A Feature Extraction based Model for Hate Speech Identification

Salar Mohtaj[1,2], Vera Schmitt[1] and Sebastian Möller[1,2]

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

Abstract

The detection of hate speech online has become an important task, as offensive language such as hurtful, obscene and insulting content can harm marginalized people or groups. This paper presents TU Berlin team experiments and results on the task 1A and 1B of the shared task on hate speech and offensive content identification in Indo-European languages 2021. The success of different Natural Language Processing models is evaluated for the respective subtasks throughout the competition. We tested different models based on recurrent neural networks in word and character levels and transfer learning approaches based on Bert on the provided dataset by the competition. Among the tested models that have been used for the experiments, the transfer learning-based models achieved the best results in both subtasks.

Keywords

Hate speech detection, Offensive Content Identification, Bert, LSTM, English

1. Introduction

Using abusive language and hate speech in social media platforms can have devastating effects on internet users by promoting racism, hatred and violence [1]. Offensive language has even the potential to shape political campaigns [2]. Due to the openness, anonymity and informal structure, social medial platforms are particularly vulnerable to ill-intentioned activities [3]. The availability of large annotated corpora from social medial platforms and the development of powerful Natural Language Processing (NLP) has the potential to remedy the challenge of detecting hate speech online [4].

Most of the research in this domain is dedicated to English datasets only. Therefore, the Hate Speech and Offensive Content Identification (HASOC) track aims to provide a platform to develop and optimize algorithms for the hate speech detection task in different languages, such as Hindi, German and English [5]. This year HASOC provides a data challenge for multilingual research on the identification of offensive speech online at the Forum for Information Retrieval Evaluation (FIRE) 2021. HASOC has defined two subtasks, whereas the first subtask contains the identification and discrimination of hate, profane and offensive posts from Twitter in English,

Hindi and Marathi. The second subtask focuses on the identification of conversational hate-speech in Code-Mixed Languages. The TU Berlin team focuses on the subtasks 1A and 1B. Subtask 1A is a coarse-grained binary classification task where tweets should be classified into two classes:

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

Subtask 1B is is a three-class classification task offered for English and Hindi, where hate-speech, profane and offensive posts from subtask 1A are further classified into the following categories:

- *(HATE Hate speech: this class contains posts which hate-speech content*
- *(OFFN Offensive: posts in this class contain offensive content*
- *(PRFN Profane: posts in this class contain profane content*

In this paper the proposed models for classifying tweets into one of the classes for the respective subtask are presented. For this purpose, the state-of-the-art NLP methods are applied to classify the posts and categorize them into the classes. Hereby the team of TU Berlin focuses on the English dataset. We used transfer learning models based on the BERT language model [6], and also Recurrent Neural Networks (RNNs), either in word and character levels to categorize tweets into the relevant classes.

The following section 2 describes some of the state-of-the-art models for the task of hate speech detection in English. Section 3 describes the provided train and test dataset, whereas section 4 contains details about data processing and the experiments and models applied. Furthermore, in section 5 the achieved results are analyzed, and section 6 summarizes and concludes the approaches and results.

2. Related Work

In this section, we overview some of the recent approaches for automatic hate speech detection from English text. Although the automated approaches for hate speech detection could be categorized into keyword-based, source metadata and machine learning based approaches [7], in this section we focus on some of the state-of-the-art machine learning based models.

Among the proposed models for the HASOC shared task on 2020, Mishra et al. has been used a Long Short-Term Memory (LSTM) based model using Glove vectors [9] for the embedding [8]. They fed the outputs of the embedding layer to a single layer LSTM network and put a fully connected layer on top. In this year's competition we tried to use a similar architecture as one of our experiments. On the other side, the *YNU_OXZ* team [10] in the HASOC 2020 competition proposed a model based on *XLM-RoBERTa* [11] and LSTMs. In their model, they concatenated the output of the last layer hidden state of *XLM-RoBERTa* and the hidden state of the last four layers of *XLM-RoBERTa* that is fed into an Ordered Neurons LSTM (ON-LSTM) [12]. Finally, they input these vectors into a fully connected network for the final classification.

