

Multilingual Hate Speech and Offensive Content Detection using Modified Cross-entropy Loss

Arka Mitra¹, Priyanshu Sankhala²

¹Indian Institute of Technology, Kharagpur, India

²National Institute of Technology Raipur, India

Abstract

The number of increased social media users has led to a lot of people misusing these platforms to spread offensive content and use hate speech. Manual tracking the vast amount of posts is impractical so it is necessary to devise automated methods to identify them quickly. Large language models are trained on a lot of data and they also make use of contextual embeddings. We fine-tune the large language models to help in our task. The data is also quite unbalanced; so we used a modified cross-entropy loss to tackle the issue. We observed that using a model which is fine-tuned in hindi corpora performs better. Our team (HNLP) achieved the macro F1-scores of 0.808, 0.639 in English Subtask A and English Subtask B respectively. For Hindi Subtask A, Hindi Subtask B our team achieved macro F1-scores of 0.737, 0.443 respectively in HASOC 2021.

Keywords

Hate speech detection, Text classification, Deep-learning, Transfer learning

1. Introduction

With the increased use of social media platform like Twitter, Facebook, Instagram, and YouTube by users around the world, the platforms have had positive aspects including but not limited to social interaction, meeting like-minded people, giving a voice to each individual to share their opinions [1]. However, as a result, social media platforms can also be used to spread hate comments, hate posts by certain individuals or groups; which can lead to having anxiety, mental illness and severe stress to people who consume that hate content [2]. It becomes necessary to be able to detect such activities at its earliest to stop it from spreading, thereby making social media a healthy place to interact and share their views without a fear of getting hate comments[3].

The hate posts can be insults or racist or discriminating on the bases of a particular gender, religion, nationality, age bracket, ethnicity. Such comments can also lead to goading of violence amongst people. With the large number of posts being shared each minute, it is not possible to manually classify each of the posts. Thus, a pre-programmed system is required to distinguish Hate speech activities quickly as hate content gains a lot of attention and is subject to be shared fast as well [4]. Direct targeted abuses and profane content are not that difficult to classify.

Forum for Information Retrieval Evaluation, December 13-17, 2021, India


✉ thearkamitra@gmail.com (A. Mitra); priyanshu.nitr.ele@gmail.com (P. Sankhala)

🌐 <https://thearkamitra.github.io/> (A. Mitra); <https://priyanshusankhala.github.io/> (P. Sankhala)

🆔 0000-0003-1071-7294 (A. Mitra); 0000-0003-2796-0039 (P. Sankhala)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

However, it is extremely hard to recognize indirect hate content often involving use of humour, irony, sarcasm even for an human annotator when the context of the posts are not provided. This makes the classification task additionally more difficult for most progressive frameworks. HASOC 2021 [5] is a shared task for the identification of hateful and offensive content in English and Indo-Aryan Languages. We participated in two sub-tasks for English and Hindi language [6].

The sub task A refers to classifying twitter samples into:

- HOF Hate and offensive :- contains hate speech/profane/offensive content.
- NOT Non Hate-offensive :- which does not contain any hate speech, profane, offensive content.

The sub task B refers to classifying twitter samples into:

- HATE Hate speech :- Posts under this class contain Hate speech content.
- OFFN Offensive :- Posts under this class contain offensive content.
- PRFN Profane :- These posts contain profane words.
- NONE Non-Hate :- These posts do not contain any hate speech content.

For tasks pertaining to English language, we experimented with large language models like fine-tuning BERT (Bidirectional Encoder Representation from Transformer) [7], RoBERTa (A Robustly Optimized BERT Pretraining Approach) [8] and XLNet (Generalized Autoregressive Pretraining for Language Understanding) [9] out of which RoBERTa outperformed others with the macro F1-score of 0.8089 while BERT and XLNet had the macro F1-score of 0.8050 and 0.7757 respectively in Subtask A and for Subtask B the macro F1-score was 0.6396 with RoBERTa model respectively. For the tasks referring to Hindi language, the authors used a model which is fine-tuned on detecting Hinglish sentiments [10] and had the macro F1-score of 0.7379 for Subtask A and macro F1-score of 0.4431 for Subtask B.

2. Related Work

In this section, we will discuss the previous state of the art methods proposed for detection of hate speech. The use of BERT and other transfer learning algorithms, and deep neural models based on LSTMs and CNNs tend to perform similar but better than traditional classifiers such as SVM [11]. The number of papers, trying to automate Hate speech detection, that have been published in Web of Science has been increasing exponentially [12]. Waseem et al. [13] have classified hate speech into different categories and led to the Offensive Language Identification Dataset (OLID) [14].

