LSV-UdS at HASOC 2019: The Problem of Defining Hate*

Dana Ruiter[1], Md. Ataur Rahman[1], and Dietrich Klakow[1]

Spoken Language Systems Group, Saarland University, Germany
{druiter, arahman, dietrich.klakow}@lsv.uni-saarland.de

Abstract. We describe our English, German and Hindi SVM and BERT-based hate speech classifiers, which includes the top-performing model for the German sub-task B. A special focus is laid on the exploration of various external corpora, the lack of mutual compatibility and the conclusions that arise from this.

Keywords: Hate Speech Detection

1 Introduction

In the participatory web, there is an ongoing influx of user contents. Depending on the policies of a web page, the netiquette allows –and disallows– a set of online behaviors towards others. This situation is further enforced by current governmental initiatives against online abuse and hate speech demanding direct action by the operators of an online service in case of law-infringing user contents.[1] However, as the amount –and psychological weight– of the data is overwhelming for human moderators, there is a growing interest in automating the identification of abusive comments online.

As a consequence, the recent years have seen a growing emergence of corpora attempting to capture hate in various facets with the aim of providing training data for text classifiers. While some corpora focus on differentiating between different targets of hate (i.e. sexism vs. racism) [17] [13], others focus on varying degrees of hate, ranging from binary distinctions such as *hate* vs. *offensive* or *abusive* speech [3], [6], over distinctions focusing on explicitness [9], all the way to multi-label corpora covering different manifestations of hate [1] (i.e. *identity hate, insult, threat* etc.). As the majority of publicly available corpora of online hate are small in size, there is an interest in merging different sources for training. However, as all of these corpora have distinct foci, the mutual compatibility between corpora is not always given. In our submission to the HASOC

[*] The project on which this paper is based was funded by the DFG under the funding code WI 4204/3-1. Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). FIRE 2019, 12-15 December 2019, Kolkata, India.

[1] https://www.theguardian.com/world/2019/jul/09/france-online-hate-speech-law-social-media

2019 shared task [12], we focus on exploring different combinations of prominent hate speech corpora for both statistical models (SVM) and neural approaches (sequence-to-label) applied to sub-tasks A and B. Notably, our simple neural approach yielded us top results for the highly low-resourced German sub-task B.

In the following sections, we will give a brief introduction to related work in hate speech classification (2), followed by a description of the data (3) and the models (4, 5). At the end, we present our results (6) as well as future work (7).

2 Related Work

In the last years, a variety of standard text classification procedures have been applied to the task of hate speech detection. These range from statistical methods such as naive-bayes [15], logistic regression [17] [19] [3] and support vector machines (SVM) [15] [14], to neural approaches such as sequence-to-label [8] or hybrid convolutional neural networks [14].

Due to the comparatively large amount of neography in user comments, subword features such as character n-grams [17] or comment embeddings [5] greatly improve classification results. Our neural approach goes in this direction, as its input is subword units, which allows it to have a high vocabulary coverage despite the noisy orthography of many comments.

While most features used for training hate speech classifiers focus on textual data, there is a recent interest in features that go beyond this by including user information via embedded user graphs [11] [10]. Further, approaches that goes beyond treating hate online as a classification task are still rare. In Salminen et al. (2018) [16], hateful parts are removed from comments with the intention of keeping the semantics of the original content intact. Instead of deleting hate from comments, Chung et al. (2019) [2] suggest a system that automatically provides counter arguments to hateful comments.

3 Data

While the pre-processing for the BERT-based models is performed using the pre-defined tokenization pipelines of each pre-trained model, the data provided to the SVM underwent various external pre-processing steps including tokenization, the removal of stopwords (excluding negations), lowercasing, stemming and lemmatization.

We explore various external hate-speech corpora and their effect on the classification performance. However, as most corpora focus on different facets of hate, a one-to-one correspondence between labels is not always given. In such cases a mapping between similar labels was performed, which are described in table 1 along with the class distributions for task A and B for each corpus.

