Fine-tune BERT to Classify Hate Speech in Hindi English Code-Mixed Text

Shikha Mundra[1*,] , Nikhil Singh [2] and Namita Mittal [3]

[1] *Malaviya National Institute of Technology, Jaipur, India*
[2] *Manipal University Jaipur, Jaipur, India*
[3] *Malaviya National Institute of Technology, Jaipur, India*

Abstract

With the exponential growth in internet technologies and social media usage, communicating and gathering information across countries is increasing at faster pace. These platforms provide opportunities to share opinions and suggestions about any socio-political events. Apart from seeking advantage, a person or community often misuses these platforms to post hate speech content. Content targeted towards casteism, racism, sexism, and insulting is called hate speech. The majority of the user are multilingual speakers knowing two or more languages. Out of them, English is the most known language. Knowing native language like Hindi and well-known languages like English, most social media users switch between Hindi and English while writing content on the social media platform. The phenomenon of mixing two languages like Hindi and English is called as Hindi English Code-mixed language. We have participated in HASOC subtask2 to classify Hi-En code mixed conversation in two classes as Hate Offensive (HOF) and Non-hate (NONE). We have experimented with various methods as Multilingual Bert (MBert), finetuned Pretrained Bert-base-uncased, an ensemble of Bert-base-uncased, XLNet and transfer learning-based approaches. We have analyzed that finetuned Bert-base-uncased has outperformed all other models. In subtask 2 of HASOC 2021, overall, 16 teams have participated, our team named “Rider” has achieved position 4 in Macro F1 score and Macro precision**.**

Keywords

Hate speech, Code Mixed Hindi English, Deep Learning, Embedding, BERT

1. Introduction

With the increasing usage of social media platform, hate speech and cyberbullying behavior has gained much attention. Some people often misuse these platforms to trigger religious, sexist, insulting, racial, abusive content against someone or community, thus, leads to depression, anxiety, and discouragement to continue the social media platform [10]. Anonymity or fake profile is one of the primary reasons that these platforms disseminate hate and offensive speech. Since these platforms are often used by multilingual people hence probability of the presence of text in more than one language is high. Indians more often mixes Hindi and English languages while writing. This phenomenon generally occurs when we do not find suitable words or phrases to write; consequently, we use the native language Hindi to complete that sentence. Understanding Hi-En code mixed text is quite difficult for a machine as it includes nonstandard words (e.g- dikkat, hme, aukat), spelling error(e.g. oxgen), screaming words(e.g. goooood, wowwwwww),switch of Hindi and English words, irrelevant text like URLs, @xxxx. In Hi-En code mixed, most often, standard Hindi (Devanagari) is written in Romanized Hindi, which increases the ambiguity in the text. For e.g. aukat can be written as aoukat, aukaat, okaat. All these words have the same semantic meaning but have different spelling due to a lack of standard

Forum for Information Retrieval Evaluation, December 13–17, 2021, India
EMAIL: a.shikha1990@gmail.com (Shikha Mundra) ; nikhil3198@gmail.com (Nikhil Singh); nmittal.cse@mnit.ac.in (Namita Mittal)
ORCID: 0000-0002-5235-7015 (A. 1)

spelling. In such a case, Back transliteration to Devanagari is also challenging due to spelling variation issues. Hence, understanding code mixed text is a significant challenge and needs to be explored.

Table 1 shows some sample sentences of the training dataset of HASOC 2021-subtask 2 written in Hi-En code mixed text and consist of nonstandard words and phrases [11]. It is a conversational dataset, in other words, posts and user comments are represented as conversational sentences. These sentences need to be classified into Hate offensive (HOF) and None classes.