There has been work in different sub fields of abuse like in sexism [15, 16], cyberbullying [17], trolling [18] and so on. There are hate comments in most of the social media sites like Youtube [19], Instagram [20] which shows the importance of having a generalized Hate detection model [13]. Work done by Yin et al. [21] gives an overall idea of the generalizability of the different models that are present for hate speech detection. For the different models, the features from the input that are used have a great impact on the performance. Xu et al. [22] showed that part-of-speech tags are quite successful for improving the model; it is further improved by

considering the sentiment values [23]. The sentences in the online platforms do not always follow the normal textual formats or correct spellings. Thus, Mehdad et al. [24] used a character level encoding rather than using the word level encoding proposed by Meyer et al. [25]. The type of architecture used also impacts on the performance on the model. Swamy et al. [26] performed a comprehensive study that shows how different models perform and generalize.

3. Methodology

HASOC 2021 [6] has been going on for two years now and a lot of different ways are uncovered to detect hate content [27, 28]. This paper covers the use of large language models for classification of hate speech content.

3.1. Languages

The Hate speech and Offensive Content Identification in English and Indo-Aryan Languages HASOC 2021 [5, 6] purposes two different tasks, in 3 different languages English, Hindi, Marathi. The authors participated in both tasks for English and Hindi languages.

3.2. Task description

The first task in all languages know as "Subtask A" refers to a classification problem of twitter samples which were labelled as HOF- Hate and offensive content and NOT- Not hate and offensive content. The second task, know as "Subtask B" refers to a classification of twitter samples which were labelled as PRFN- Profane Words, HATE- Hate speech, OFFN- Offensive Content, and NONE- Non-hate content. The detailed description of all columns present in a dataset is given in Table 1 and the number of twitter samples corresponding to each label is given in Table 2.

3.3. Approach

The dataset that is provided in all the subtasks has an unequal number of samples per class. Table. 2 shows the overall distribution. For subtask A for English, the ratio of the classes (HOF and NOT) is around 2:1 while for Hindi it is around 1:2. Again, for subtask B, the ratio of the classes (PRFN, HATE, OFFN, NONE) is about 2:1:1:2 for english and approximately 2:5:6:30 for Hindi. From the ratio, one can understand that it would be unjust for the loss for each class to be the same. The cross entropy loss assigns same value to a probability score irrespective of the number of times it is present in the training set. To mitigate this, the authors have used modified cross-entropy loss as shown in Eqn. 1; it assigns a greater loss whenever a class with smaller frequency is misclassified. The weights factor in Eqn. 1 has a higher value for a class if the class has a lower frequency. This penalizes the model whenever that class is wrongly predicted and helps to improve the performance of the model.

$$loss(logits, class) = weight[class] * (-logits[class] + \log(\sum_j exp(logits[j]))) \quad (1)$$

Table 1

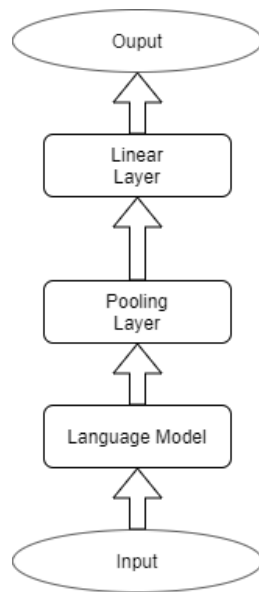
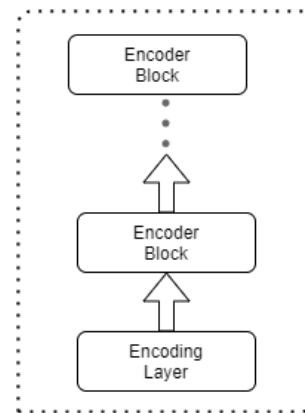
The detailed data description is given in table below:-

Columns	Description
tweet_id	unique value for the tweets
text	full text of the tweets
task1	label, either tweet is HOF or NOT for Subtask A
task2	label, either tweet is HATE, OFFN or PRFN for Subtask B
ID	unique hasoc ID for each tweet for Hindi data set

Table 2

Class division of both subtasks for Train and Test Dataset

Subtasks	No. of posts	Train set		Test set	
		English	Hindi	English	Hindi
Subtask A	HOF	2501	1433	798	1027
	NOT	1342	3161	483	505
Subtask B	PRFN	1196	213	224	74
	HATE	683	566	379	215
	OFFN	622	654	95	216
	NONE	1342	3161	483	1027
Total		3843	4594	1281	1532

**Figure 1:** Overall Pipeline**Figure 2:** Language Model

The authors used large-language models since the models are trained on a large amount of data and thus can understand the semantic structure of sentences and the tokens that are sent as inputs to these models have a contextual embedding associated with them. The output of the model is taken and then pooled. The resulting output is then passed through a linear layer and

a argmax is used to find the expected class of the sentence as shown in Figure. 1.

