Improving Cross-Domain Hate Speech Detection by Reducing the False Positive Rate

Ilia Markov
CLIPS Research Center
University of Antwerp, Belgium
ilia.markov@uantwerpen.be

Walter Daelemans
CLIPS Research Center
University of Antwerp, Belgium
walter.daelemans@uantwerpen.be

Abstract

Hate speech detection is an actively growing field of research with a variety of recently proposed approaches that allowed to push the state-of-the-art results. One of the challenges of such automated approaches – namely recent deep learning models – is a risk of false positives (i.e., false accusations), which may lead to over-blocking or removal of harmless social media content in applications with little moderator intervention. We evaluate deep learning models both under in-domain and cross-domain hate speech detection conditions, and introduce an SVM approach that allows to significantly improve the state-of-the-art results when combined with the deep learning models through a simple majority-voting ensemble. The improvement is mainly due to a reduction of the false positive rate.

1 Introduction

A commonly used definition of hate speech is a communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics (Nockleby, 2000). The automated detection of hate speech online and related concepts, such as toxicity, cyberbullying, abusive and offensive language, has recently gained popularity within the Natural Language Processing (NLP) community. Robust hate speech detection systems may provide valuable information for police, security agencies, and social media platforms to effectively counter such effects in online discussions (Halevy et al., 2020).

Despite the recent advances in the field, mainly due to a large amount of available social media data and recent deep learning techniques, the task remains challenging from an NLP perspective, since on the one hand, hate speech, toxicity, or offensive language are often not explicitly expressed through the use of offensive words, while on the other hand, non-hateful content may contain such terms and the classifier may consider signals for an offensive word stronger than other signals from the context, leading to false positive predictions, and further removal of harmless content online (van Aken et al., 2018; Zhang and Luo, 2018).

Labelling non-hateful utterances as hate speech (false positives or type II errors) is a common error even for human annotators due to personal bias. Several studies showed that providing context, detailed annotation guidelines, or the background of the author of a message improves annotation quality by reducing the number of utterances erroneously annotated as hateful (de Gibert et al., 2018; Sap et al., 2019; Vidgen and Derczynski, 2020).

We assess the performance of deep learning models that currently provide state-of-the-art results for the hate speech detection task (Zampieri et al., 2019b, 2020) both under in-domain and cross-domain hate speech detection conditions, and introduce an SVM approach with a variety of engineered features (e.g., stylometric, emotion, hate speech lexicon features, described further in the paper) that significantly improves the results when combined with the deep learning models in an ensemble, mainly by reducing the false positive rate.

We target the use cases where messages are flagged automatically and can be mistakenly removed, without or with little moderator intervention. While existing optimization strategies (e.g., threshold variation) allow to minimize false positives with a negative effect on overall accuracy, our method reduces the false positive rate without decreasing overall performance.

2 Methodology

Hate speech detection is commonly framed as a binary supervised classification task (hate speech vs. non-hate speech) and has been addressed using both deep neural networks and methods based on manual feature engineering (Zampieri et al., 2019b,

Proceedings of the 4th NLP4IF Workshop on NLP for Internet Freedom, pages 17–22

2020). Our work evaluates and exploits the advantages of deep neural networks as means for extracting discriminative features directly from text and of a conventional SVM approach taking the advantage of explicit feature engineering based on task and domain knowledge. In more detail, we focus on the approaches described below.

2.1 Baselines

Bag of words (BoW) We use a tf-weighted lowercased bag-of-words (BoW) approach with the liblinear Support Vector Machines (SVM) classifier. The optimal SVM parameters (penalty parameter (C), loss function (loss), and tolerance for stopping criteria (tol)) were selected based on grid search.

Convolutional neural networks (CNN) We use a convolutional neural networks (CNN) approach (Kim, 2014) to learn discriminative word-level hate speech features with the following architecture: to process the word embeddings (trained with fastText (Joulin et al., 2017)), we use a convolutional layer followed by a global average pooling layer and a dropout of 0.6. Then, a dense layer with a ReLU activation is applied, followed by a dropout of 0.6, and finally, a dense layer with a sigmoid activation to make the prediction for the binary classification.

Long short-term memory networks (LSTM) We use an LSTM model (Hochreiter and Schmidhuber, 1997), which takes a sequence of words as input and aims at capturing long-term dependencies. We process the sequence of word embeddings (trained with GloVe (Pennington et al., 2014)) with a unidirectional LSTM layer with 300 units, followed by a dropout of 0.2, and a dense layer with a sigmoid activation for predictions.