Corpora	Lang.	Source	Task A	Task B	Mappings (A)	Mappings (B)
Kaggle	en	Wikipedia	143.3/16.2	0.3/14.5/1.4	$\forall c = 0$ →NOT $\exists c = 1$ →HOF	obsc →PRFN id.hate →HATE rest →OFFN
Davidson	en	Twitter	2.5/12.3	0/11.5/0.8	none →NOT hate, off →HOF	offn →OFFN hate →HATE
Founta	en	Twitter	53.9/32.1	0/27.2/5.0	none →NOT hate, off →HOF	offn →OFFN hate →HATE
TRAC	en	Facebook	7.4/9.8		none →NOT aggr →HOF	
TRAC	hi	Facebook	3.4/13.8		none →NOT aggr HOF	
GermEval	de	Twitter	3.3/1.7	0.1/0.6/1.0	none →NOT hate →HOF	prfn →PRFN ins →OFFN hate →HATE

Table 1. Language and source of the comments collected for each of the additional data sets explored. The distribution of labels for task 1 (NOT/HOF) and task 2 (PRFN/OFFN/HATE) are reported in thousands. The mappings between original labels {*obscene* (obs), *identity hate* (id.hate), *none*, *offense* (offn), *hate*, *other*, *overtly/covertly agressive* (aggr), *profane* (prfn), *insult* (ins)} to HASOC compatible labels is given.

For **English**, we use four different external corpora. The *Kaggle* (KA) [1] corpus[2] is a large corpus of Wikipedia comments and includes several hate-related non-exclusive labels ranging from toxic, severe toxic and obscene to threat, insult and identity hate.

The *Davidson* (DA) [3] corpus[3] and the *Founta* (FO) [6] corpus[4] are both twitter corpora focusing on hate as well as offensive speech.

Lastly, we used the *TRAC* (TR) [9] corpus[5], focusing on overtly and covertly aggressive Facebook comments. Note that we also used the **Hindi** version of this dataset.

For **German**, we explored the GermEval 2018 (GE) [18] corpus[6] as additional data.

Most of the corpora have unbalanced classes. For task 1, the NOT class is often over-represented, which in its extremes leads to a ratio of 1:8835 hate to non-hate labels in the case of KA. However, for DA and TR, this unbalance is reversed, where more samples are marked as hateful than not. This unbalance is also present in task 2, where PRFN is heavily under-represented, followed by HATE for most corpora except GE. This unbalance, which can also be observed in the

[2] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data

[3] https://github.com/t-davidson/hate-speech-and-offensive-language

[4] https://github.com/ENCASEH2020/hatespeech-twitter

[5] https://sites.google.com/view/trac1/shared-task

[6] https://github.com/uds-lsv/GermEval-2018-Data

official HS training data, leads to special difficulties when training a classifier on these datasets.

4 BERT Classifier

We use pre-trained BERT [4] models to encode a tweet into a single vector. For this, we use monolingual cased BERT-base for English (BERT_{en})[7] and German (BERT_{de})[8] as well as multilingual cased BERT (BERT_{multi})[9]. The classifier is a linear layer of depth 1, mapping the encoded tweets to labels.

To deal with the unbalanced nature of the training data, we perform randomized weighted re-sampling of the data at each epoch. The weights given to a class is calculated such that underrepresented classes are given a larger weight and vice versa.

5 SVM Classifier

We have used a linear SVM classifier for task 1. Here, we explored different features including tf-idf, word and character n-grams as well as byte pair encoding (BPE) [7].

Table 2 depicts the best combination of features for each selection, used for the SVM submissions in sub-task A:

Lang.	**Feature Combination**
English	tf-idf + BPE + word n-grams(1, 3) + stopword
German	tf-idf + BPE + word n-grams(1, 3) + stopword
Hindi	tf-idf + BPE + word n-grams(1, 3)

Table 2. Best feature combination used for the SVM models for different languages.

6 Results

We train both BERT-based and SVM classifiers on different combinations of corpora, while 10-fold cross validation is performed on the official HASOC training data only. The results are reported in table 3.