4. Results

The authors submitted four groups of results Table 3 gives the final results for our submission. The results has been evaluated on a test dataset, which is about one-third of the training data size, using the Macro F1 scores.

Table 3

Results from the official Test set from the leaderboard published from 15% the data set

Task	Our Score (Macro average F1)	Best Score	Rank
English Subtask A	0.8089	0.8177	4
English Subtask B	0.6396	0.6657	6
Hindi Subtask A	0.7379	0.7825	22
Hindi Subtask B	0.4431	0.5603	16

The experiments showed that large cased BERT performed the best followed by RoBERTa and the lowest scores were obtained from the BERT base model. The maximum sequence length that is used has a direct impact on the performance; with a larger length having a better performance, with the training time increases at the same time.

The methodology followed for both English and Hindi are the same, but the performance obtained for the English subtask is quite better than that for the Hindi subtask. This shows that the language models are pretty good in understanding the semantics for English but fail to do so for a low resource language like Hindi. The modified cross-entropy loss provided a better F1 score as compared to training with equal importance given to all of the separate classes.

5. Experimental Details

For English language we experimented with RoBERTa base pre-trained model [8], fine tuned BERT large cased architecture[7], and XLNet [9]- all for the same configuration, i.e, max length is set to 120, batch size to 8 and trained with 4 number of epochs. AdamW optimizer [29] with an initial learning rate of $2e-5$ is used for training. Similarly for hindi language tasks we used a pre-trained model [10] from the Hugging face [30] library. The Max length has been set to 200, batch size was 8, and number of epochs was set to 4.

There is a trade-off between the accuracy and the total number of tokens. The amount of time the model takes for training is proportional to the square of the number of tokens. As the number of tokens increases, the amount of time increases. However, when we truncate the maximum length, some of the information present in the sentence gets lost and the prediction for the sentence might be wrong. We had to consider a trade-poff between the accuracy and the time it takes for the model to train. For deciding the maximum sentence length, about 99% percentile of number of tokens in sentences is considered. For generating predictions we made a split of 90 % for training and 10 % validation to compare the performance of different models, for

each specific task and based on F1 scores of a particular epoch we updated the model weights. The weights corresponding to the best validation scores have been selected for inferring the test values. We observed that usually 3, 4 trained epochs had a higher F1 score. For reproducibility, the codes have been uploaded to github ¹. The random seed has been set to 42.

6. Conclusions

In this paper, we explain the shared tasks presented by HASOC in English and Indo-Aryan languages. We used large language models which are pre-trained on large corpora for hate speech detection tasks and to evaluate predictions by different models a validation dataset was created. In future work, we hope to try out more different fine tuned models.

Acknowledgments

The authors would like to thank the organizers of Hate Speech and Offensive Content Identification in Indo-Aryan Languages 2021 [5] for conducting this data challenge. The authors gratefully acknowledge google colab for providing GPU's to do the computation. All pre-trained models is based upon work supported by Hugging Face [30].

References

- [1] O. Istaiteh, R. Al-Omouh, S. Tedmori, Racist and Sexist Hate Speech Detection: Literature Review, 2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA) (2020) 95–99.
- [2] S. Kawate, K. Patil, Analysis of foul language usage in social media text conversation, *Int. J. Soc. Media Interact. Learn. Environ.* 5 (2017) 227–251.
- [3] S. Jaki, T. D. Smedt, M. Gwózdź, R. Panchal, A. Rossa, G. D. Pauw, Online hatred of women in the Incels.me forum 7 (2019) 240–268. URL: <https://doi.org/10.1075%2Fj1ac.00026.jak>. doi:10.1075/j1ac.00026.jak.
- [4] B. Mathew, R. Dutt, P. Goyal, A. Mukherjee, Spread of Hate Speech in Online Social Media, *Proceedings of the 10th ACM Conference on Web Science* (2019).
- [5] 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.
- [6] 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 Lan-

