MULTILINGUAL HATE SPEECH AND OFFENSIVE LANGUAGE DETECTION

Puspendu Biswas[1*], Donavalli Haritha[2]

Abstract

Hate speech and offensive language are phenomena that spread with the growing reputation of social media boards. Computerized detection of such content is vital for predicting conflicts among social communities and blocking off inappropriate content from social media boards. This paper aims to explain our group SSN_NLP_MLRG submission to HASOC 2021: Hate speech and offensive language detection in English and Indo-Aryan language, wherein we discover one of a kind models to carry out the subtask1 includes subtask A: To discover the remarks is Hate speech and offensive (HOF) or now not and subtask B: to categorize the HOF remarks into profanity (PRFN), Hate speech (HATE), Offensive (OFFN) in English, Hindi language and subtask A in Marathi language. The experiments cowl unique gaining knowledge of strategies that consist of gadget getting to know, transfer studying, and Multilingual pre-educated models. Our exceptional fashions are Roberta for English subtask A, BERT for English subtask B, and MBERT for the Hindi subtask A, Hindi subtask B, and Marathi subtask A. Our crew carried out the macro-averaged F1 scores of 0.7919, zero.7320, zero.8223, zero.6242, and 0.5110 within the English subtask A, Hindi subtask A, Marathi subtask A, English subtask B, and Hindi subtask B, respectively.

Keywords: Transfer learning, Code-Mixed language, Machine learning, Language modeling, Low-resource language

1*,2Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India

***Corrsponding Author:** Puspendu Biswas
*Department Of Computer Science And Engineering, Koneru Lakshmaiah Education, Foundation, Vaddeswaram, AP, India

DOI: 10.48047/ecb/2023.12.si10.00430

1. Introduction

Social media is a tremendous on-line verbal exchange discussion board that enables the public to explicit themselves without difficulty, at instances, anonymously. while expressing their opinion oneself is a proper of people that is cherished, inducing, and spreading offensive content in the direction of another social network is an abuse of this liberty [1, 2]. consequently, social media forums and other manner of online communication systems have began to play a bigger function in hate and offensive crimes. Many on line social media forums including Twitter, fb, Instagram, and YouTube don't forget hate speech and offensive content material harmful and feature the coverage to dispose of such content. because of societal subject and how tremendous offensive content material is becoming at the internet, there may be a strong motivation to discover hate speech and offensive content material in social media forums. Hate speech1 defines the assaults towards someone or organization network, primarily based on these attributes as race, gender, ethnicity, religion, sexual orientation, age, physical or intellectual incapacity, and others. Offensive content2 is a language that could significantly offend an man or woman or group based totally on their age, religious or political beliefs, marital or parental repute, sexual orientation, physical capabilities, country wide origin, or disability.

Hindi is an Indo-Aryan language with the legit languages of India and spoken mainly in North India. Marathi is an Indo-Aryan language spoken predominantly by using Marathi human beings of the Maharashtra nation in India and authentic and co-reputable language within the Maharashtra and Goa states of Western India. Code-combined language is a phenomenon that mixes one or greater languages and additionally native language written in roman script. The detection and categorization of hate speech and offensive language within the indirect feedback [3] of the code-blended are challenging responsibilities no longer only inside the English language. consequently, there is an open studies place in the subject of a code-combined multilingual community which include Hindi, Marathi languages, etc.

The HASOC 2019 [4] and 2020 competitions intention to teach the structures capable of detecting hate speech and offensive content in social media forums for the English, Hindi, and German languages. within the HASOC 20193, the organizers offered sub-responsibilities A, B, C for English, Hindi, and sub-responsibilities A, B for German languages. furthermore, whether or not the content is not or HOF (sub-task A), what are the traits of the HOF content (sub-undertaking B), and who's the target of the HOF message (sub-project C). in the HASOC 20204, they organized the two duties for English, German, and Hindi Languages, particularly subtask A: To become aware of the given comment is HOF or now not and subtask B: To categorize the HOF remark into offensive, hate, and profanity.

