Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2020”

Moscow, June 17–20, 2020

TOXIC COMMENTS DETECTION IN RUSSIAN

Smetanin S. I. (sismetanin@gmail.com)
National Research University Higher School of Economics, Russia

Currently, social network sites tend to be one of the major communication platforms in both offline and online space. Freedom of expression of various points of view, including toxic, aggressive, and abusive comments, might have a long-term negative impact on people’s opinions and social cohesion. As a consequence, the ability to automatically identify and moderate toxic content on the Internet to eliminate the negative consequences is one of the necessary tasks for modern society. This paper aims at the automatic detection of toxic comments in the Russian language. As a source of data, we utilized anonymously published Kaggle dataset and additionally validated its annotation quality. To build a classification model, we performed fine-tuning of two versions of Multilingual Universal Sentence Encoder, Bidirectional Encoder Representations from Transformers, and ruBERT. Fine-tuned RuBERT achieved F_1 = 92.20%, demonstrating the best classification score. We made trained models and code samples publicly available to the research community.

Keywords: toxic comments detection, language models fine-tuning, text classification, natural language processing

DOI: 10.28995/2075-7182-2020-19-1149-1159

АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ ТОКСИЧНЫХ КОММЕНТАРИЕВ НА РУССКОМ ЯЗЫКЕ

Сметанин С. И. (sismetanin@gmail.com)
Национальный исследовательский университет
«Высшая школа экономики», Россия

В настоящее время социальные сети стали один из основных инструментов для коммуникаций. Свобода выражения различных точек зрения, включая токсичные, агрессивные и оскорбительные комментарии, может оказать долгосрочное негативное влияние на как непосредственно на мнения людей, так и социальную сплоченность в целом. В данной статье рассматривается разработка подхода для автоматической классификации токсичных комментариев на русском языке. В качестве источника данных был использован набор русскоязычных комментариев Russian Language Toxic Comments, анонимно опубликованный на Kaggle. На основе предобученных моделей Multilingual Universal Sentence Encoder (M-USE), Bidirectional Encoder Representations from Transformers (BERT) и ruBERT были построены классификаторы. Наилучшие результаты в задачи бинарной классификации текстов достиг ruBERT-Toxic, продемонстрировав F_1 = 92,20 %. Предобученные модели и примеры кода для использования были опубликованы на GitHub.

Ключевые слова: определение токсичных комментариев, классификация текстов, обработка естественных языков

1. Introduction

Nowadays, social network sites have become one of the key ways to express opinions online. The rapid growth of content has led to the fact that the amount of unverified information is increasing every day. Freedom of expression of various points of view, including toxic, aggressive, and abusive comments, might have a long-term negative impact on people's opinions and social cohesion. Thus, the ability to automatically identify toxic speech and inappropriate content on the Internet to eliminate the negative consequences is one of the necessary tasks for modern society. A significant amount of studies have already been conducted by large companies [23], [26], [39], [47], however, for social acceptance of such systems limiting the right of Free Speech a good understanding and publicly available research is necessary.

A growing number of evaluation tracks such as [3], [21], [42] were organized in recent years, and the best detection approaches were evaluated. Currently, advanced deep learning techniques tend to be the superior method for this task [1], [35]. While some papers directly examined the detection of toxic language, abusive and hate speech for Russian-language [2], [8], [17], there is only one publicly available dataset of Russian-language toxic comments [5]. This dataset was published at Kaggle without any details about the annotation process, so it can be unreliable to use this dataset in academic and applied projects without deep examination.

This paper focuses on the automatic detection of toxic comments in Russian-language texts. To do this, we performed annotation validation of Russian Language Toxic Comments Dataset [5]. Next, we build classification models by exploring transfer learning of pre-trained Multilingual version of pre-trained Multilingual Universal Sentence Encoder (M-USE) [48], Bidirectional Encoder Representations from Transformers (M-BERT) [13] and ruBERT [22]. The top-performing model ruBERT-Toxic achieved F_1 = 92.20% in a binary classification task. We made the sample code and fine-tuned M-BERT and M-USE models publicly available at https://github.com/sismetanin/toxic-comments-detection-in-russian.

The rest of the article is organized as follows. In **Section 2**, we present a brief overview of the related work, including a summary of existing annotated datasets in Russian. In **Section 3**, we provide a general overview of Russian Language Toxic Comments Dataset and describes the annotation validation process. In **Section 4**, we describe the adoption of language models for the text classification task. In **Section 4**, we describe the classification experiment. In conclusion, we present the performance of our system and further ways of research.