¹<https://github.com/priyanshusankhala/hasoc-hnlp>

- guages, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021. URL: <http://ceur-ws.org/>.
- [7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: NAACL, 2019.
- [8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, ArXiv abs/1907.11692 (2019).
- [9] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. V. Le, XLNet: Generalized Autoregressive Pretraining for Language Understanding, in: NeurIPS, 2019.
- [10] M. Bhangе, N. Kasliwal, Hinglishnlp: Fine-tuned language models for hinglish sentiment detection (2020).
- [11] S. Modha, T. Mandl, P. Majumder, D. Patel, Tracking Hate in Social Media: Evaluation, Challenges and Approaches, SN Comput. Sci. 1 (2020) 105.
- [12] M. A. Paz, J. Montero-Díaz, A. Moreno-Delgado, Hate speech: A systematized review, SAGE Open 10 (2020).
- [13] Z. Waseem, T. Davidson, D. Warmsley, I. Weber, Understanding abuse: A typology of abusive language detection subtasks, in: Proceedings of the First Workshop on Abusive Language Online, Association for Computational Linguistics, Vancouver, BC, Canada, 2017, pp. 78–84. URL: <https://aclanthology.org/W17-3012>. doi:10.18653/v1/W17-3012.
- [14] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, 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, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 1415–1420. URL: <https://aclanthology.org/N19-1144>. doi:10.18653/v1/N19-1144.
- [15] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on Twitter, in: Proceedings of the NAACL Student Research Workshop, Association for Computational Linguistics, San Diego, California, 2016, pp. 88–93. URL: <https://aclanthology.org/N16-2013>. doi:10.18653/v1/N16-2013.
- [16] S. Jaki, T. de Smedt, M. Gwózdź, R. Panchal, A. Rossa, G. D. Pauw, Online hatred of women in the Incels.me forum, Journal of Language Aggression and Conflict (2019).
- [17] M. Dadvar, D. Trieschnigg, R. Ordelman, F. de Jong, Improving cyberbullying detection with user context, in: P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, E. Yilmaz (Eds.), Advances in Information Retrieval, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 693–696.
- [18] R. Kumar, A. K. Ojha, M. Zampieri, S. Malmasi (Eds.), Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018), Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018. URL: <https://aclanthology.org/W18-4400>.
- [19] K. Dinakar, R. Reichart, H. Lieberman, Modeling the Detection of Textual Cyberbullying, in: The Social Mobile Web, 2011.
- [20] H. Zhong, H. Li, A. C. Squicciarini, S. M. Rajtmajer, C. Griffin, D. J. Miller, C. Caragea, Content-Driven Detection of Cyberbullying on the Instagram Social Network, in: IJCAI, 2016.
- [21] W. Yin, A. Zubiaga, Towards generalisable hate speech detection: a review on obstacles

- and solutions, PeerJ. Computer science 7 (2021) e598–e598.
- [22] J.-M. Xu, K.-S. Jun, X. Zhu, A. Bellmore, Learning from bullying traces in social media, in: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Montréal, Canada, 2012, pp. 656–666. URL: <https://aclanthology.org/N12-1084>.
 - [23] T. Davidson, D. Warmsley, M. W. Macy, I. Weber, Automated Hate Speech Detection and the Problem of Offensive Language, in: ICWSM, 2017.
 - [24] Y. Mehdad, J. Tetreault, Do characters abuse more than words?, in: Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Association for Computational Linguistics, Los Angeles, 2016, pp. 299–303. URL: <https://aclanthology.org/W16-3638>. doi:10.18653/v1/W16-3638.
 - [25] J. S. Meyer, B. Gambäck, A platform agnostic dual-strand hate speech detector, in: Proceedings of the Third Workshop on Abusive Language Online, Association for Computational Linguistics, Florence, Italy, 2019, pp. 146–156. URL: <https://aclanthology.org/W19-3516>. doi:10.18653/v1/W19-3516.
 - [26] S. D. Swamy, A. Jamatia, B. Gambäck, Studying generalisability across abusive language detection datasets, in: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 940–950. URL: <https://aclanthology.org/K19-1088>. doi:10.18653/v1/K19-1088.
 - [27] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandalia, A. Patel, Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages, Proceedings of the 11th Forum for Information Retrieval Evaluation (2019).
 - [28] 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, Proceedings of the 12th Forum for Information Retrieval Evaluation (2020).
 - [29] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, in: ICLR, 2019.
 - [30] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, HuggingFace’s Transformers: State-of-the-art Natural Language Processing, 2020. arXiv:1910.03771.