2.2 Models

BERT and RoBERTa Pretrained language models, i.e., Bidirectional Encoder Representations from Transformers, BERT (Devlin et al., 2019) and Robustly Optimized BERT Pretraining Approach, RoBERTa (Liu et al., 2019b), currently provide the best results for hate speech detection, as shown by several shared tasks in the field (Zampieri et al., 2019b; Mandl et al., 2019; Zampieri et al., 2020). We use the BERT-base-cased (12-layer, 768-hidden, 12-heads, 110 million parameters) and RoBERTa-base (12-layer, 768-hidden, 12-heads, 125 million parameters) models from the huggingface library[1] fine-tuning the models on the training data. The implementation was done in PyTorch (Paszke et al., 2019) using the simple transformers library[2].

Support Vector Machines (SVM) The Support Vector Machines (SVM) algorithm (Cortes and Vapnik, 1995) is commonly used for the hate speech detection task (Davidson et al., 2017; Salminen et al., 2018; MacAvaney et al., 2019; Del Vigna et al., 2017; Ljubešić et al., 2020).

Following Markov et al. (2021), we lemmatize the messages in our data and represent them through universal part-of-speech (POS) tags (obtained with the Stanford POS Tagger (Toutanova et al., 2003)), function words (words belonging to the closed syntactic classes)[3], and emotion-conveying words (from the NRC word-emotion association lexicon (Mohammad and Turney, 2013)) to capture stylometric and emotion-based peculiarities of hateful content. For example, the phrase *@USER all conservatives are bad people* [OLID id: 22902] is represented through POS, function words, and emotion-conveying words as 'PROPN', 'all', 'NOUN', 'be', 'bad', 'NOUN'. From this representation n-grams (with n = 1–3) are built.

We use the NRC lexicon emotion associations (e.g., *bad* = 'anger', 'disgust', 'fear', 'negative', 'sadness') and hate speech lexicon entries (De Smedt et al., 2020) as additional feature vectors, word unigrams, and character n-grams for the in-domain setting (with n = 1–6), considering only those n-grams that appear in ten training messages (min_df = 10).

We use tf-idf weighting scheme and the liblinear scikit-learn (Pedregosa et al., 2011) implementation of the SVM algorithm with optimized parameters (penalty parameter (C), loss function (loss), and tolerance for stopping criteria (tol)) selected based on grid search.

Ensemble We use a simple ensembling strategy, which consists in combining the predictions produced by the deep learning and machine learning approaches: BERT, RoBERTa, and SVM, through a hard majority-voting ensemble, i.e., selecting the label that is most often predicted.

[1] https://huggingface.co/
[2] https://simpletransformers.ai/
[3] https://universaldependencies.org/u/pos/

3 Experiments and Results

3.1 Data

To evaluate the approaches discussed in Section 2 we conducted experiments on two recent English social media datasets for hate speech detection:

FRENK (Ljubešić et al., 2019) The FRENK datasets consist of Facebook comments in English and Slovene covering LGBT and migrant topics. The datasets were manually annotated for fine-grained types of socially unacceptable discourse (e.g., violence, offensiveness, threat). We focus on the English dataset and use the coarse-grained (binary) hate speech classes: hate speech vs. non-hate speech. We select the messages for which more than four out of eight annotators agreed upon the class and use training and test partitions splitting the dataset by post boundaries in order to avoid comments from the same discussion thread to appear in both training and test sets, that is, to avoid within-post bias.

OLID (Zampieri et al., 2019a) The OLID dataset has been introduced in the context of the SemEval 2019 shared task on offensive language identification (Zampieri et al., 2019b). The dataset is a collection of English tweets annotated for the type and target of offensive language. We focus on whether a message is offensive or not and use the same training and test partitions as in the OffensEval 2019 shared task (Zampieri et al., 2019b).

The statistics of the datasets used are shown in Table 1. For cross-domain experiments, we train (merging the training and test subsets) on FRENK and test on OLID, and vice versa.

		FRENK		OLID	
		# messages	%	# messages	%
Train	HS	2,848	35.9	4,400	33.2
	non-HS	5,091	64.1	8,840	66.8
Test	HS	744	35.5	240	27.9
	non-HS	1,351	64.5	620	72.1
Total		10,034		14,100	

Table 1: Statistics of the datasets used.

3.2 Results

The performance of the models described in Section 2 in terms of precision, recall, and F1-score (macro-averaged) in the in-domain and cross-domain settings is shown in Table 2. Statistically significant gains of the ensemble approach (BERT, RoBERTa, and SVM) over the best-performing individual model for each of the settings according to McNemar's statistical significance test (McNemar, 1947) with $\alpha < 0.05$ are marked with '*'.