This paper affords our techniques to HASOC-2021. we have participated in subtask1 shared undertaking such as English subtask A, English subtask B, Hindi subtask A, Hindi subtask B, and Marathi subtask A. The goal of subtask A is to become aware of and locate the social media remarks are hate speech and offensive (HOF) or no longer. The subtask B targets to categorize the traits of the HOF content material into hate, offensive, and Profanity. We used the system learning algorithms, BERT, MBERT, ALBERT, RoBERTa, DistilBERT model with ktrain library, ULMFiT to adapt and first-class-song the machine. We used the NLTK library for pre-processing the education set and trying out set for all three languages. The paper outlines as follows. The survey of relevant works describes in phase 2. The info of the experiment information and approach of our fashions are in Sections 3 and 4. segment five describes the evaluation a part of the test results. eventually, section 6 suggests the concluded work and discusses in addition paintings.

2. Related work

The OffensEval 2019 [5] is the shared challenge to become aware of the offensive content inside the English language. most of the teams hired BERT with unique parameters to detect offensive language. inside the SemEval 2019 shared task, the researchers used gadget getting to know procedures, bi-directional LSTM fashions to classify offensive language [6]. by and large, the researchers have tailored and best-tuned the BERT, GPT, and ULMFiT models for detecting offensive language in the shared mission. The OffensEval 2020 [7, 8] is the Shared assignment on Multilingual Offensive

Language identity for the English, Greek, Danish, Turkish, Arabic languages. maximum groups used con-textualized Transformers, ELMo embeddings, BERT, RoBERTa, and the multilingual mBERT to come across and categorize the offensive language in 5 distinct languages.

For the low-resource language, the authors used the move-lingual data augmentation tech- nique for the enter context [9]. within the Gemeval2018 shared mission [10], They obtained the top of the line answer by the use of the most entropy meta-degree classifier version to discover the micro-posts of offensive language in the German language. inside the HatEval [11], the top team used the SVM model to stumble on the Twitter feedback is hate speech towards ladies and immigrants in multilingual language. the general public of groups used deep getting to know techniques in that LSTM model to locate hate speech and offensive language inside the shared challenge of HASOC 2019.

in the HASOC 2020, the high-quality-done groups have used the variations of the BERT trans-formers model [12, 13] to pick out and categorize the dislike speech and offensive language within the English, German, and Hindi (code-blended) languages. From the remark, most of the researchers have been used device learning techniques, deep mastering techniques, variation of pre-educated transformer models to stumble on hate speech and offensive language. we are nevertheless coping with the problem of spotting offensive content in low-aid languages, as well as the hassle of handling an imbalanced dataset in diverse code-mixed languages. This problem opens new research in exceptional low-resource languages other than English. HASOC 20215 shared undertaking organizers provide the resource for the English, Hindi (code-blended), and Marathi languages [14].

3. Experiment data

This section presents the task description, data pre-processing techniques in the shared task HASOC 2021.

3.1. Data description

The organizers offer two shared tasks, namely subtask1 and subtask2. In subtask1, they supplied the HASOC 2021 dataset of the English, Hindi (code-mixed), and Marathi languages. In subtask2, the organizers furnished the Hindi code-combined conversation dataset. Our group SSN_NLP_ MLRG participated within the subtask1 for all 3 languages. desk 1 provides the annotated tweets for the English, Marathi, Hindi HASOC 2021 dataset. For education and checking out the gadget, the English dataset has 3832 and 1281 posts. The Hindi code-mixed dataset incorporates 4594 posts, 1532 posts for training and checking out the machine. The Marathi code-combined dataset consists of 1863 posts for the train device and 625 feedback for testing the model machine. desk 2 suggests the statistics of the dataset for all three languages.