When high-quality **monolingual** pre-trained models are available, these often yielded better results than their multilingual counterparts, i.e. +0.02 for HS_{en} and HS_{de} in task A, with the biggest gain in Macro F1 being +0.14 in

[7] `https://storage.googleapis.com/bert_models/2018_10_18/cased_L-12_H-768_A-12.zip`

[8] `https://deepset.ai/german-bert`

[9] `https://storage.googleapis.com/bert_models/2018_11_23/multi_cased_L-12_H-768_A-12.zip`

			Task A			Task B		
Lang.	Model	Data	F1	Macro	Micro	F1	Macro	Micro
en	BERT$_{en}$	HS$_{en}$	0.74/0.54	0.64	0.66	0.71/0.31/0.75	0.59	0.65
		+ KA	**0.76/0.56**	**0.66**	**0.68**	0.69/0.36/0.73	0.59	0.65
		+ DA	0.73/0.54	0.64	0.66	0.68/0.36/0.73	0.59	0.64
		+ FO	0.75/0.55	0.65	0.67	0.68/0.34/0.74	0.59	0.64
		+ TR$_{en}$	0.73/0.56	0.65	0.66			
	BERT$_{multi}$	HS$_{en}$	0.72/0.53	0.62	0.65	**0.72/0.37/0.75**	**0.61**	**0.66**
		HS$_{en+de+hi}$	0.73/0.54	0.63	0.65	0.71/0.36/0.74	0.60	0.66
	SVM	HS$_{en}$ + TR$_{en}$	0.65/0.48	0.56	0.65	–	–	–
de	BERT$_{de}$	HS$_{de}$	0.95/0.27	0.61	0.87	**0.40/0.69/0.38**	**0.49**	**0.54**
		+ GE	**0.94/0.40**	**0.67**	**0.88**	0.39/0.61/0.47	0.49	0.52
	BERT$_{multi}$	HS$_{de}$	0.94/0.23	0.59	0.87	0.12/0.65/0.27	0.35	0.46
		HS$_{en+de+hi}$	0.94/0.30	0.62	0.87	0.38/0.61/0.38	0.46	0.50
	SVM	HS$_{de}$ + GE	**0.94**/0.34	0.64	**0.88**	–	–	–
hi	BERT$_{multi}$	HS$_{hi}$	**0.79/0.81**	**0.80**	**0.80**	0.76/0.50/0.39	0.55	0.61
		HS$_{en+de+hi}$	**0.79/0.81**	**0.80**	**0.80**	**0.82/0.49/0.47**	**0.59**	**0.65**
	SVM	HS$_{hi}$ + TR$_{hi}$	0.72/0.81	0.77	0.80	–	–	–

Table 3. Scores as calculated using 10-fold cross validation: F1 for task 1 (NOT/HOF) and task 2 (PRFN/OFFN/HATE) labels as well as micro and macro F1 scores for several corpus combinations and models. Top scores are in bold.

the case of the monolingual HS$_{de}$ as opposed to its multilingual counterpart in task B. This comes to show that high quality monolingual models –if language model training data is available in abundance– can lead to great improvements over multilingual baselines. In fact, the usage of a high-quality monolingual pre-trained model applied to the severely low-resourced task B, yielded top results for German. Nevertheless, for HS$_{en}$, we observe a slightly better performance of the multilingual model in task B. One reason for this may be due to the nature of the training data, as the HS$_{en}$ contains India-related content as well as some Hinglish and code-switched sentences. This, together with the general enforced data sparsity in task 2, might have lead to the slight gain in macro F1 for the multilingual model.