Table 1
A glimpse of HASOC 2021(subtask 2) training dataset

Conversational Sentences	Label
No HINDU is still going to attack Charlie Hebdo for this Pun, just not because we aren't fools like izlamic bigots. No one may love us but HINDUTVA Loves all.... @Aadarsh_VAJPAI woh sirf hijda hoga jo charlie hebdo ko support karega. #CharlieHebdo should be taught a lesson @broadcastmyview @Aadarsh_VAJPAI No bro... Charlie Hebdo se hme koi dikkat nhi h.....kyunki he made clear ki no god can produce oxgen it is we people who will help each other.	NONE
No HINDU is still going to attack Charlie Hebdo for this Pun, just not because we aren't fools like izlamic bigots. No one may love us but HINDUTVA Loves all.... @Aadarsh_VAJPAI Ye sab wahi west se aaya hai jo hamare dharm par sanatan k khilaf hamesha sajish karte hai. Tune dusre dharm ka mazaak udaane k liye yehi west ka saath diya jo hamari sankriti k dushman hai. https://t.co/v37HXZZOLf	NONE
No HINDU is still going to attack Charlie Hebdo for this Pun, just not because we aren't fools like izlamic bigots. No one may love us but HINDUTVA Loves all.... @Aadarsh_VAJPAI No, we will not attack Charlie Hebdo, instead we'll lynch few innocent muslims, and force them up to chant "Jay ShreeRam" by wear and tear.	HOF
No HINDU is still going to attack Charlie Hebdo for this Pun, just not because we aren't fools like izlamic bigots. No one may love us but HINDUTVA Loves all.... @Aadarsh_VAJPAI Because tumhari aukat hi nhi h	HOF

Apart from understanding the code-mixed text, classification of hate offensive content is also a challenging task. As shown in table 1, sentence 1 and sentence 3 share most of the words, yet, they belong to different classes as NONE and HOF, respectively. Since they share a similar pattern of writing, it is difficult for a machine to classify them. Hence, this problem is associated with two challenges: First is to understand Hi-En code mixed text, and second is to build efficient features which can recognize hate offensive content. Further, for text understanding and feature extraction, we experimented with multiple deep learning-based and transformer-based models to identify the best model.

2. Related Work

Text representation or text understanding is the crucial phase for a machine to understand the text. Recently, several algorithms emerged which are focused on text representation only. Some popular text representation algorithms are based on statistical methods, lexicon-based methods, and deep learning methods.

Lexicon Based Features
Most of the prior work[1][2] has extracted features from the lexicon. Publicly available lexicon like indosentiwordnet, wordnet has been used widely to extract features from the available lexicon.

However, according to [2], lexicon-based features are not effective for code mixed text due to their nonstandard writing. Also, lexicon-based features require handcrafted rules and do not consider context while creating features. Hence, these approaches are not significantly applicable for code mixed text.

Statistical feature extraction

Most of the prior work in text classification is based on statistical features like n-gram and term frequency-inverse document frequency (tf-idf). However, the statistical feature is not effective in Hi-En code mixed text due to the presence of nonstandard spelling variation. For e.g: aukat, okat, aaukat words have same meaning with spelling variation hence, the statistical feature would not be effective. The baseline model is based on tf-idf weighted n gram feature and classified using SVM (Support Vector Machine) as shown in table 5.

Deep Learning and transformer-based features extraction

Recently, deep learning features have shown a significant improvement in understanding text at sementic level. Deep learning features are dense in comparison to traditional features. Some popular architectures of deep learning algorithms are CNN (convolutional neural network) and LSTM (Long Short Term Memory). Transformer-based architecture like Bert-base-uncased[3], XLNet [5] are based on deep learning algorithms, but they require high computational resources.

3. Proposed methodology

It was found that most of the words present in the training dataset of HASOC 2021-subtask2 were in English. Hence, we have used transformer-based embedding for text representation known as BERT (Bidirectional Encoder Representations from Transformers). We have used base model of the pretrained BERT model known as bert-base-uncased[3]. It is a transformer model pretrained on a large corpus of English data in a self-supervised fashion. Pretrained BERT Base model uses 12 layers of transformers block with a hidden size of 768 and the number of self-attention heads as 12 and has around 110M trainable parameters as shown in figure 1. It is pretrained with two objectives:Masked language modeling (MLM) and Next Sentence Prediction (NSP) as explained below.