3.2. Task description

The shared task of HASOC 2021 is to identify the hate speech and offensive content in English and Indo-Aryan languages [15, 16]. The English language includes two tasks, namely subtasks

Table 1 Sample annotated comments - HASOC2021

Comments	Label/Subtask A	Label/Subtask B
found the little bastard now the fun begins	**HOF**	**PRFN**
my first time seeing report about this is very heart breaking	NOT	NONE
technically that is still turning back the clock dick head	**HOF**	**OFFN**
india has got the worst finance minister and health minister ever	**HOF**	**HATE**

Table 2 Train dataset of HASOC 2021

Language	Label/Subtask A	No of Comments	Label/Subtask B	No of Comments	Total
English	HOF	2491	HATE	680	**3832**
English			OFFN	619	
English			PRFN	1192	
English	NOT	1341	NONE	1341	
Hindi	HOF	1433	HATE	566	**4594**
Hindi			OFFN	654	
Hindi			PRFN	213	
Hindi	NOT	3161	NONE	3161	
Marathi	HOF	663	-	-	**1863**
Marathi	NOT	1200	-	-	

A and B as same as Hindi language. The Marathi language offers one task, namely subtask A. Subtask A is a binary text classification task that focuses the systems able to classify the given social media comments into two classes, namely, HOF and NOT.

Hate and Offensive content (HOF): The social

media comments contain harassment, profane, insults, threatening words.

Non-Hate and Offensive (NOT): The social media comments do not include hate and offensive content.
Subtask B is a multi-text classification that focuses the systems able to classify the given online comments into three classes, namely HATE, OFFN, PRFN.

Hate speech (HATE): The social media comments which contain hate words. **Offensive (OFFN):** The social media posts which contain offensive content.

Profane (PRFN): The social media posts contain profanity words.

3.3. Data pre-processing

The data pre-processing is to clean the social media comments from the unnecessary noisy content is present in the given dataset and transform it into a coherent form, which can be portable for English, Hindi code-mixed, and Marathi languages. We used the NLTK[6] for data cleaning, data duplication from the HASOC 2021 dataset [17]. First, we remove @ symbol with a string denoted as user-id because it does not have any meaningful expressions. Next, we remove the hashtag with a text as the user's name because it affects the performance of our model. The example of data pre-processing like "@AjeebBharti @BeraJaykrishna @khan_nainam @Policy @Twitter Prove it !!! What evidence you have bloody hell prove kar" and after that "prove it what evidence you have bloody hell prove kar". After that, we removed the punctuation, numerals, symbols, URLs, and emojis and then converted the upper case text into small case text. Finally, we replaced the misspelling offense words and string with * into appropriate matched words presented in the collected vocabulary words.

Table 3 Validation accuracy of the BERT-based and language models

Model	English A	English B	Hindi A	Hindi B	Marathi A
BERT	0.84	0.66	-	-	-
ALBERT	0.81	0.66	-	-	-
MBERT	-	-	0.79	0.69	0.88
DistilBERT	0.82	-	-	-	-
ULMFiT	0.75	-	-	-	-
RoBERTa	0.83	-	-	-	-

4. Methodology

This section presents the experimental analysis of the various methods used for the validation process.

4.1. BERT-based model and Language Model

we have experimented with various pre-trained fashions, particularly BERT uncased, DistilBERT base-uncased, ALBERT (Albert-base-v2), RoBERTa base, ULMFiT language modeling, and gadget studying strategies for subtask A of English language. We used the BERT, ALBERT pre-educated fashions and system learning strategies for subtask B of the English language. For the Hindi language subtask A and B, We used the multi-cased BERT transformers to evolve and quality-track the system to classify the detest speech and offensive content material from the given dataset. For the Marathi language subtask A, we used the multi-cased BERT (MBERT) for the binary text type venture. For the validation manner of the gadget, we take 25% of the records from the training dataset for the 3 languages. We used the above-cited pre-skilled fashions with the ktrain7 library that is beneficial to construct the machine using machine mastering, neural network, and deep studying strategies. we've analyzed the training machine to set the numerous batch length to six, 32 and gaining knowledge of prices as 2e-5, 3e-5, and the epochs to six, 7, 9, and 10. We used the ULMFiT[18] framework in that common-SGD Weight-Dropped LSTM (AWD-LSTM) structure version to predict the dislike speech and offensive content and their characteristics for the English language dataset. table three presents the validation results for the BERT-primarily based models of the three languages.