For task A, adding **external data** either lead to slightly improved or unchanged results. For English, adding KA yielded an improvement of +0.02, which given the large size of the KA corpus is a modest increase. For German we observe a large increase in macro F1 (+0.06) when adding GE. This is most likely due to the larger amount of HOF-labeled data in the otherwise very similarly defined GE corpus. In general, the simplicity of the binary decision task still allows for external data to be of use –or at least not destructive– for the described task. However, when moving to the more complex task of identifying different shades of hate, external data quickly becomes reduced to additional noise during training, leading to either decayed or unchanged results for all external data in task B. This is especially interesting for GE, which has a very similar three-class corpus design (`profane`, `insult` and `abuse`). This comes to show that, as defini-

Task	Language	Run	Model	F1	Macro	Micro
A	en	1	BERT$_{en}$ on HS$_{en}$ + KA	0.63/0.52	0.58	0.60
		2	SVM on HS$_{en}$+TR$_{en}$	0.68/0.54	0.61	0.64
		3	Ensemble	0.78/0.59	0.69	0.73
	de	1	BERT$_{de}$ on HS$_{de}$ + GE	0.89/0.32	0.61	0.80
		2	SVM on HS$_{de}$ + GE	0.89/0.19	0.54	0.78
		3	Ensemble	0.87/0.32	0.59	0.78
	hi	1	BERT$_{multi}$ on HS$_{en+de+hi}$	0.67/0.75	0.71	0.71
		2	SVM on HS$_{hi}$ + TR$_{hi}$	0.78/0.79	0.78	0.78
		3	Ensemble	0.80/0.80	0.80	0.80
B	en	1	BERT$_{multi}$ on HS$_{en}$ (pipe)	0.63/0.62/0.23/0.27	0.44	0.57
		2	BERT$_{multi}$ on HS$_{en}$ (4-class)	0.68/0.65/0.25/0.29	0.47	0.61
		3	Ensemble	0.58/0.62/0.21/0.27	0.42	0.53
	de	1	BERT$_{de}$ on HS$_{de}$ (pipe)	0.89/0.17/0.18/0.15	**0.35**	**0.77**
		2	BERT$_{de}$ on HS$_{de}$ (4-class)	0.89/0.00/0.06/0.08	0.26	0.75
		3	Ensemble	0.66/0.04/0.36/0.15	0.28	0.58
	hi	1	BERT$_{multi}$ on HS$_{en+de+hi}$ (pipe)	0.67/0.79/0.40/0.29	0.54	0.60
		2	BERT$_{multi}$ on HS$_{en+de+hi}$ (4-class)	0.78/0.80/0.43/0.27	0.57	0.66
		3	Ensemble	0.78/0.79/0.40/0.35	0.58	0.66

Table 4. Scores on the official HASOC test sets: F1 for task 1 (NOT/HOF) and task 2 (PRFN/OFFN/HATE) labels as well as micro and macro F1 scores of submitted runs. Top scoring runs are in bold.

tions of hate and its sub-classes differ, and final annotations depend not only on the definitions provided but also on subjective choices of the annotators, different hate speech corpora become incompatible, thus enforcing the data sparsity in this field.

For all models in task A, the SVM models are outperformed by their BERT counterparts by margins between +0.03 (English and Hindi) and +0.09 (German).

6.1 Submitted Models

For **task A**, both BERT and SVM models as well as an ensemble of both are submitted. For each language, **run 1** is the ensemble of all 10 folds of the top-scoring BERT model. As the recall of the `NOT` class is generally low, it was boosted by labeling a test sample as `NOT` whenever any of the folds suggested this label. For **run 2**, an SVM version trained on the whole dataset was submitted. Lastly, **run 3** is the ensemble of all ten BERT folds and the SVM, using the same voting scheme as for run 1.

For **task B**, only BERT models were taken into consideration. As the test data provided still contained non-hateful comments, **run 1** uses the BERT ensemble from task A (run 1) to pre-select hateful comments, which are then further classified by an ensemble of the 10 folds of the top scoring BERT model in task 2. Here, a majority vote approach was taken, such that the label with the most votes is accepted. For **run 2**, alternative models trained on HS data only but also covering the `NONE` label have been trained and were ensembled using the majority vote approach. Finally, **run 3** is the ensemble of both run 1 and 2. The results of each run are reported in table 4.