Masked Language Modeling (MLM): taking a sentence, the model randomly masks 15% of the words in the input, then run the entire masked sentence through the model and must predict the masked words

Next Sentence Prediction (NSP): the model concatenates two masked sentences as inputs during pretraining. The model then must predict if the two sentences were following each other or not. Hence, it can create embedding having contextual knowledge as well.

The proposed framework is divided into two stages: In stage 1, we used some preprocessing steps to remove irrelevant data. In the next stage, processed data is fed to finetuned Bert-base-uncased model to generate the embedding, and further, it is finetuned concerning the classification task.

- *Preprocessing Step*

  During preprocessing phase, we have removed URLs and numerics as they were irrelevant. We have replaced @xxxxxx with @user. We have replaced all characters to lowercase. All the screaming words were reduced in length. We did not remove any stop words because they were necessary for the sentence's semantics. We have not performed any transliteration.

- *Fine-tune BERT*

  We have finetuned the Pretrained Bert-base-uncased for the training dataset. To validate the model, we have splitted the training dataset into 80:20 as training and validation data. Further, we have used validation data to verify the performance of each experimental approach. The model was finetuned with a dropout layer with 30% as the probability of retention. The dropout layer was followed by a fully connected layer and a softmax layer to output the logits containing the probability of each input sequence belonging to either of the two classes. We set the maximum length as 196. Every sequence greater than 196 would get truncated to a length of 196, and all sequences less than 196 would get padded till they are 196 in length. We used a batch size of 32 and a learning rate of 2e-5 and used the Adam optimizer for optimizing the model during training. We trained the model for four epochs and used a Tesla P100 GPU with a memory of 16 GB for training it.

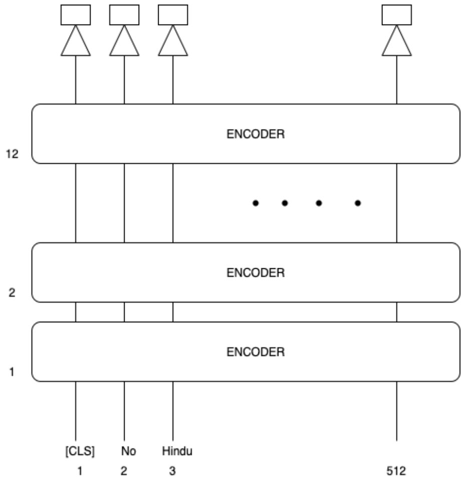

Figure 1: Bert

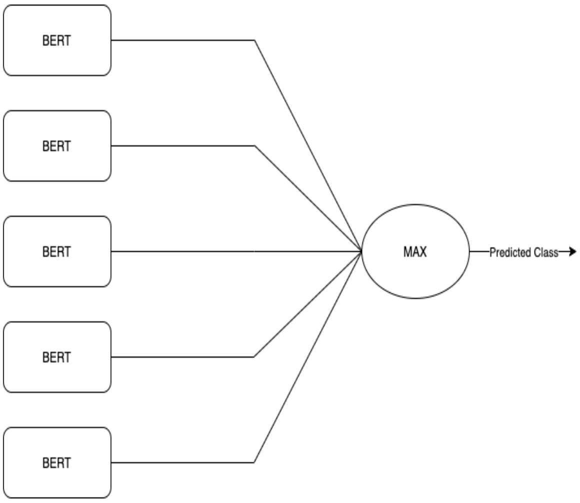

Figure 2: Ensemble Bert

4. Experimental Results and Error Analysis

We have experimented with and compared our proposed method with several other approaches: Multilingual Bert (Mbert), XLNet, Transfer of supervised features from a prominent English supervised dataset, and Ensemble Bert. To validate each method, we have divided the training set in 80:20 as the training and validation set. Further, we have applied the following approaches on the validation set and the test dataset.

