TUB at HASOC 2020: Character based LSTM for Hate Speech Detection in Indo-European Languages

Salar Mohtaj[a,b], Vinicius Woloszyn[a] and Sebastian Möller[a,b]

[a]*Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany*
[b]*German Research Centre for Artificial Intelligence (DFKI), Projektbüro Berlin, Berlin, Germany*

Abstract

This paper presents TU Berlin team experiments and results on the task 1 of the shared task on hate speech and offensive content identification in Indo-European languages. Recently, hate speech has become an important problem that is seriously affecting online social media. Large scale social platforms are currently investing important resources to automatically detect and classify toxic language. The competition evaluates the success of different natural language processing models on detecting hate speech in different languages, automatically. Among the state-of-the-art deep learning models that have been used for the experiments, the character based LSTM achieved the best results on detecting hate speech contents in tweets.

Keywords

Hate speech detection, Offensive Content Identification, Bert, LSTM

1. Introduction

With a massive increase of content generation on online social media, there has also been an increase of hateful and offensive language in online posts. It is possible to automate a part or the whole process of toxic language detection among the content that is generated by users by using Natural Language Processing (NLP).

HASOC (2020) at FIRE[1] provides a shared task and a data challenge for multilingual research on the identification of toxic content. HASOC offers the task of hate speech detection on English, German and Hindi languages, includes 2 sub-tasks, on annotated tweets from Twitter [1].

Sub-task A of HASOC (2020) is a binary classification task for identifying hate, offensive and profane content. Two classes include:

- ***(NOT)*** *Non Hate-Offensive*: These posts do not contain any hate speech, profane, offensive content

FIRE '20, Forum for Information Retrieval Evaluation, December 16–20, 2020, Hyderabad, India
salar.mohtaj@tu-berlin.de (S. Mohtaj); woloszyn@tu-berlin.de (V. Woloszyn); sebastian.moeller@tu-berlin.de (S. Möller)
https://salarmohtaj.github.io/ (S. Mohtaj);
https://www.qu.tu-berlin.de/menue/team/senior_researchers/vinicius_woloszyn/ (V. Woloszyn);
https://www.qu.tu-berlin.de/menue/team/professur/ (S. Möller)
0000-0002-0032-3833 (S. Mohtaj)

[1]Forum for Information Retrieval Evaluation (FIRE 2020) - http://fire.irsi.res.in/fire/2020/home

Language	Total # of Instances	Sub-task A		Sub-task B		
		NOT	HOF	HATE	OFFN	PRFN
English	5852	3591	2261	1143	451	667
German	3819	3412	407	111	210	86
Hindi	4665	2196	2469	556	676	1237

Table 1
Statistics of the *HASOC2020* **train** dataset

- ***(HOF)*** *Hate and Offensive*: These posts contain hate, offensive, and profane content

On the other hand, sub-task B is a three classes classification task to Discrimination between hate, profane and offensive posts in order to further classify tweets from the sub-task A into three categories.

- ***(HATE)*** *Hate speech*: Posts under this class contain Hate speech content
- ***(OFFN)*** *Offenive*: Posts under this class contain offensive content
- ***(PRFN)*** *Profane*: Posts contain profane words

The *TU Berlin* team take part in the sub-task A, where the state-of-the-art methods on text classification are applied on the tweets to categorize them into one of the aforementioned classes. In this paper the applied methods to pre-process and process the content are described in details.

The paper is organized as follows; Section 2 present a short description on the proposed datasets for train and test in the competition. The proposed approaches for data preprocessing and the experiments are described in details in Section 3. Section 4 contains the achieved results on the test data that is reported by the competition's organizers. Finally, in Section 5 we conclude the approaches and the results.

2. Data

In this section we briefly described the proposed dataset by task organizers in order to train and test models for the task of hate speech detection.

The HASOC dataset is sampled from Twitter and partially from Facebook in English, German and Hindi languages [2]. Some statistics of the train and test datasets are presented in Table 1 and 2, respectively. The content would contains hashtags, emojis, links and usernames that refer to a user on Twitter or Facebook. Moreover, some samples from the English dataset are shown in Table 3.

3. Experiments

In this section we describe the approaches that have been used in order to pre-process the data and also the state-of-the-art models that have been trained on the resulting text.

Language	Total # of Instances	Sub-task A		Sub-task B		
		NOT	HOF	HATE	OFFN	PRFN
English	814	391	423	25	82	293
German	526	392	134	24	36	88
Hindi	663	466	197	56	87	27

Table 2
Statistics of the *HASOC2020* **test** dataset

Sample Tweets	Classes	
	Sub-task A	Sub-task B
@realDonaldTrump TRAITOR #TrumpIsATraitor https://t.co/Ip9XqS0U3c	HOF	OFFN
If that Bengali Jihadi $lu7 catches any illness, no doctor must treat her. She should suffer from non-treatment and face consequences! #DoctorsFightBack #DoctorsUnderOppression #DoctorsProtest	HOF	HATE
@brianstelter @OANN I went looking for dickheads this weekend and I found you. #douchebag	HOF	PRFN
@HuffPost Is she wearing clothes? #TrumpIsATraitor	NOT	NONE

Table 3
Samples of tweets from the English train dataset in different classes

3.1. Data preprocessing

As mentioned in Section 1, the proposed data in the share task contains hashtags, mentioned usernames, links and emojis. To clean up the data by reserving important information from tweets and removing unimportant ones, we follow some of the steps that was used in [3] for the preprocessing. The following changes have been applied as the preprocessing steps:

- **Username mentions** (e.g., terms starting with @) are replaced with *'username'* phrase. Although the username itself do not contain important information for the task of hate speech detection, pointing in the tweet that it contains a username would improve the overall performance of the classifier.
- **Emojis** (i.e. smileys) are replaced with a short textual description that express the corresponding emotion, using *demoji*[2] package.
- **Links** are replaced with a *'link'* phrase. Like username, although individual links don't contain important information to be kept, referring to the model that a tweet includes links would improve the performance of the model. To this end all the terms started with *http* have been replaced with the *'link'* term.
- **Multiple white spaces** are replaced with a single white space.