2. Related Work

Much work on the toxic comments detection been carried out regarding different data sources. For example, Prabowo and colleagues evaluated Naive Bayes (NB), Support Vector Machine (SVM), and Random Forest Decision Tree (RFDT) algorithms for detecting hate speech and abusive language on Indonesian Twitter [34]. The experimental results demonstrated an accuracy of 68.43% for the hierarchical approach with word uni-gram features and the SVM model. In the paper [15], Founta et al. proposed a deep GRU-based neural network with pre-trained GloVe embeddings for toxic texts classification. The developed model achieved high performance across five abusive texts datasets, with the AUC value to ranged from 92% to 98%.

A growing body of workshops and competitions has dedicated to the tasks of toxic language, hate speech, and offensive language detection. For instance, HatEval and OffensEval at SemEval-2019; HASOC at FIRE-2019; Shared Task on the Identification of Offensive Language at GermEval-2019 and GermEval-2018; TRAC at COLING-2018. The models used in the task submissions varies from traditional machine learning, e.g., SVM and logistic regression, to deep learning, e.g., RNN, LSTM, GRU, CNN, CapsNet, including attention mechanism [45], [49], to state-of-the-art deep learning models such as ELMo [31] BERT [13], and USE [9], [48]. A considerable amount of the top-performing teams [18], [24], [27], [28], [30], [36], [38] exploited embeddings from the listed pre-trained language models. Since representations from pre-trained language models demonstrated high classification scores, they were widely used in further studies. For example, scholars from the University of Lorraine performed a multi-class and binary classification of tweets using two approaches: training DNN classifier with pre-trained word embeddings and fine-tuning the pre-trained BERT model [14]. They observed that BERT fine-tuning performed much better than CNN and Bidirectional LSTM neural networks build on the top of FastText embeddings.

While a significant amount of studies examined toxic and aggressive behaviour in Russian-language social media source [7], [33], [41], there are a limited amount of research papers directly exploring the automatic classification of toxicity of the texts. Gordeev utilized Convolutional Neural Networks (CNNs) and Random Forest Classifier (RFC) for detecting the state of aggression in English-language and Russian-language texts [17]. The corpus of aggression-annotated messages consisted of about 1000 annotated messages for Russian and about 1,000 for English; however, it was not made publicly available. The trained CNN model achieved an accuracy score of 66.68% in the binary classification of aggression in Russian-language texts. Based on the results, the authors considered that CNNs and deep learning approaches seem

to be more perspective and promising in the aggression detection task. Andrusyak and co-workers proposed an unsupervised probabilistic approach with a seed dictionary for classifying abusive comments from YouTube, written in Ukrainian and Russian languages [2]. The authors published a manually labelled dataset of 2,000 comments, but it contains comments in both Russian and Ukrainian languages. Consequently, it can not be directly applied for the research about Russian-language content.

Several recent studies were aimed at automatic identifying of attitudes to migrants and ethnic groups in Russian-language social media, including identification of identity-based attacks. Bodrunova and coworkers studied attitudes toward resettlers from the Post-Soviet South versus other nations by analyzing 363,000 posts from the Russian-language LiveJournal [8]. They found that migrants neither provoked the significant volume of discussion nor experience the worst treatment in Russian blogs. Furthermore, North Caucasians and Central Asians were treated quite differently. Bessudnov's research group found that traditionally Russians are more hostile to immigrants from the Caucasus and Central Asia; meanwhile, they generally accept Ukrainians and Moldovans as their potential neighbours [6]. However, according to Koltsova and coworkers, various Central Asians and Ukrainians take the lead in a negative attitude [19]. Even though some discussed academic studies aim at the detection of toxic language, abusive, and hate speech, none of them made their Russian-language datasets publicly available for the research community. To the best of our knowledge, Russian Language Toxic Comments Dataset [5] is the only publicly available dataset of Russian-language toxic comments. However, this dataset was published at Kaggle without any description of the creation and annotation process, so it can be unreliable to use this dataset in academic and applied projects without in-depth examination.

Thus, since there is a minimal amount of research focused on toxicity detection in Russian, we decided to evaluate deep learning models on the Russian Language Toxic Comments Dataset [5]. To best of our knowledge, there is no research dedicated to toxic comments classification based on this source of data. We identified Multilingual BERT and Multilingual USE as one of the most common and successful language models in recent text classification papers. Moreover, only these language models officially support the Russian language. We decided to utilize fine-tuning as a transfer learning approach since recent fine-tuning studies reported the best classification results [13], [22], [43], [48].