Since code mixed language consists of Hindi and English, we have experimented with Pretrained MBert (Multilingual Bert). This model is trained in 104 languages in advance. [8] has finetuned MBert to detect hate speech in Hi-En code mixed text. However, it has the limitation of corrupting the input with masks, which ignores dependency between the masked positions and suffers from a pretrain-finetune discrepancy. XLNet is also a transformer approach that is proposed to overcome BERT's limitations [5]. Hence, to experiment with XLNet, we have finetuned it for our classification task.

Inspired from [6], We have investigated an approach related Transfer of supervised features (Transfer Learning): In this approach, we took advantage of the existing labeled dataset of Hate speech in English [4]. We trained (Davidson T et al.) hate speech dataset using English pretrained word2vec and CNN (convolutional neural network) in a supervised fashion and transferred the supervised feature to Hasoc 2021 training dataset. These transferred features are further finetuned for classification tasks using CNN. However, this approach has not performed well (as shown in table 3) because of the contamination of features.

Inspired by [7], we experimented with Ensemble Bert using the same hyperparameter as Bert. We created an ensemble with five random states, as shown in figure 2. Any five states have been randomly chosen between (1-500) as seed and finetuned with the classification task. At last, a classifier during the inference would take a vote of all five finetuned models in order to label a particular text sequence as a specific class. Further, we have discussed the validation and test results.

Table 2 shows validation accuracy on a 20 % stand-out validation dataset. Among all discussed approaches, Ensemble Bert-base-uncased has achieved the highest validation accuracy on the validation dataset.

Table 2

Evaluation on (20% standout) Validation dataset

Models	Validation Accuracy
MBert	70.56
XLNet	73.47
Transfer learning	72
Ensemble BERT	**75.95**
Fine tune BERT	75.26

We have applied each approach to the Test dataset, and we found that finetuned Bert-base-uncased outperformed other approaches (as shown in table 3).

Table 3

Comparative analysis using Test Dataset of Hasoc 2021(subtask 2)

Models	F1 score	Precision	Recall	Accuracy
MBert [9]	54.59	54.93	54.86	54.674
XLNet [5]	56.31	56.63	56.55	56.379
Transfer learning [6]	67.04	67.03	67.04	67.062
Ensemble BERT [7]	68.37	68.37	68.38	68.398
Fine tune BERT [3]	**68.90**	**69.12**	**69.05**	**68.917**

4.1. Error Analysis

As shown in table 4, our proposed methodology could detect hate speech to a certain extent. There is still room left for improvement. We have analyzed that the proposed approach has not handled Romanized English data. Also, the proposed approach is pretrained on English corpus. Hence, it will ignore Devanagari text.

Table 4

Classification of Test Dataset of Hasoc 2021(subtask 2)

		Actual Labels	Predicted Labels
1	#नीला_बकलोल_दिवस reservation bhik h https://t.co/b8U6qUy1Xc@HawkeyeAvenger1 Agar reservation bhik h to tum logo se bada es duniya me koi **bhikhari** nhi h😃😃😃	NONE	HOF
2	#नीला_बकलोल_दिवस reservation bhik h https://t.co/b8U6qUy1Xc@HawkeyeAvenger1 Agar reservation bhik h to tum logo se bada es duniya me koi bhikhari nhi h😃😃😃@GautamOmbeer 5000BC ki descendency h kshatriyaon ki **bhik** department belongs 2u reservation / aarakshan jivi	HOF	NONE
3	#नीला_बकलोल_दिवस reservation bhik h https://t.co/b8U6qUy1Xc@HawkeyeAvenger1 Je beem	HOF	HOF
4	#नीला_बकलोल_दिवस reservation bhik h https://t.co/b8U6qUy1Xc@HawkeyeAvenger1 Jiska meme bana rhe	HOF	HOF