Badjatiya et al. did different experiments based on three different neural network architecture to detect hate speech tweets in Twitter [13]. They used convolutional Neural Networks (CNNs), LSTM, and FastText [14], with either random embeddings or GloVe embeddings.The proposed models categorize tweets as racist, sexist or neither. Their experiments show that the model based on LSTM, random embedding and Gradient Boosted Decision Trees outperforms the other models in terms of precision, recall, and F1 score.

3. Data

The English dataset of HASOC 2021 for the subtasks 1A and 1B, contains the text content of tweets in English, IDs, and the labels for subtask 1A and 1B, respectively. the statistics of the training dataset is presented in Table 1. Moreover, the test dataset contains **1281** tweets which should be categorized into one of the classes based on the subtask.

The content would contains hashtags, emojis, links and usernames that refer to a user on Twitter. A sample of the dataset in different categories is presented in Table 2. More details about the datasets are provided in [15, 16].

4. Experiments

This section contains a short description on the used pre-processing steps and also the developed models and experiments for the task of hate speech detection.

4.1. Data Processing

For pre-processing of the raw data, we followed the same procedure as the experiments on the last year's competition [17]. The data pre-processing mainly includes the replacement of mentions with the phrase *'username'*, replacement of emojis with short textual descriptions, links are also replaced with the phrase *'link'*, and the replacement of multiple white spaces with a single white space. These steps are applied to both, the train and test datasets in order to facilitate the training process.

4.2. Models

The best performance of the last year HASOC competition for the English dataset have been achieved by [18] with a LSTM using GloVe embeddings [9] as input. Furthermore, transformer based language models such as BERT [6], DistilBERT and RoBERTa [19], and also ELMO [20]

Language	Total # of Instances	Subtask 1A		Subtask 1B			
		HOF	NOT	HATE	OFFN	PRFN	NONE
English	3843	2501	1342	683	622	1196	1342

Table 1
Statistics of the *HASOC2021* **training** dataset for subtasks 1A and 1B

Sample Tweets	Classes	
	Sub-task 1A	Sub-task 1B
This is enough of yours Modi This is not skill India it is kill India @narendramodi #ExitModi #Resign_PM_Modi https://t.co/m9FZyU4Lfg	HOF	OFFN
Please, abdicate! You failed us. You failed everyone. Everyone is suffering. EVERYONE! #ModiKaVaccineJumla	HOF	HATE
@Feisty_Waters Ok. What did you do to piss off the universe?	HOF	PRFN
@ndtv Nothing gonna help you please #Resign_PM_Modi	NOT	NONE

Table 2
Samples of tweets from the English train dataset in different classes

showed also promising results for similar task. Therefore, the TU Berlin team focuses on BERT based transfer learning approaches for the proposed subtasks. We also did some experiments on character level LSTM models which achieved our best results on the last year's competition [17].

4.3. LSTM based models

We developed two different models based on LSTM networks. We developed a smaller, character based architecture, Char_LSTM hereinafter, and a deeper and more complex network based on words, Word_LSTM hereinafter. Since people sometimes do minor changes on the words (e.g., by repeating some characters) when they express hate speech, a word based model may not signal those terms properly. As a result, we also developed a character based model to compare the outcomes of the models.

For the Char_LSTM, we tried out different hyper-parameters that includes:

- Embedding dimension [50, 100, 200]
- Hidden dimension [16, 32, 64, 128]
- Dropout [0.25, 0.5, 0.75]

The range of the above mentioned hyper-parameters for the Word_LSTM model are as follow:

- Embedding dimension [100, 300]
- Hidden dimension [32, 64, 128, 256, 512]
- Dropout [0.25, 0.5, 0.75]

In our experiments, the batch size of 32 and, the Adam optimizer [21] and the Binary Cross Entropy (BCE) loss function have been used in both models. In the Word_LSTM model, we tested either using Glove pre-trained vectors and training the embedding layer from scratch. The detailed results of the proposed models are presented in the section 5.

4.4. Bert based models

In addition to the models based on Recurrent Neural Networks (RNNs), we tested two transfer learning based models using BERT language model [22]. In one of the experiments, we fine-tuned English Bert for the task of hate speech identification. For this purpose, we followed the recommended hyper-parameters by the authors [22].