4.2. Machine Learning Techniques

For the machine learning techniques, we have conducted two experiments.

4.2.1. Experiment 1

in the first test, we have used the subsequent fashions, particularly aid vector machine classifier (SVM), Naive Bayes classifier (NB), random wooded area classifier (RF), and severe gradient boosting ensemble classifier (XGB), and used to are expecting the hate speech and offensive

content within the given English dataset. We used the sci-kit research library for the implementation of the gadget gaining knowledge of classifiers. For the use of time period frequency-inverse file frequency (TF-IDF) vectorization, we extracted the Ngram, man or woman level, word-stage features from the given dataset. For the usage of the sklearn CountVectorizer, we build vocabulary for regarded phrases and additionally tokenize the gathered facts. FastText is the pre-educated vector for 157 languages trained on common crawl and Wikipedia. We used the FastText pre-trained word embedding vectors for the English language, specifically Wikipedia Tamil vectors (wiki.ta.vec).

4.2.2. Experiment 2

inside the 2d experiment, we used the Gensim library for vector embeddings. Gensim is the quickest library for schooling the device of vector embedding. We test with the logistic regression, Multinomial Naive Bayes (NB), Random forest, and Linear guide vector system (SVC) models to are expecting the gadget. we've got utilized the genism for pre-processing and lemmatized the schooling dataset for this experiment 2. we have applied the phrase cloud for categorised the training dataset. we have used a file to vector transformer (Doc2vec) and text to TFIDF transformer for extracted the functions with the aid of the use of the genism version. desk four affords the validation outcomes for the gadget learning fashions of the English language.

in the end, we've used the MBERT model to predict the dislike speech and offensive content material and got a macro F1-score of 0.8223 with the epochs to ten and the learning price as 2e-5 for the Marathi subtask A. We got macro F1-rankings of 0.7320, zero.511 of the Hindi subtask A, Hindi subtask B with the epochs to ten, and the studying rate as 2e-5 for the MBERT model. For English subtask A, We were given a macro F1 rating of 0.7919 for the RoBERTa model with the 07 epochs and the gaining knowledge of charge as 3e-5. We were given a macro F1 score of zero.624 with the 09 epochs and the learning fee as 2e-five for the BERT version of the English subtask B.

5. Result Analysis

This phase provides the assessment of our model. for instance, we used the evaluation metrics like precision, take into account, macro-averaged F1-rating. we've got submitted our exceptional model after compar- ing the overall performance of our strategies for English, Marathi, and Hindi code-combined languages. The HASOC 2021 organizers supplied the check information for English subtask A, English subtask B, Hindi subtask A, Hindi subtask B, and Marathi subtask A. From the performance of the validation gadget, the RoBERTa model completed an accuracy of zero.eighty three and Precision, consider, and macro F1-score of 0.eighty one, 0.eighty, and 0.eighty, when compared with the overall performance of the opposite system

Table 4 Validation accuracy of the machine learning models

Model	English A	English B
Experiment 1:		
NB Count vector	0.75	-
NB WordLevel TF-IDF	0.71	-
NB N-gram	0.69	-
NB CharLevel	0.69	-
SVM N-gram	0.68	-
RF, Count vector	0.74	-
RF, Word level	0.75	-
XGB, Count vector	0.72	-
XGB, Word level	0.72	-
XGB, CharLevel	0.73	-
Experiment 2:		
LR doc2vec	0.67	0.36
RF doc2vec	0.64	0.36
XGB doc2vec	0.64	0.35
NB doc2vec	0.48	0.21
SVM doc2vec	0.65	0.36
LR TFIDF	0.74	0.65
RF TFIDF	0.75	0.62
XGB TFIDF	0.69	0.62
NB TFIDF	0.70	0.58
SVM TFIDF	0.73	0.59

learning approaches and pre-trained language models. The F1-score for the Not-offensive comments and hate speech offensive comments for the RoBERTa model is 0.74, 0.87 respectively. Therefore, the RoBERTa model performs well than other models for the English subtask A. The accuracy of the English subtask B is 0.66, and the F1-score for the OFFN, HATE, PRFN, and NONE comments for the BERT model are 0.54, 0.73, 0.41, and 0.77.