We can observe that the in-domain trends are similar across the two datasets: BERT and RoBERTa achieve the highest results, outperforming the baseline methods and the SVM approach. The results on the OLID test set are in line with the previous research on this data (Zampieri et al., 2019a) and are similar to the best-performing shared task systems when the same types of models are used (i.e., 80.0% F1-score with CNN, 75.0% with LSTM, and 82.9% with BERT (Zampieri et al., 2019b)), while the results on the FRENK test set are higher than the results reported in (Markov et al., 2021) for all the reported models.[4] We can also note that the SVM approach achieves competitive results compared to the deep learning models. A near state-of-the-art SVM performance (compared to BERT) was also observed in other studies on hate speech detection, e.g., (MacAvaney et al., 2019), where tf-idf weighted word and character n-gram features were used. The results for SVM on the OLID test set are higher than the results obtained by the machine learning approaches in the OffensEval 2019 shared task (i.e., 69.0% F1-score (Zampieri et al., 2019b)). Combining the SVM predictions with the predictions produced by BERT and RoBERTa through the majority-voting ensemble further improves the results on the both datasets. We also note that the F1-score obtained by the ensemble approach on the OLID test set is higher than the result of the winning approach of the OffensEval 2019 shared task (Liu et al., 2019a): 83.2% and 82.9% F1-score, respectively.

The cross-domain results indicate that using out-of-domain data for testing leads to a substantial drop in performance by around 5–10 F1 points for all the evaluated models. BERT and RoBERTa remain the best-performing individual models in the cross-domain setting, while the SVM approach shows a smaller drop than the baseline CNN and LSTM models, outperforming these models in the cross-domain setup, and contributes to the ensemble approach.

Both in the in-domain and cross-domain settings, combining the predictions produced by BERT and RoBERTa with SVM through the majority-voting

[4] Markov et al. (2021) used multilingual BERT and did not used pretrained embedding for CNN and LSTM to address multiple language covered in the paper.

In-domain						
	FRENK			OLID		
Model	Precision	Recall	F1	Precision	Recall	F1
BoW	71.0	70.8	70.9	75.9	70.9	72.5
CNN	76.8	76.6	76.7	81.8	77.8	79.4
LSTM	73.3	72.5	72.8	78.2	75.1	76.4
BERT	78.3	78.4	78.3	82.3	82.0	82.2
RoBERTa	78.4	78.7	78.5	80.2	79.7	80.0
SVM	77.8	76.4	77.0	82.3	76.1	78.3
Ensemble	**80.0**	**79.5**	**79.7***	**84.7**	**82.0**	**83.2**

Cross-domain						
	OLID – FRENK			FRENK – OLID		
Model	Precision	Recall	F1	Precision	Recall	F1
BoW	70.3	64.9	65.5	66.3	63.1	63.8
CNN	70.8	65.6	66.3	65.9	67.6	66.0
LSTM	68.0	66.1	66.6	67.5	65.9	66.5
BERT	70.5	68.8	69.4	71.7	72.7	72.1
RoBERTa	**73.9**	68.2	69.2	71.9	73.6	72.4
SVM	70.2	67.0	67.7	70.2	68.4	69.0
Ensemble	73.1	**68.8**	**69.7***	**73.5**	**73.9**	**73.6***

Table 2: In-domain and cross-domain results for the baselines, individual models and the ensemble.

	In-domain				Cross-domain			
	FRENK		OLID		OLID – FRENK		FRENK – OLID	
Model	FPR	PPV	FPR	PPV	FPR	PPV	FPR	PPV
CNN	15.8	70.6	7.3	77.0	11.0	68.2	31.2	51.0
LSTM	17.0	66.7	9.4	71.1	17.2	61.5	17.3	58.2
BERT	15.6	71.8	9.7	74.7	16.8	64.3	21.1	60.7
RoBERTa	16.0	71.7	10.6	71.8	**9.5**	**72.8**	23.7	59.5
SVM	**13.2**	73.3	**5.8**	79.4	14.0	65.6	**15.7**	62.1
Ensemble	13.3	**74.9**	6.8	**80.2**	11.4	70.5	18.3	**63.9**

Table 3: False positive rate (FPR) and positive predictive value (PPV) for the examined models.

ensemble approach improves the results over the individual models incorporated into the ensemble.[5] This improvement is significant in all cases, except for the OLID in-domain setting, where only 860 messages are used for testing. A more detailed analysis presented below provides deeper insights into the nature of these improvements.

4 Error Analysis

We performed a quantitative analysis of the obtained results focusing on the false positive rate: $FPR = FP/(FP + TN)$, the probability that a positive label is assigned to a negative instance; we additionally report positive predictive value: $PPV = TP/(TP + FP)$, the probability a predicted positive is a true positive, for the examined models in the in-domain and cross-domain settings (Table 3).