[2]https://pypi.org/project/demoji/

- All the token in tweets are **Lower-cased**.

The same steps of preprocessing are applied on the train and the test datasets to empower the model in order to generalize information from the tweets, in all of the three languages (e.g., English, German and Hindi).

3.2. Models

Transformer based language models (e.g., *BERT* [4] and *ELMO* [5]) received lots of attention during last years and achieved stunning results in many NLP tasks. They have been also used by some of the participants of *HASOC (2019)* for the task of hate speech detection [6, 7].

We used a BERT based architecture and also a character based LSTM [8] model in our experiments. For the BERT based transfer learning approach, we fine-tuned weights from the pre-trained models based on the proposed data for the task of hate speech detection. We used both *bert-based* and *bert-large* from the huggingface[3] package with different sets of hyperparameters for the English tweets. Moreover, the corresponding German and Hindi models from the same package have been used for the two other languages.

In addition to the state-of-the-art BERT architecture, a character based LSTM model is also trained on the training datasets. For this end, a bidirectional two layer LSTM with embedding size of 200 and hidden layer size of 256 is trained.

For measuring the performance of the two models for the task of hate speech detection, a part of the training dataset has been separated for the test purpose. Our experiments on different hyperparameters show that the LSTM model outperform the BERT based model, from the F1 measure point of view. So, the LSTM model is submitted as the team's model for the competition. The result of the model on the test data for different languages is reported in the next section.

4. Results

In this section we present the achieved results on the test dataset in all of the three languages. In addition to applying the models on the test datasets that are published by the organizers, the model are also applied on approximately 15% of a private test dataset. The proposed model on the English dataset achieved a F1 accuracy of **0.504** (in macro average) and ranked 6th among 35 participated teams. The final results of top 10 teams on the English data is presented in Table 4.

The proposed model for the German dataset does not generated any positive label (HOF), neither on train and test datasets. The reason would be the high-class imbalance (10%-90% for HOF and NOT classes, respectively) in the German data, comparing to the English and Hindi datasets. The model achieved a F1 accuracy of **0.427** and ranked 19th among 20 participated teams in the German hate speech detection task. Finally, the proposed Hindi model achieved a F1 accuracy of **0.467** and ranked 19th among 24 teams in the Hindi task.

[3] https://huggingface.co/

#	Team Name	F1 Macro average
1	IIIT_DWD	0.5152
2	CONCORDIA_CIT_TEAM	0.5078
3	AI_ML_NIT_Patna	0.5078
4	Oreo	0.5067
5	MUM	0.5046
6	Huiping Shi	0.5042
7	**TU Berlin**	**0.5041**
8	NITP-AI-NLP	0.5031
9	JU	0.5028
10	HASOCOne	0.5018

Table 4
Final results on sub-task A English data (top 10 teams)

5. Conclusion and future work

In this paper, we presented the proposed models on the task 1 of the shared task on hate speech and offensive content identification in English, German and Hindi languages. We used a BERT based architecture and a character based LSTM model for training classifiers to detect offensive language among tweets. Our experiments show that LSTM model outperform the BERT based approach. The proposed model achieved the 6th best performance in the English data for task 1.

The achieved result can be improved by making the training data more balance, using up-sampling approaches. There is also design space based on Bert, for the specific architectures for our task that is a future direction of our work.

Acknowledgments

We would like to thank the organizers of *HASOC2020* shared task for organizing the competition and taking time on the inquiries.

References

[1] T. Mandl, S. Modha, G. K. Shahi, A. K. Jaiswal, D. Nandini, D. Patel, P. Majumder, J. Schäfer, Overview of the HASOC track at FIRE 2020: Hate Speech and Offensive Content Identification in Indo-European Languages), in: Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, CEUR, 2020.

[2] S. Modha, T. Mandl, P. Majumder, D. Patel, Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in indo-european languages, in: P. Mehta, P. Rosso, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019, volume 2517 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 167–190. URL: http://ceur-ws.org/Vol-2517/T3-1.pdf.

[3] B. Wang, Y. Ding, S. Liu, X. Zhou, Ynu_wb at HASOC 2019: Ordered neurons LSTM with attention for identifying hate speech and offensive language, in: P. Mehta, P. Rosso,

P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019, volume 2517 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 191–198. URL: http://ceur-ws.org/Vol-2517/T3-2.pdf.

[4] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.

[5] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: M. A. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), Association for Computational Linguistics, 2018, pp. 2227–2237. URL: https://doi.org/10.18653/v1/n18-1202. doi:10.18653/v1/n18-1202.

[6] T. Ranasinghe, M. Zampieri, H. Hettiarachchi, BRUMS at HASOC 2019: Deep learning models for multilingual hate speech and offensive language identification, in: P. Mehta, P. Rosso, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019, volume 2517 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 199–207. URL: http://ceur-ws.org/Vol-2517/T3-3.pdf.

[7] S. Mishra, S. Mishra, 3idiots at HASOC 2019: Fine-tuning transformer neural networks for hate speech identification in indo-european languages, in: P. Mehta, P. Rosso, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019, volume 2517 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019, pp. 208–213. URL: http://ceur-ws.org/Vol-2517/T3-4.pdf.

[8] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780. URL: https://doi.org/10.1162/neco.1997.9.8.1735. doi:10.1162/neco.1997.9.8.1735.