As the other transfer learning based model, we used Bert for extracting features from textual data. In other words, in this approach, the Bert language model was used to convert text data into vectors. The resulting vectors inputted into a Gated Recurrent Units (GRU) network. Different hyper-parameters tested on the data to choose the best parameters. The range of different hyper-parameters which had been used in the feature extraction approach are as follow:

- Hidden dimension [32, 64, 128, 256, 512]
- Dropout [0.25, 0.5, 0.75]

Like the LSTM based models, the batch size of 32 and, the Adam optimizer [21] and the Binary Cross Entropy (BCE) loss function have been used in this experiment. We present the detailed results by the different architectures in section 5.

5. Results

In this section the achieved results on the training data are presented. For doing the experiments, the training dataset has been divided into train, validation and test datasets. The train part contains 70% of the whole data, the validation part consist of 10% of the data, and the test part contains the remaining 20% of the provided dataset.

We tested all of the mentioned models with different hyper-parameters. The best achieved results are shown in tables 3 - 5. In order to determine the impact of the pre-processing steps on the final results, we've repeated the experiments with the same hyper-parameters without applying the pre-processing steps. Although the runs without applying pre-processing could achieve competitive results in some cases, the experiments based on the pre-processed data outperforms the other ones in most of the cases. The performance of the submitted models for both sub-tasks are reported in details in [15].

The same architectures have been trained on the data for the sub-task 1B. The best achieved results on the second task were applied on the sub-task 1B test dataset and submitted to the shared task.

6. Conclusion and Future Work

In this paper, we presented the proposed models on the task 1A and 1B of the shared task on hate speech and offensive content identification in English. We used a BERT based architecture and word and character based LSTM models for training a model to classify tweets into offensive and not offensive categories. Our experiments show that Bert based model outperform the other approaches.

| Model name | Pre-processed | Hyper-parameters | | | F1 |
		Embedding dimension	Hidden dimension	dropout	
Char_LSTM	yes	50	256	0.5	0.75
	yes	50	128	0.75	0.78
	yes	100	64	0.5	0.76
	yes	200	16	0.5	**0.79**
	no	200	16	0.75	0.75
	no	100	128	0.75	0.77
Word_LSTM	yes	100	512	0.25	0.81
	yes	300	256	0.25	**0.83**
	yes	300	256	0.75	0.80
	no	300	256	0.25	0.79

Table 3
The achieved results by the **character based** and **word based LSTM** models for the sub-task 1A

| Model name | Pre-processed | Hyper-parameters | | | F1 |
		Bert model	Hidden dimension	dropout	
BERT feature extraction	yes	base	256	0.25	**0.86**
	yes	base	128	0.25	0.83
	yes	large	256	0.5	0.84
	yes	large	128	0.25	0.80
	no	base	128	0.25	0.79

Table 4
The achieved results by the **BERT feature extraction based** model for the sub-task 1A

| Model name | Pre-processed | Hyper-parameters | F1 |
		Bert model	
BERT fine-tuning	yes	base	0.81
	no	base	**0.83**

Table 5
The achieved results by the **BERT fine-tuning** model for the sub-task 1A

Since over-fitting was one of the main issues for training different models during the competition, enriching the training data by adding data samples from different resources could be a possible solution for improving the results. Moreover, the proposed transfer learning based results could be compared with the results from the the other state-of-the-art language models like GPT-3 to check if there is a significant difference in the performances.

Acknowledgments

and taking time on the inquiries.

References

[1] J. Kim, C. Ortiz, S. Nam, S. Santiago, V. Datta, Intersectional bias in hate speech and abusive language datasets, CoRR abs/2005.05921 (2020). URL: https://arxiv.org/abs/2005.05921. arXiv:2005.05921.

[2] I. Gagliardone, M. Pohjonen, Z. Beyene, A. Zerai, G. Aynekulu, M. Bekalu, J. Bright, M. Moges, M. Seifu, N. Stremlau, et al., Mechachal: Online debates and elections in ethiopia-from hate speech to engagement in social media, Available at SSRN 2831369 (2016).

[3] H. S. Alatawi, A. M. Alhothali, K. M. Moria, Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert, IEEE Access 9 (2021) 106363–106374.

[4] K. Florio, V. Basile, M. Polignano, P. Basile, V. Patti, Time of your hate: The challenge of time in hate speech detection on social media, Applied Sciences 10 (2020). URL: https://www.mdpi.com/2076-3417/10/12/4180. doi:10.3390/app10124180.