We note that the SVM approach shows the lowest FPR and the highest PPV in all the considered settings, except when training on the OLID dataset and testing on the FRENK dataset. Combining BERT and RoBERTa with SVM through the ensemble approach reduces the false positive rate in three out of four settings, when compared to BERT and RoBERTa in isolation, and contributes to the overall improvement of the results in all the considered settings. The improvement brought by combining BERT and RoBERTa with SVM is higher in the majority of cases than combining BERT and RoBERTa with either CNN or LSTM. Measuring the correlation of the predictions of different models using the Pearson correlation coefficient revealed that SVM produces highly uncorrelated predictions when compared to BERT and RoBERTa. An analogous effect for deep learning and shallow approaches was observed in (van Aken et al., 2018).

The majority of the erroneous false positive predictions produced by the SVM approach contain

[5] We also examined other ensemble approaches, e.g., Gradient Boosting, AdaBoost, soft majority voting, achieving similar results and trends under the cross-domain conditions.

offensive words used in a non-hateful context (avg. 78.8% messages over the four settings), while for BERT and RoBERTa this percentage is lower in all the settings (avg. 68.7% and 69.7%, respectively), indicating that BERT and RoBERTa tend to classify an instance as belonging to the hate speech class even if it is not explicitly contains offensive terms.

Our findings suggest that the SVM approach improves the results mainly by reducing the false positive rate when combined with BERT and RoBERTa. This strategy can be used to address one of the challenges that social media platforms are facing: removal of content that does not violate community guidelines.

5 Conclusions

We showed that one of the challenges in hate speech detection: erroneous false positive decisions, can be addressed by combining deep learning models with a robust feature-engineered SVM approach. The results are consistent within the in-domain and cross-domain settings. This simple strategy provides a significant boost to the state-of-the-art hate speech detection results.

Acknowledgements

This research has been supported by the Flemish Research Foundation through the bilateral research project FWO G070619N "The linguistic landscape of hate speech on social media". The research also received funding from the Flemish Government (AI Research Program).

References

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*, pages 512–515, Montreal, QC, Canada. AAAI Press.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online*, pages 11–20, Brussels, Belgium. ACL.

Tom De Smedt, Pierre Voué, Sylvia Jaki, Melina Röttcher, and Guy De Pauw. 2020. Profanity & offensive words (POW): Multilingual fine-grained lexicons for hate speech. Technical report, TextGain.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on Facebook. In *Proceedings of the First Italian Conference on Cybersecurity*, pages 86–95, Venice, Italy. CEUR-WS.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, MN, USA. ACL.

Alon Halevy, Cristian Canton Ferrer, Hao Ma, Umut Ozertem, Patrick Pantel, Marzieh Saeidi, Fabrizio Silvestri, and Ves Stoyanov. 2020. Preserving integrity in online social networks. *CoRR*, abs/2009.10311.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. ACL.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751, Doha, Qatar. ACL.

Ping Liu, Wen Li, and Liang Zou. 2019a. 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*, pages 87–91, Minneapolis, Minnesota, USA. ACL.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2019. The FRENK datasets of socially unacceptable discourse in Slovene and English. In *Proceedings of the 22nd International Conference on Text, Speech, and Dialogue*, pages 103–114, Ljubljana, Slovenia. Springer.

Nikola Ljubešić, Ilia Markov, Darja Fišer, and Walter Daelemans. 2020. The LiLaH emotion lexicon of Croatian, Dutch and Slovene. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 153–157, Barcelona, Spain (Online). ACL.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PLOS ONE*, 14(8):1–16.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, pages 14–17, New York, NY, USA. ACM.

Ilia Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. 2021. Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–159, Kyiv, Ukraine (Online). ACL.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Saif Mohammad and Peter Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29:436–465.

John Nockleby. 2000. Hate speech. *Encyclopedia of the American Constitution*, pages 1277–1279.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8026–8037. Curran Associates, Inc.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. CL.

Joni Salminen, Hind Almerekhi, Milica Milenković, Soon Gyo Jung, Jisun An, Haewoon Kwak, and Bernard J. Jansen. 2018. Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*, pages 330–339, Palo Alto, California, USA. AAAI press.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. ACL.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 252–259, Edmonton, Canada. ACL.

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. *CoRR*, abs/1809.07572.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data: Garbage in, garbage out. *CoRR*, abs/2004.01670.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1415–1420. ACL.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. ACL.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *CoRR*, abs/2006.07235.

Ziqi Zhang and Lei Luo. 2018. Hate speech detection: A solved problem? The challenging case of long tail on Twitter. *CoRR*, abs/1803.03662.