From the observation, the BERT model performs well than other approaches of machine learning techniques for the English subtask B. For the Hindi language, the MBERT model achieved an accuracy of 0.79 for subtask A, 0.69 for subtask B, and F1-score of HOF and NOT comments are 0.65, 0.86 for the subtask A, and the F1-score for the OFFN, HATE, PRFN, and NONE comments for the subtask B task are 0.83, 0.19, 0.40 and 0.43 respectively. For Marathi Language, MBERT achieved an accuracy of 0.88, the macro F1-score is 0.87, and the F1-score of HOF and NOT comments are 0.91 and 0.83. We adapted and fine-tuned the BERT-based model to build and predict the data and its characteristics for all the languages. Our team submitted three runs for the English subtask A and one runs for the subtasks of the other languages.

The final results[8] of our team for the three languages are present in Table 5. Our team SSN_NLP_MLRG submission got the 19*th*, 12*th*, 25*th*, 9*th*, 19*th*rank in the shared task for English subtask A, English subtask B, Hindi subtask A, Hindi subtask B, Marathi subtask A respectively.

Table 5 Final results for the three languages

Language	Model	Accuracy	Precision	Recall	Macro F1
English A	RoBERTa	0.80	0.80	0.78	**0.79**
English A	ALBERT	0.79	0.78	0.77	0.77
English A	BERT	0.80	0.80	0.77	0.78
English B	ALBERT	0.65	0.61	0.60	0.60
English B	BERT	0.66	0.62	0.62	**0.62**
Hindi A	MBERT	0.77	0.74	0.72	**0.73**
Hindi B	MBERT	0.70	0.49	0.53	**0.51**
Marathi A	MBERT	0.84	0.82	0.82	**0.82**

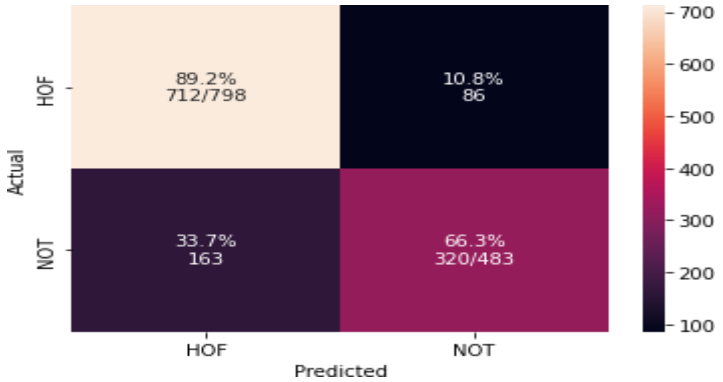

Figure 1: Confusion matrix of English subtask A - RoBERTa

We classify the performance of the model for all the three languages by using the confusion matrix are presented in the Figure 1 for English subtask A for the RoBERTa model, Figure 2 for the English subtask B for the BERT model, Figure 3 for the Hindi subtask A for the MBERT model, Figure 4 for the Hindi subtask B for the MBERT model, Figure 5 for the Marathi subtask A for the MBERT model. From the confusion matrix, we noticed that many test cases were classified as HOF comments by the RoBERTa model for the English subtask A. For English A, the Precision, Recall, and F1-score for the HOF and NOT comments are 0.81, 0.89, 0.85, and 0.79, 0.66, 0.62 respectively. For English B, the F1-score for the OFFN, HATE, PRFN, and NONE comments are 0.72, 0.46, 0.75, and 0.56. For Hindi A, the Precision, Recall, and F1-score for the HOF and NOT comments are 0.81, 0.87, 0.84, and 0.69, 0.58, 0.63 respectively. For Hindi B, the F1-score for the OFFN, HATE, PRFN, and NONE comments are 0.83, 0.40, 0.50, and 0.31. For Marathi A, the Precision, Recall, and F1-score for the HOF and NOT comments are 0.89, 0.88, 0.88, and 0.75, 0.77, 0.76 respectively. Overall, the hate speech and offensive comments perform well by Bert-based models for all three languages.