3. Toxic Comments Dataset

Kaggle Russian Language Toxic Comments Dataset[1] [5] is the collection of annotated comments from 2ch[2] and Pikabu[3], which was published on Kaggle in 2019. It consists of 14,412 comments, where 4,826 texts were labeled as toxic, and 9,586 as non-toxic. The average length of comments is 175 characters; the minimum length is 21, and the maximum is 7,403.

---

1 https://www.kaggle.com/blackmoon/russian-language-toxic-comments

2 https://2ch.hk/

3 https://pikabu.ru/

To validate the annotation quality of the dataset, we decided to manually annotate a subset of comments and compare original labels and our labels using inter-annotator agreement metrics. We decided to assume the dataset annotation as valid in case of a substantial or high level of the inter-annotator agreement will be achieved. To begin with, we performed human annotation of a part of this dataset (3000 comments) and then compared our class labels with the original ones. The annotation was performed by Russian-language speakers on the crowdsourcing platform Yandex. Toloka, which had already been used in several academic studies about Russian-language texts [10], [29], [32], [44]. As an annotation guideline, we used the annotation instructions for toxicity with sub-attributes from Jigsaw Toxic Comment Classification Challenge. According to guidelines, annotators were requested to detect the toxicity of texts in a collection of online comments. For each comment provided, annotators were required to select the level of toxicity in the comment. In order to get more accurate responses from annotators and restrict access to tasks for cheating annotators, we utilized the following techniques: assigning a skill to the annotators based on their responses to control tasks and baning the performers who give incorrect responses; restricting the pool access for annotators who respond too quickly; restricting access to tasks for annotators who fail to enter captcha several times in a row. Each post was annotated by 3 to 8 annotators using a dynamic overlap technique[4]. Next, the results were aggregated using the Dawid-Skene method [12] based on the Yandex.Toloka's recommendations[5]. The annotators demonstrated a high inter-annotator agreement as according to the Krippendorff's alpha value of 0.81. Finally, Cohen's kappa coefficient between original and our aggregated labels constituted 0.68, which is a substantial level of the inter-annotator agreement, according to Cohen [11]. As a result, we assumed the dataset annotation as valid, especially taking into account potential differences in the annotation instructions.

4. Machine Learning Models

4.1. Baselines

As baseline approaches, we selected one basic machine learning-based approach and one modern neural network-based approach. In both cases, we applied the following preprocessing techniques: replacing URLs and usernames with keywords, removing punctuation marks, and converting strings into lowercase. The first one was Multinomial Naive Bayes (MNB), which tended to perform well in the text classification task [16, 40]. To build the MNB model, we used the Bag-of-Words model and the TF-IDF vectorization. The second one was Bidirectional Long Short-Term Memory (BiLSTM) neural network, which demonstrated high classification scores in recent sentiment analysis studies. For the embeddings layer of the neural network, we pre-trained Word2Vec embeddings ($dim = 300$) [25] on the collection of Russian

[4] https://yandex.ru/support/toloka-requester/concepts/dynamic-overlap.html

[5] https://yandex.ru/support/toloka-requester/concepts/categorization.html?lang=en

language tweets from RuTweetCorp [37]. On the top of the Word2Vec embeddings, we added two stacked Bidirectional LSTM layers. Next, we added a hidden fully connected layer and sigmoid output layer. To reduce overfitting, regularization layers with Gaussian noise and Dropout layers were also added to the neural network. We used Adam optimizer with the initial learning rate of 0.001 and categorical binary cross-entropy as a loss function. We trained our network with frozen embeddings for the 10 epochs. We tried to unfrozen embeddings on the different epoch with the simultaneous reduction of learning rate but failed to get better results. It was probably connected with the size of the training dataset [4].

