. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Multilingual universal sentence encoder for semantic retrieval. In *arXiv preprint arXiv:1907.04307*.