CFILT IIT Bombay at HASOC 2020: Joint multitask learning of multilingual hate speech and offensive content detection system

Pankaj Singh[a], Pushpak Bhattacharyya[a]

[a]*Indian Institute of Technololgy, Bombay, India*

Abstract

This paper describes our system submitted to HASOC FIRE 2020. The goal of the shared tasks was to detected hate speech and offensive content in three languages namely Hindi, English, and German. The first subtask was a binary classification of a sentence into hate and offensive and normal. In the second subtask, a more granular classification of hate/offensive sentences was required. So overall there were 6 subtasks, 2 per language for 3 languages. We propose a system that performs all these tasks with a single model by jointly training a multilingual system on a combined corpus for all languages. It is relatively easy to fine-tune a model per task but it can pose various problems during deployment. These days most of the online platform supports multiple languages and it is not practical to deploy one model per language or per task. There are so many languages and tasks to cover and the online system will quickly run into memory and latency issues if there were multiple models handling the same task for different languages. Our system is capable of handling all subtasks for three languages with a single deep learning model. On the test set, we achieved a weighted average f1-score of 0.62, 0.85, 0.75 on subtask A and 0.35, 0.51, 0.43 on subtask B for Hindi, English, and German respectively.

Keywords

BERT, Multilingual Hate Speech and Offensive Content Detection, Multi-task learning

1. Introduction

With the rising popularity of the internet, there has been the growth of various online platforms and social media platforms being among them. Currently, we have multiple social media platforms operating globally across many countries and regions. There is a need for automatic monitoring systems for these social media platforms to detect unsocial elements such as hate speech and offensive content which can quickly disrupt the harmony in societies. This problem becomes more challenging when we want a platform to support multiple languages to fulfill the needs and enhance the experience of users from various backgrounds. The solution needs to scalable across multiple languages and multiple tasks. HASOC 2020 provides a platform to test hate speech and offensive content detection through the competition organized by them. We proposed and evaluated a system capable of supporting multiple languages and performing multiple tasks using a single deep learning model. The requirement of such systems is becoming obvious day by day as many social media platforms started supporting a large number of

FIRE '20, Forum for Information Retrieval Evaluation, December 16–20, 2020, Hyderabad, India.

pankajsingh7@iitb.ac.in (P. Singh); pb@cse.iitb.ac.in (P. Bhattacharyya)

https://www.cse.iitb.ac.in/~pb/ (P. Bhattacharyya)

languages and it is very inefficient to have one model per language during deployment as it increases the memory requirement and also adds up the training time.

In HASOC 2019 [1], a similar shared task competition was organized and received many submissions from researchers across the globe. Although various machine learning model was proposed but the transformer-based model seemed to be a popular choice and also best performing. Many other shared task such as GermEval 2018 [2], HatEval [3] and OffensEval [4] has been organized to push the research in this area.

In HASOC: FIRE 2020 [5], organizers have presented two challenges and provided the training and test datasets. These two subtasks were for three Indo-European Languages- Hindi, English, and German. The dataset was curated by collecting tweets and manually annotating the dataset for hate speech and offensive content. The challenge consists of the following two tasks for each of the three languages mentioned above:

- **Sub-task A- Identifying Hate, offensive, and profane content:** This was basically a binary classification task where each of the tweets was required to be classified either as a normal tweet or hate speech/offensive.
- **Sub-task B- Discrimination between Hate, profane and offensive posts:** This was a further granular classification of hate speech and offensive tweets. Each of the hate speech or offensive tweets was required to be classified as either hate or offensive or profane. This subtask was relatively more challenging than the first one due to an increase in the number of classes and a reduction in dataset size as only tweets labeled as hate speech or offensive are relevant for training.

We propose a joint multitask learning approach to perform all the six subtasks in the challenge using a single deep learning model. We combined the datasets of all three languages and fine-tuned a multilingual BERT [6] on two subtask A and B together. The results of this system were pretty competitive and reduced the resource requirements during training and inference. In section 2 we explain our system and training method in detail. In section 3 we report the performance of our system for various subtasks on the test dataset.

2. Materials and method

In this section, we provide the dataset description and statistics, deep learning network architecture, and its joint training on multilingual corpora on both subtasks.

2.1. Dataset

Organizes have collected tweets and annotated them for subtask A subtask B in all three languages. This dataset was made available to the participants along with the dataset of HASOC: 2019. However, we used only HASOC 2020 dataset to train and evaluate our system. Since, the dataset was for Hindi, English, and German languages so it contained both Roman and Devanagari scripts. Hindi language dataset also contained some amount of code-mixing and transliteration. For subtask A, each tweet was labeled as either (NOT) Non-Hate-Offensive or (HOF) Hate and Offensive and for subtask B, each of the (HOF) labeled tweets was further

Table 1
Dataset statistics for subtask A and B

Language	Total tweets	NOT	HOF	HATE	OFF	PRFN
Hindi	2963	2116	847	465	234	148
English	3708	1852	1856	1377	321	158
German	2373	1700	673	387	146	140

categorized into (HATE) Hate speech, (OFFN) Offensive, and (PRFN) Profane. Table 1 provides details about the number of tweets per class in the training dataset for three languages. This also shows class imbalance in the dataset which is an important issue to tackle while building robust hate speech and offensive content detection systems. We split the provided dataset into five-folds and performed 5-fold cross-validation.