	ho uske wajah se hi aaj tujh jaise **napunsak** yeh tweet karne ka freedom hai.		
5	#नीला_बकलोल_दिवस reservation bhik h https://t.co/b8U6qUy1Xc@HawkeyeAvenger1 Jiska meme bana rhe ho uske wajah se hi aaj tujh jaise **napunsak** yeh tweet karne ka freedom hai.@ToshanJharia Mind your language first no one is your **dad servant** here first of all I don't make meme n all or freedom of speech kiske constitution se liya hai batao zara	HOF	NONE
6	#नीला_बकलोल_दिवस reservation bhik h https://t.co/b8U6qUy1Xc@HawkeyeAvenger1 Baba sahab ko na kaho kuch .... Jo kuvh **bhadka** tha h vo h chandra sekhar azad	HOF	HOF
7	I am COVID positive and recovering. Your love and blessings sustain me. 🙏 https://t.co/HU0wz4xaH8@sudhirchaudhary For the first time, Sudhir Chaudhary shared any positive news. Stay safe Corona@ReJaulhaq6 @sudhirchaudhary तुम जैसे **घटिया** लोग ऐसे ही बोलेंगे	HOF	NONE
8	I am COVID positive and recovering. Your love and blessings sustain me. 🙏 https://t.co/HU0wz4xaH8@sudhirchaudhary #GodiMedia modi ji miss you only....@Kumar10Rahul @sudhirchaudhary अबे चुप कर **कुतरे**	NONE	HOF
9	Democracy doesn’t have “Kings” ...Dynasties have Kings .. While Democracy has "Karyakartas” you all will always continue to be “Durbaris” of the Past Kings &amp; Future Clown Princes...@sambitswaraj Oooo **ye pappu** banega Prince Princess nhi banega	NONE	HOF
10	Democracy doesn’t have “Kings” ...Dynasties have Kings .. While Democracy has "Karyakartas” you all will always continue to be “Durbaris” of the Past Kings &amp; Future Clown Princes...@sambitswaraj **Nalla** doctor	NONE	HOF
11	Democracy doesn’t have “Kings” ...Dynasties have Kings .. While Democracy has "Karyakartas” you all will always continue to be “Durbaris” of the Past Kings &amp; Future Clown Princes...@sambitswaraj इस वीडियो में एक दरबारी है जिसने सीधा टोंटी में ही मुहँ लगा दिया है।आप भी देखिए। https://t.co/TvHGnmqQVX	HOF	HOF
12	Democracy doesn’t have “Kings” ...Dynasties have Kings .. While Democracy has "Karyakartas” you all will always continue to be “Durbaris” of the Past Kings &amp; Future Clown Princes...@sambitswaraj इस वीडियो में एक दरबारी है जिसने सीधा टोंटी में ही मुहँ लगा दिया है।आप भी देखिए। https://t.co/TvHGnmqQVX@rahuldbarman @sambitswaraj ये तो राहुल बाबा की तस्वीर है	NONE	HOF

Table 4 shows the actual label and predicted label. After comparing ground truth against predicted label, we have noticed following points:

-Proposed approach has classified sentences 3,4,6,11 correctly. It is to note here that Devanagari script is not known (unk) for proposed approach.In(3,4,6,11) sentences,inclusion of Devanagari text has not modified the sementic meaning of text, hence these sentences are classified correctly.

-Proposed approach has not predicted sentences 1,2,5 correctly due to presence of Romanized English.

-Sentence 7 is predicted incorrectly as Devanagari text present in it is not recognized by the proposed approach. Therefore, after ignoring Devanagari text from sentence 7, sentence will not infer any Hate Hence, model predicted NONE label.

-In our opinion, Sentences 8,9,10,12 are labeled incorrectly. These sentences have targeted insulting words still, they have been assigned to NONE class.

4.2. Result of HASOC 2021(Leaderboard)

Initially, we were given a baseline model. In the Baseline model, preprocessed data using Hindi lemmatizer is fed in the first stage and the traditional tf-idf weighted n-gram method is used to extract features [11]. The baseline model classified these features using SVM and achieved 63.15% Macro F1 score and 63.16% Macro Precision as shown in table 5.