7 Conclusion

We have trained both SVM and BERT-based classifiers on several combinations of external corpora. Primarily, we observed that finer-grained detection of different types of hate does not benefit from external corpora due to the incompatibility between different definitions of hate –and its subtypes– as well as the subjectivity of the matter, reducing external resources to added noise during training. This further enforces the data sparsity in the field of hate speech detection also for higher-resourced languages with several corpora available. We therefore want to underline two directions for future research in abusive language detection and similar fields: a) A special focus on low-resource text classification for improved results despite the lack of large amounts of mutually compatible labeled data and b) creating corpora of hate speech which go beyond ambiguously defined sub-categories of hate. For the latter, we plan to create a corpus which focuses on identifying different objective features within a comment –i.e. the targets of a sentiment (positive or negative), pragmatic cues such as the existence of an accusation, swear words or capitalization etc., — which in their sum will help to identify hateful content based on different subsets of such features.

References

1. van Aken, B., Risch, J., Krestel, R., Löser, A.: Challenges for toxic comment classification: An in-depth error analysis. In: Proceedings of the 2nd Workshop on Abusive Language Online, EMNLP 2018, Brussels, Belgium, October 31, 2018. pp. 33–42 (2018)
2. Chung, Y.L., Kuzmenko, E., Tekiroglu, S.S., Guerini, M.: CONAN - COunter NArratives through Nichesourcing: a multilingual dataset of responses to fight online hate speech. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 2819–2829. Association for Computational Linguistics, Florence, Italy (Jul 2019)
3. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Eleventh International AAAI Conference on Web and Social Media (May 2017)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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)

5. Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., Bhamidipati, N.: Hate speech detection with comment embeddings. In: Proceedings of the 24th International Conference on World Wide Web. pp. 29–30. ACM, New York, NY, USA (2015)
6. Founta, A.M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., Kourtellis, N.: Large scale crowdsourcing and characterization of twitter abusive behavior. In: Twelfth International AAAI Conference on Web and Social Media (2018)
7. Heinzerling, B., Strube, M.: BPEmb: Tokenization-free pre-trained subword embeddings in 275 Languages. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 7-12 2018)
8. Jha, A., Mamidi, R.: When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In: Proceedings of the second workshop on NLP and computational social science. pp. 7–16 (2017)
9. Kumar, R., Ojha, A.K., Malmasi, S., Zampieri, M.: Benchmarking aggression identification in social media. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). pp. 1–11 (2018)
10. Mishra, P., Del Tredici, M., Yannakoudakis, H., Shutova, E.: Abusive language detection with graph convolutional networks. In: Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT). pp. 2145–2150 (2019)
11. Mishra, P., Tredici, M.D., Yannakoudakis, H., Shutova, E.: Author profiling for abuse detection. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 1088–1098 (Aug 2018)
12. Modha, S., Mandl, T., Majumder, P., Patel, D.: Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In: Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (2019)
13. Mondal, M., Silva, L.A., Benevenuto, F.: A measurement study of hate speech in social media. In: Proceedings of the 28th ACM Conference on Hypertext and Social Media. pp. 85–94. ACM, New York, NY, USA (2017)
14. Park, J.H., Fung, P.: One-step and two-step classification for abusive language detection on twitter. In: ALW1: 1st Workshop on Abusive Language Online, Association for Computational Linguistics, Vancouver, Canada (2017)
15. Saleem, H.M., Dillon, K.P., Benesch, S., Ruths, D.: A web of hate: tackling hateful speech in online social spaces (2016)
16. Salminen, J., Luotolahti, J., Almerekhi, H., Jansen, B.J., Jung, S.g.: Neural network hate deletion: Developing a machine learning model to eliminate hate from online comments. In: Bodrunova, S.S. (ed.) Internet Science. pp. 25–39. Lecture Notes in Computer Science, Springer International Publishing (2018)
17. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: Proceedings of the NAACL student research workshop. pp. 88–93 (2016)
18. Wiegand, M., Siegel, M., Ruppenhofer, J.: Overview of the germeval 2018 shared task on the identification of offensive language (2018)
19. Wulczyn, E., Thain, N., Dixon, L.: Ex machina: Personal attacks seen at scale. In: Proceedings of the 26th International Conference on World Wide Web. pp. 1391–1399 (2017)