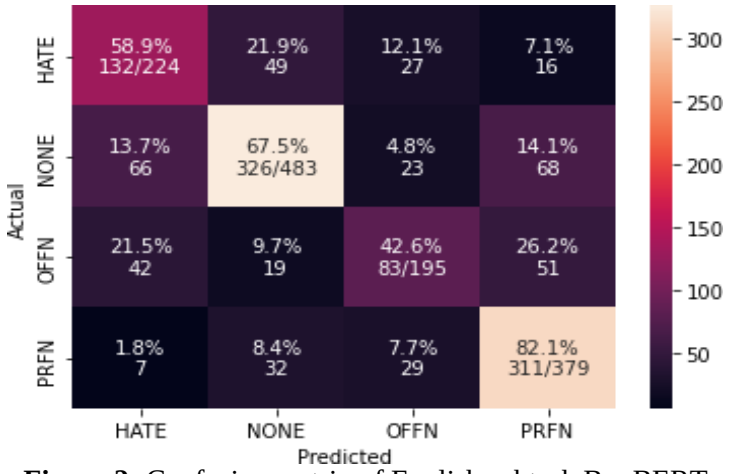

Figure 2: Confusion matrix of English subtask B – BERT

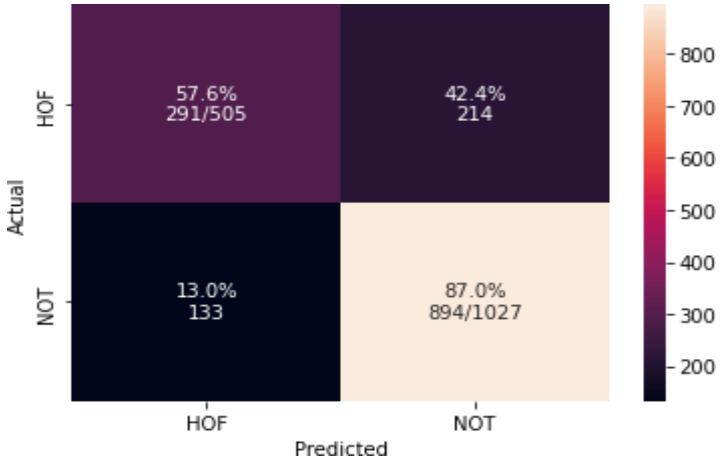

Figure 3: Confusion matrix of Hindi subtask A - MBERT

6. Conclusion

This paper presents the team submitted runs for the hate speech and offensive language identification for the HASOC 2021 subtask1 shared task in the Forum for Information Retrieval Evaluation (FIRE) 2021. We experimented with different approaches such as machine learning techniques, pre-trained BERT-based models. The results show the RoBERTa models perform well than the other BERT-based models and the machine learning approaches for the English subtask A. The BERT uncased performs well in the English subtask B. MBERT performs well in Hindi subtask A, Hindi subtask B, and Marathi subtask A. Based on the evaluation, the Overall BERT-based model performs well for the three languages. Our team submission had a macro F1-score of 0.8223, for the Marathi subtask A, macro F1-score of 0.7320, 0.511 for the Hindi subtask A and Hindi subtask B code-mixed language, and macro F1-score of 0.7919, 0.624 for the