Overall, 16 teams have participated at the international level in HASOC 2021(subtask 2). Our team named 'Rider' has achieved the fourth position with 68.90% and 69.12% of Macro F1 score and Macro Precision, respectively. Our approach based on fine-tune Bert-base-uncased has significantly improved by 5.75 % in f1 score and 5.96 % in Macro precision against the baseline model.

Table 5
Rank in HASOC 2021 (subtask 2) in Macro F1 score

RANK	TEAM	MACRO F1	MACRO PRECISION
1	MIDAS-IIITD	0.7253	0.7267
2	Super Mario	0.7107	0.7117
3	PreCog IIIT Hyderabad	0.7038	0.7069
4	**Rider**	**0.6890**	**0.6912**
5	Hasnuhana	0.6866	0.6912
6	IRLab@IITBHU	0.6795	0.6805
7	r1_2021	0.6742	0.6761
8	TeamBD	0.6656	0.6658
9	Chandigarh_Concordia	0.6551	0.6578
10	PC1	0.6537	0.6672
11	MUM	0.6476	0.6483
12	**HASOC(Baseline)**	0.6315	0.6316

6. Conclusion

In Hasoc 2021, subtask 2, we experimented with MBert, XLNet, Supervised Transfer learning, finetune Bert-base-uncased, Ensemble Bert-base-uncased to classify hate offensive text from Hindi English code mixed social media text. Our validation and experimental result shows that Finetuned Bert-base-uncased and Ensemble Bert-base-uncased are better than other approaches. On comparing with baseline model, it is depicted that indepth preprocessing and statistical features are not enough to detect hate speech. Mbert did not perform well on validation and test data due to the presence of the majority of the text in English. In the future, we will enhance existing features which can include the Devanagari script too.

7. References

[1] Samghabadi NS, Mave D, Kar S, Solorio T. RiTual-uh at TRAC 2018 shared task: aggression identification. arXiv preprint arXiv:1807.11712. 2018 Jul 31.

[2] Sharma S, Srinivas PY, Balabantaray RC. Sentiment analysis of code-mix script. In2015 international conference on computing and network communications (CoCoNet) 2015 Dec 16 (pp. 530-534). IEEE.

[3] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pretraining of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.

[4] Davidson T, Warmsley D, Macy M, Weber I. Automated hate speech detection and the problem of offensive language. In Proceedings of the International AAAI Conference on Web and Social Media 2017 May 3 (Vol. 11, No. 1).

[5] Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems. 2019;32.

[6] Mathur P, Shah R, Sawhney R, Mahata D. Detecting offensive tweets in Hindi English code-switched language. in Proceedings of the Sixth International Workshop on Natural Language Processing for social media 2018 Jul (pp. 18-26).

[7] Risch J, Krestel R. Bagging BERT models for robust aggression identification. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying 2020 May (pp. 55-61).

[8] Malte A, Ratadiya P. Multilingual cyber abuse detection using advanced transformer architecture. In TENCON 2019-2019 IEEE Region 10 Conference (TENCON) 2019 Oct 17 (pp. 784-789). IEEE.

[9] Libovický J, Rosa R, Fraser A. How language-neutral is multilingual BERT?. arXiv preprint arXiv:1911.03310. 2019 Nov 8.

[10] S. Modha, T. Mandl, G. K. Shahi, H. Madhu, S. Satapara, T. Ranasinghe, and M. Zampieri,“Overview of the HASOC Subtrack at FIRE 2021: Hate Speech and Offensive Content Identifica-tion in English and Indo-Aryan Languages and Conversational Hate Speech,” inFIRE 2021: Forumfor Information Retrieval Evaluation, Virtual Event, 13th-17th December 2021, ACM, December2021

[11] S. Satapara, S. Modha, T. Mandl, H. Madhu, and P. Majumder, “ Overview of the HASOCSubtrack at FIRE 2021: Conversational Hate Speech Detection in Code-mixed language ,” inWorking Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021