2.2. System description

Since the dataset was directly scraped from Twitter and raw tweets contained various unnecessary features we processed the text before feeding it to the deep learning model. As a pre-processing step we performed the following tasks:

- Removed @mentions and RT from the tweets
- Replaced website URLs with string URL
- Removed # character from words used as hashtags
- Removed multiple spaces, if present any sentence

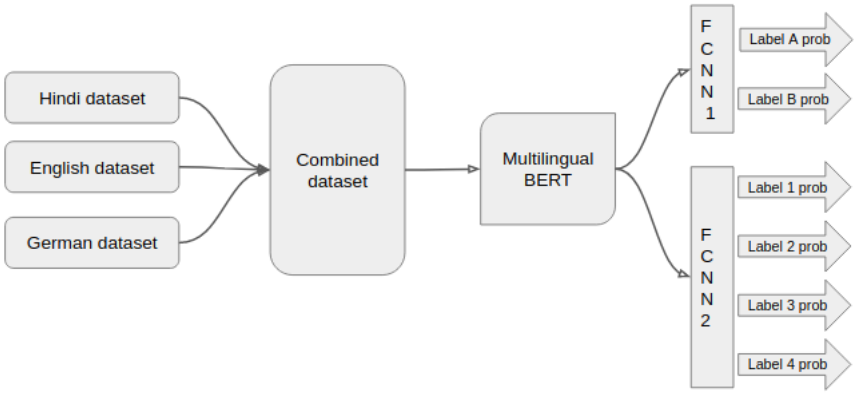

Figure 1: Overview of the proposed system

As a deep learning model, we choose multilingual BERT [6] and trained jointly on both subtasks on a combined corpus of all three languages. The final hidden state vector of special token [CLS] is taken as an aggregate representation [7] of the entire tweet and this 768-dimensional vector is passed throughout two different fully connected neural networks. One neural network

Table 2
Performance of our system on test dataset provided by organizers

Language	Subtask	Precision	Recall	F-Score	Accuracy
Hindi	A	0.65	0.61	0.62	0.72
	B	0.38	0.35	0.35	0.70
English	A	0.85	0.85	0.85	0.85
	B	0.57	0.51	0.51	0.80
German	A	0.77	0.74	0.75	0.82
	B	0.45	0.43	0.43	0.75

ends with a softmax layer having two heads responsible for subtask A and the second neural network ends with a softmax layer having four heads responsible for subtask B. In subtask B we did four-class classification by also considering NOT labeled tweets for training along with HATE, OFF and PRFN labels. Figure 1 depicts the overview of the proposed multilingual deep learning system. We combined the loss from both the network heads and the back-propagate average of these two losses. The entire network was then jointly trained on both subtasks by gradually unfreezing the layers of the multilingual BERT model.

3. Experiments and results

We did extensive hyperparameter tuning to get the best performance from the system. The deep learning model was trained on a combined corpus of all three languages and jointly fine-tuned for both the subtasks of each language. Since there was a class imbalance in the training dataset, hence we employed weighted cross-entropy loss giving more weight to counter the under-representation of some classes.

In table 2, we report the performance of our system on the test set provided by the organizers. Subtask A was a binary classification system having two class labels, NOT and HOF. Subtask B was a four-class classification system with four labels, NONE, HATE, OFF, and PRFN. We used the macro average f1-score as an evaluation metric. We also report accuracy, macro average precision, and recall for each subtask of six languages. In leaderboard scores, on average there was an absolute difference of 0.0356 in the macro average f-score of our system and top-3 best-performing systems in individual tasks. Given that we have trained a single deep learning model for every subtask this trade-off between f-score and resources (memory and latency) seems promising. In table 3, we compare the the performance of our system with top-3 performing system as per leaderboard published by organizers.

4. Conclusion

With an increase in diversity and number of users, online platforms have to support multiple languages. This leads to the demand for language scalable solutions if we want to perform hate speech and offensive content detection. Having one deep learning model per language or per task will be a very inefficient solution in deployment if our platform has to support hundreds of

Table 3
Performance of our system relative to top-3 performing submissions published in leaderboard (F1 macro average)

Task	Top-1 System	Average of Top-3 Systems	Our System
Hindi Subtask A	0.5337	0.5332	0.4834
Hindi Subtask B	0.3345	0.2885	0.2355
English Subtask A	0.5152	0.5102	0.4889
English Subtask B	0.2652	0.2640	0.2229
German Subtask A	0.5235	0.5220	0.5028
German Subtask B	0.2920	0.2890	0.2594

languages and multiple tasks. We purposed and established the efficacy of a multilingual and multitask system that can support three languages and perform two tasks for each language. The performance of our system was very competitive and at par with single individual models fine-tuned for one task.

In the future, we would like to expand our system to more languages and increase the number of tasks it can perform. We will also explore the use of other multilingual transformer models and do a comparative analysis.

Acknowledgments

We thank all the organizers of HASOC, FIRE 2020 for arranging this opportunity to push the research in multilingual hate speech and offensive content detection. We also express our gratitude towards them for their continuous support throughout the competition and for being very accommodating towards the requests from participants.

References

[1] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, A. Patel, 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, FIRE '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 14–17. URL: https://doi.org/10.1145/3368567.3368584. doi:10.1145/3368567.3368584.

[2] M. Wiegand, M. Siegel, J. Ruppenhofer, Overview of the germeval 2018 shared task on the identification of offensive language, Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria – September 21, 2018, Austrian Academy of Sciences, Vienna, Austria, 2018, pp. 1 – 10. URL: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-84935.

[3] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, 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,

Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63. URL: https://www.aclweb.org/anthology/S19-2007. doi:10.18653/v1/S19-2007.

[4] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval), in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 75–86. URL: https://www.aclweb.org/anthology/S19-2010. doi:10.18653/v1/S19-2010.

[5] 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.

[6] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual BERT?, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4996–5001. URL: https://www.aclweb.org/anthology/P19-1493. doi:10.18653/v1/P19-1493.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://www.aclweb.org/anthology/N19-1423. doi:10.18653/v1/N19-1423.