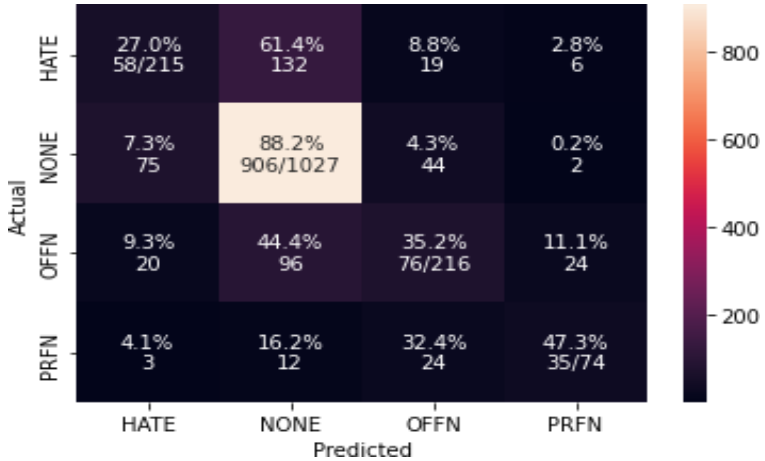

Figure 4: Confusion matrix of Hindi subtask B - MBERT

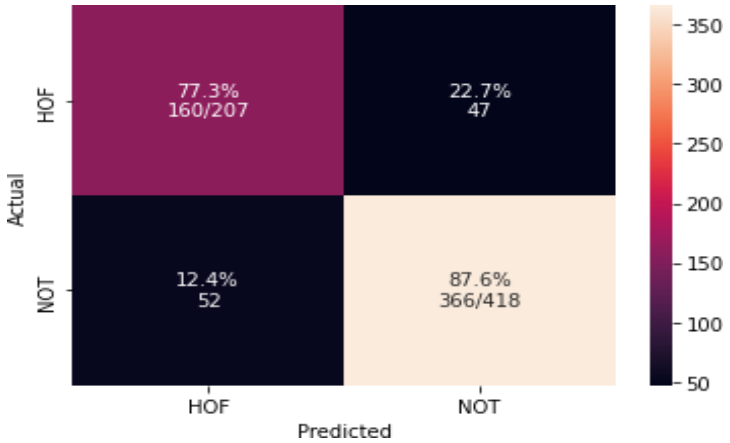

Figure 5: Confusion matrix of Marathi subtask A - MBERT

English subtask A and English subtask B. For future work, we will handle the sarcastic feature and imbalanced dataset to avoid misclassification and extend this work into other low-resource languages.

References

1. S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, O. Frieder, Hate speech detection: Challenges and solutions, PloS one 14 (2019) e0221152.
2. A. Kalaivani, D. Thenmozhi, Sentimental Analysis using Deep Learning Techniques, International Journal of Recent Technology and Engineering (IJRTE) 7 (2019) 600–606.
3. A. Kalaivani, D. Thenmozhi, Sarcasm Identification and Detection in Conversion Context using BERT, in: Proceedings of the Second Workshop on Figurative Language Processing,
4. Association for Computational Linguistics, Online, 2020, pp. 72–76. URL: https://www.aclweb.org/anthology/2020.figlang-1.10. doi:10.18653/v1/2020.figlang-1.10.
5. 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.
6. 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), 2019. arXiv:1903.08983.
7. D. Thenmozhi, B. Senthil Kumar, S. Sharavanan, A. Chandrabose, SSN_NLP at SemEval- 2019 Task 6: Offensive Language Identification in Social Media using Traditional and Deep Machine Learning Approaches, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 739–744. URL: https://www.aclweb.org/anthology/S19-2130. doi:10.18653/ v1/S19-2130.
8. M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczyn- ski, Z. Pitenis, Çağrı Çöltekin, SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020), 2020. arXiv:2006.07235.
9. A. Kalaivani, D. Thenmozhi, SSN_NLP_ MLRG at SemEval-2020 Task 12: Offensive Lan- guage Identification in English, Danish, Greek Using BERT and Machine Learning Approach, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 2161–2170. URL: https://aclanthology.org/2020.semeval-1.287.
10. J. Singh, B. McCann, N. S. Keskar, C. Xiong, R. Socher, XLDA: Cross-Lingual Data Augmentation for Natural Language Inference and Question Answering, CoRR abs/1905.11471 (2019). URL: http://arxiv.org/abs/1905.11471. arXiv:1905.11471.