4.2. Bidirectional Encoder Representations from Transformers

Currently, two multilingual versions of $\text{BERT}_{\text{BASE}}$ are officially available, but only Cased version is officially recommended[6]. $\text{BERT}_{\text{BASE}}$ takes a sequence of no more than 512 tokens and outputs the representation of this sequence. The tokenization is performed by WordPiece tokenizer [46] with the preliminary text normalization and punctuation splitting. Based on the $\text{BERT}_{\text{BASE}}$ Cased, researchers from the Moscow Institute of Physics and Technology pre-trained and published ruBERT model for the Russian language [22]. We used pre-trained Multilingual $\text{BERT}_{\text{BASE}}$ Cased and ruBERT, which support 104 languages, including Russian, with 12 stacked Transformer blocks, a hidden size of 768, 12 self-attention heads, and 110M parameters in general. The fine-tuning stage was performed with the recommended parameters from the paper [43] and the official repository[7]: a number of train epochs of 3, a number of warm-up steps of 10%, a max sequence length of 128, a batch size of 32, and a learning rate of 5e-5.

4.3. Multilingual Universal Sentence Encoder

As input data, Multilingual $\text{USE}_{\text{Trans}}$ takes a sequence of no more than 100 tokens, while Multilingual USE_{CNN} takes a sequence of no more than 256 tokens. SentencePiece tokenization [20] is used for all supported languages. We used pre-trained Multilingual $\text{USE}_{\text{Trans}}$ which support 16 languages including Russian, with Transformer encoder with 6 transformer layers, 8 attention heads, filter size of 2048, hidden size of 512, and 16 parameters in general. We also used pre-trained Multilingual USE_{CNN} which support N languages including Russian, with CNN encoder with 2 CNN layers, filter width of (1, 2, 3, 5), filter size of 256, and N parameters in general. For both models, we used recommended parameters from and the TensorFlow Hub page[8]: a number of train epochs of 100, a batch size of 32, and a learning rate of 3e-4.

---

6 https://github.com/google-research/bert/blob/master/multilingual.md

7 https://github.com/google-research/bert

8 https://www.tensorflow.org/hub/tutorials/text_classification_with_tf_hub

5. Experiment

We evaluated the following baseline and transfer learning approaches: Multinomial Naive Bayes classifier, Bidirectional Long Short-Term Memory (BiLSTM) neural network, Multilingual version of Bidirectional Encoder Representations from Transformers (M-BERT), ruBERT, two versions of Multilingual Universal Sentence Encoder (M-USE). The classification performance of trained models on the test subset (20%) can be found in Table 2. All fine-tuned language models exceed baseline approaches in terms of precision, recall, and F_1-measure. According to the results, ruBERT achieved $F_1 = 92.20\%$, demonstrating the best classification score.

Table 1: Binary classification of toxic comments in Russian

System	P	R	F_1
MNB	87.01%	81.22%	83.21%
BiLSTM	86.56%	86.65%	86.59%
$M - BERT_{BASE} - Toxic$	91.19%	91.10%	91.15%
ruBert − Toxic	**91.91%**	**92.51%**	**92.20%**
$M - USE_{CNN} - Toxic$	89.69%	90.14%	89.91%
$M - USE_{Trans} - Toxic$	90.85%	91.92%	91.35%

6. Conclusion

In this paper, we fine-tuned two versions of Multilingual Universal Sentence Encoder [48], Multilingual Bidirectional Encoder Representations from Transformers [13] and RuBERT [22] for toxic comments detection in Russian. Fine-tuned RuBERT-Toxic achieved $F_1 = 92.20\%$, demonstrating the best classification score. The contributions of this study to practice and research are threefold. Firstly, we outlined an existing knowledge base regarding toxic comments detection in Russian content. In doing so, we identified the only existing toxic comments dataset in Russian, which is publicly available. Secondly, we performed the validation of annotation quality for this dataset, since it was anonymously published at Kaggle. Lastly, to provide further studies with strong classification baselines, we made pre-trained Multilingual BERT-based, ruBERT-based and Multilingual USE-based models publicly available to the research community: https://github.com/sismetanin/toxic-comments-detection-in-russian.

References

1. *Aken, B. van et al.:* Challenges for toxic comment classification: An in-depth error analysis. In: Proceedings of the 2nd workshop on abusive language online (ALW2). pp. 33–42. Association for Computational Linguistics, Brussels, Belgium (2018).
2. *Andrusyak, B. et al.:* Detection of abusive speech for mixed sociolects of russian and ukrainian languages. In: The 12th workshop on recent advances in slavonic natural languages processing, RASLAN 2018, karlova studanka, czech republic, december 7–9, 2018. pp. 77–84 (2018).