[5] T. Mandl, S. Modha, A. K. M, B. R. Chakravarthi, Overview of the HASOC track at FIRE 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german, in: P. Majumder, M. Mitra, S. Gangopadhyay, P. Mehta (Eds.), FIRE 2020: Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, ACM, 2020, pp. 29–32. URL: https://doi.org/10.1145/3441501.3441517. doi:10.1145/3441501.3441517.

[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceeding sof the 2019 Conference of the North American Chapter of the Association for ComputationalLinguistics: Human Language Technologies, NAACL-HLT 2019, volume 1 (Long and Short Papers), 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:doi:10.18653/v1/n19-1423.

[7] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, O. Frieder, Hate speech detection: Challenges and solutions, PloS one 14 (2019) e0221152.

[8] A. K. Mishra, S. Saumya, A. Kumar, Iiit_dwd@hasoc 2020: Identifying offensive content in indo-european languages, in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 139–144. URL: http://ceur-ws.org/Vol-2826/T2-5.pdf.

[9] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2014, pp. 1532–1543. URL: https://doi.org/10.3115/v1/d14-1162. doi:10.3115/v1/d14-1162.

[10] X. Ou, H. Li, Ynu_oxz at HASOC 2020: Multilingual hate speech and offensive content identification based on xlm-roberta, in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad,

India, December 16-20, 2020, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 121–127. URL: http://ceur-ws.org/Vol-2826/T2-3.pdf.

[11] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 8440–8451. URL: https://doi.org/10.18653/v1/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

[12] Y. Shen, S. Tan, A. Sordoni, A. C. Courville, Ordered neurons: Integrating tree structures into recurrent neural networks, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net, 2019. URL: https://openreview.net/forum?id=B1l6qiR5F7.

[13] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in: R. Barrett, R. Cummings, E. Agichtein, E. Gabrilovich (Eds.), Proceedings of the 26th International Conference on World Wide Web Companion, Perth, Australia, April 3-7, 2017, ACM, 2017, pp. 759–760. URL: https://doi.org/10.1145/3041021.3054223. doi:10.1145/3041021.3054223.

[14] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, in: M. Lapata, P. Blunsom, A. Koller (Eds.), Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers, Association for Computational Linguistics, 2017, pp. 427–431. URL: https://doi.org/10.18653/v1/e17-2068. doi:10.18653/v1/e17-2068.

[15] S. Modha, T. Mandl, G. K. Shahi, H. Madhu, S. Satapara, T. Ranasinghe, M. Zampieri, Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages and Conversational Hate Speech, in: FIRE 2021: Forum for Information Retrieval Evaluation, Virtual Event, 13th-17th December 2021, ACM, 2021.

[16] T. Mandl, S. Modha, G. K. Shahi, H. Madhu, S. Satapara, P. Majumder, J. Schäfer, T. Ranasinghe, M. Zampieri, D. Nandini, A. K. Jaiswal, Overview of the HASOC subtrack at FIRE 2021: Hate Speech and Offensive Content Identification in English and Indo-Aryan Languages, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021. URL: http://ceur-ws.org/.

[17] S. Mohtaj, V. Woloszyn, S. Möller, TUB at HASOC 2020: Character based LSTM for hate speech detection in indo-european languages, in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 298–303. URL: http://ceur-ws.org/Vol-2826/T2-26.pdf.

[18] A. K. Mishra, S. Saumya, A. Kumar, Iiit_dwd@hasoc 2020: Identifying offensive content in indo-european languages, in: P. Mehta, T. Mandl, P. Majumder, M. Mitra (Eds.), Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 139–144. URL: http://ceur-ws.org/Vol-2826/T2-5.pdf.

[19] R. Kumar, B. Lahiri, A. K. Ojha, A. Bansal, Comma@fire 2020: Exploring multilingual joint training across different classification tasks, in: P. Mehta, T. Mandl, P. Majumder,

M. Mitra (Eds.), Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 16-20, 2020, volume 2826 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 823–828. URL: http://ceur-ws.org/Vol-2826/T10-3.pdf.

[20] 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.

[21] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL: http://arxiv.org/abs/1412.6980.

[22] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.