3. *Basile, V. et al.:* SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In: Proceedings of the 13th international workshop on semantic evaluation. pp. 54–63. Association for Computational Linguistics, Minneapolis, Minnesota, USA (2019).
4. *Baziotis, C. et al.:* DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In: Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017). pp. 747–754. Association for Computational Linguistics, Vancouver, Canada (2017).
5. *Belchikov, A.:* Russian language toxic comments, https://www.kaggle.com/blackmoon/russian-language-toxic-comments.
6. *Bessudnov, A., Shcherbak, A.:* Ethnic discrimination in multi-ethnic societies: Evidence from russia. European Sociological Review. (2019).
7. *Biryukova, E. V. et al.:* READER'S comment in on-line magazine as a genre of internet discourse (by the material of the german and russian languages). Philological Sciences. Issues of Theory and Practice. 12, 1, 79–82 (2018).
8. *Bodrunova, S. S. et al.:* Who's bad? Attitudes toward resettlers from the post-soviet south versus other nations in the russian blogosphere. International Journal of Communication. 11, 23 (2017).
9. *Cer, D. M. et al.:* Universal sentence encoder. ArXiv. abs/1803.11175, (2018).
10. *Chernyak, E. et al.:* Char-rnn for word stress detection in east slavic languages. CoRR. abs/1906.04082, (2019).
11. *Cohen, J.:* A coefficient of agreement for nominal scales. Educational and psychological measurement. 20, 1, 37–46 (1960).
12. *Dawid, A. P., Skene, A. M.:* Maximum likelihood estimation of observer error-rates using the em algorithm. Journal of the Royal Statistical Society: Series C (Applied Statistics). 28, 1, 20–28 (1979).
13. *Devlin, J. et al.:* BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019).
14. *d'Sa, A. G. et al.:* BERT and fastText embeddings for automatic detection of toxic speech. In: SIIE 2020-information systems and economic intelligence. (2020).
15. *Founta, A. M. et al.:* A unified deep learning architecture for abuse detection. In: Proceedings of the 10th acm conference on web science. pp. 105–114. Association for Computing Machinery, New York, NY, USA (2019).
16. *Frank, E., Bouckaert, R.:* Naive bayes for text classification with unbalanced classes. In: Fürnkranz, J. et al. (eds.) Knowledge discovery in databases: PKDD 2006. pp. 503–510. Springer Berlin Heidelberg, Berlin, Heidelberg (2006).
17. *Gordeev, D.:* Detecting state of aggression in sentences using cnn. In: International conference on speech and computer. pp. 240–245. Springer (2016).
18. *Indurthi, V. et al.:* FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in twitter. In: Proceedings of the 13th international workshop on semantic evaluation. pp. 70–74. Association for Computational Linguistics, Minneapolis, Minnesota, USA (2019).

19. *Koltsova, O. et al.:* FINDING and analyzing judgements on ethnicity in the russian-language social media. AoIR Selected Papers of Internet Research. (2017).
20. *Kudo, T., Richardson, J.:* SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations. pp. 66–71. Association for Computational Linguistics, Brussels, Belgium (2018).
21. *Kumar, R. et al. eds:* Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018). Association for Computational Linguistics, Santa Fe, New Mexico, USA (2018).
22. *Kuratov, Y., Arkhipov, M.:* Adaptation of deep bidirectional multilingual transformers for Russian language. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". pp. 333–340. RSUH, Moscow, Russia (2019).
23. *Lenhart, A. et al.:* Online harassment, digital abuse, and cyberstalking in america. Data; Society Research Institute (2016).
24. *Liu, P. et al.:* NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In: Proceedings of the 13th international workshop on semantic evaluation. pp. 87–91. Association for Computational Linguistics, Minneapolis, Minnesota, USA (2019).
25. *Mikolov, T. et al.:* Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th international conference on neural information processing systems—volume 2. pp. 3111–3119. Curran Associates Inc., Red Hook, NY, USA (2013).
26. *Mishra, P. et al.:* Abusive language detection with graph convolutional networks. In: Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers). pp. 2145–2150 (2019).
27. *Mishra, S., Mishra, S.:* 3Idiots at HASOC 2019: Fine-tuning transformer neural networks for hate speech identification in indo-european languages. In: Working notes of FIRE 2019—forum for information retrieval evaluation, kolkata, india, december 12–15, 2019. pp. 208–213 (2019).
28. *Nikolov, A., Radivchev, V.:* Nikolov-radivchev at SemEval-2019 task 6: Offensive tweet classification with BERT and ensembles. In: Proceedings of the 13th international workshop on semantic evaluation. pp. 691–695. Association for Computational Linguistics, Minneapolis, Minnesota, USA (2019).
29. *Panchenko, A. et al.:* RUSSE'2018: A Shared Task on Word Sense Induction for the Russian Language. In: Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". pp. 547–564. RSUH, Moscow, Russia (2018).
30. *Paraschiv, A., Cercel, D.-C.:* UPB at germeval-2019 task 2: BERT-based offensive language classification of german tweets. In: Preliminary proceedings of the 15th conference on natural language processing (konvens 2019). Erlangen, germany: German society for computational linguistics & language technology. pp. 396–402 (2019).

31. *Peters, M. et al.:* Deep contextualized word representations. In: Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers). pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (2018).
32. *Ponomareva, M. et al.:* Automated word stress detection in Russian. In: Proceedings of the first workshop on subword and character level models in NLP. pp. 31–35. Association for Computational Linguistics, Copenhagen, Denmark (2017).
33. *Potapova, R., Komalova, L.:* Lexico-semantical indices of "deprivation–aggression" modality correlation in social network discourse. In: International conference on speech and computer. pp. 493–502. Springer (2017).
34. *Prabowo, F. A. et al.:* Hierarchical multi-label classification to identify hate speech and abusive language on indonesian twitter. In: 2019 6th international conference on information technology, computer and electrical engineering (icitacee). pp. 1–5 (2019).
35. *Risch, J., Krestel, R.:* Toxic comment detection in online discussions. In: Deep learning-based approaches for sentiment analysis. pp. 85–109. Springer (2020).
36. *Risch, J. et al.:* HpiDEDIS at germeval 2019: Offensive language identification using a german bert model. In: Preliminary proceedings of the 15th conference on natural language processing (konvens 2019). Erlangen, germany: German society for computational linguistics & language technology. pp. 403–408 (2019).
37. *Rubtsova, Y.:* A method for development and analysis of short text corpus for the review classification task. Proceedings of conferences Digital Libraries: Advanced Methods and Technologies, Digital Collections (RCDL'2013). Pp. 269–275 (2013).
38. *Ruiter, D. et al.:* LSV-uds at HASOC 2019: The problem of defining hate. In: Working notes of FIRE 2019—forum for information retrieval evaluation, kolkata, india, december 12–15, 2019. pp. 263–270 (2019).
39. *Sambasivan, N. et al.:* "They don't leave us alone anywhere we go": Gender and digital abuse in south asia. In: Proceedings of the 2019 chi conference on human factors in computing systems. Association for Computing Machinery, New York, NY, USA (2019).
40. *Sang-Bum Kim et al.:* Some effective techniques for naive bayes text classification. IEEE Transactions on Knowledge and Data Engineering. 18, 11, 1457–1466 (2006).
41. *Shkapenko, T., Vertelova, I.:* Hate speech markers in internet comments to translated articles from polish media. Political Linguistics. 70, 4, Pages 104–111 (2018).
42. *Strus, J. M. et al.:* Overview of germeval task 2, 2019 shared task on the identification of offensive language. Presented at the (2019).
43. *Sun, C. et al.:* How to fine-tune bert for text classification? In: Sun, M. et al. (eds.) Chinese computational linguistics. pp. 194–206. Springer International Publishing, Cham (2019).
44. *Ustalov, D., Igushkin, S.:* Sense inventory alignment using lexical substitutions and crowdsourcing. In: 2016 international fruct conference on intelligence, social media and web (ismw fruct). (2016).

45. *Vaswani, A. et al.:* Attention is all you need. In: Proceedings of the 31st international conference on neural information processing systems. pp. 6000–6010. Curran Associates Inc., Red Hook, NY, USA (2017).
46. *Wu, Y. et al.*: Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144. (2016).
47. *Yang, F. et al.:* Exploring deep multimodal fusion of text and photo for hate speech classification. In: Proceedings of the third workshop on abusive language online. pp. 11–18. Association for Computational Linguistics, Florence, Italy (2019).
48. *Yang, Y. et al.:* Multilingual universal sentence encoder for semantic retrieval. CoRR. abs/1907.04307, (2019).
49. *Yang, Z. et al.:* Hierarchical attention networks for document classification. In: Proceedings of the 2016 conference of the north American chapter of the association for computational linguistics: Human language technologies. pp. 1480–1489. Association for Computational Linguistics, San Diego, California (2016).