Detection of Hate or Offensive Phrase using Magnified Tf-Idf

Palash Nandi, Dipankar Das

Jadavpur University, Kolkata, India

Abstract

The non-negotiable challenge that social media platforms are facing nowadays is the abundant presence of hate speeches in text messages. Thus, automatic hate speech detection becomes an important ethical concern and research should be carried out to overcome this challenge. In the present paper, we propose a tf-idf based binary classification framework that manipulates the scores obtained as the differences between hate and offensive (HOF) words and non-HOF (NOT) words. Employing this framework, we have achieved a Macro F1 score of 0.6813 and 0.6762 for the English and Hindi test datasets, respectively provided in subtask-1A of the HASOC 2021[13] shared task.

Keywords

hate speech, tf-idf, HOF, NOT, MagTIDS, NonMagTIDS, magnification factor, knowledge-base.

1. Introduction

Usage of HOF content is considered a major threat on online social media platforms. Saha et al. [1] presented that the users exposed to HOF react differently due to varying psychological endurance to hate exposure. Users with low mental endurance are more vulnerable to emotional instability than people with higher mental endurance. Mathew et al. [2] confirm that hateful content reaches farther, wider, faster, and has a greater impact, popularity than the content of non-hateful or neutral users. HOF can be a cause of individual to large-scale violence [2]. Therefore detection of HOF in social media platforms has become a priority.

In this paper, we represent a HOF detection framework on behalf of the subtask-1A of HASOC 2021 based on tf-idf and a manually created knowledge base of hate-words for English and Hindi.

2. Related Work

The evolution of research on HOF detection extends from keyword [3,4], distributional semantics based classifiers [5,6,7] to deep learning based classifiers [8,9,10]. Sood et al. [3] used a list of profane words, being able to identify 40% of words that are profane and then correctly identifying 52% as HOF or NOT. Mondal et al. [4] used sentence structures and a Hatebase[2] to identify hate targets. Nobata et al. [5] detected hate speech, profanity, and derogatory language in social media using n-grams as well as linguistic, syntactic, and distributional semantics. Djuric et al. [6] detected online hate using word embeddings from a neural network called Paragraph2vec to compare with the Bag of Words (BOW) model. Saleem et al. [7] used Labeled Latent Dirichlet Allocation (LLDA) to

[1]Forum for Information Retrieval Evaluation, December 13-17, 2021, India
EMAIL: sondhanil1@gmail.com (P. Nandi); dipankar.das@jadavpuruniversity.in (D. Das)
ORCID: 0000-0002-0775-0723 (P. Nandi)

[2]Structured repository of regionalized, multilingual hate speech: https ://hatebase.org/

automatically infer topics for the classifier. Park et al. [8] detected racist and sexist language through a two-step approach with convolutional neural networks. They used three CNN models (CharCNN, WordCNN, and HybridCNN) on 20K tweets, achieving the best performance with HybridCNN and the worst with CharCNN. Zhang et al. [9] used a pre-trained word embedding layer to map the text into vector space, which was then passed through a convolution layer with a max-pooling downsampling technique. Badjatiya et al. [10] classified the hatefulness of tweets using deep neural networks.

3. Task Description

Subtask-1A of HASOC 2021 strictly focuses on the binary classification of HOF and NOT classes. The definitions of HOF and NOT class are given below,

- NOT - A NOT statement does not contain any hate speech, profane, offensive content.
- HOF - A HOF statement contains hate, offensive, and profane content.

Given a Twitter post, subtask-1A expects the participating system to identify whether it is HOF or NOT. For example, the Twitter post '*@TheRealOJ32 Of all the retired NFL players, why is it that you DON'T suffer from CTE? You should be at the bottom of a pool you mistook for an elevator. #murderer'* is expected to be identified as HOF as a person or a group of people is targeted with hateful, offensive statements whereas the Twitter post '*Empty podiums make too much noise #ToryLeadershipDebate #UKPM #BorisJohnsonShouldNotBePM #Leadersdebate #GTTO #JC4PM2019 #frightnight https://t.co/aDgCqhdDTl'* should be labeled as NOT. The data for subtask-1A of HASOC 2021 is available for English, Hindi, and Marathi. We will use the English and Hindi dataset for our research work.

3.1. Data Analysis

In this section, we discuss the dataset used for creating the knowledge base, HOF_knowledge_base for both English and Hindi. We have used individual as well as the merged dataset from HASOC 2019, HASOC 2020 [12], and HASOC 2021 [11]. Table 1 represents information about datasets for both languages.

Table 1
Statistics of the dataset

Name of the dataset	HOF	NOT	TOTAL
English			
HASOC_EN_2019	2261	3591	5852
HASOC_EN_2020	288	865	1153
HASOC_EN_2021	2051	1342	3843
HASOC_EN_COMBINED	5050	5798	10848
Hindi			
HASOC_HI_2019	2469	2196	4665

HASOC_HI_2020	605	713	1318
HASOC_HI_2021	1494	3171	4665
HASOC_HI_COMBINED	4568	6080	10648

From the HOF posts of the combined dataset i.e. HASOC_EN_COMBINED, HASOC_HI_COMBINED, we have manually identified 277 and 174 offensive words for English and Hindi respectively.

4. Proposed Approach

4.1. Preprocessing

Since most of the time Twitter posts do not follow grammatically correct conventions, raw Twitter posts are not to be directly used for classification. Therefore we opted for a preprocessing pipeline to refine Twitter data. The steps in the preprocessing pipeline are explained below with help of an English and a Hindi Twitter post.

- ***Convert words into the lower case:*** HOF words are insensitive to letter cases. For that reason, each word of each sentence is turned into the lower case for the English dataset whereas it is inapplicable for the Hindi dataset except for the user mentions. For example, in English the Twitter post ‘@realDonaldTrump Technically that's still turning back the clock, you FatHead https://t.co/jbKaPJmpt1’ is turned into ‘‘@realdonaldtrump technically that's still turning back the clock, you fathead https://t.co/jbkapjmpt1’’ and in Hindi ‘@AskAnshul, आसमानी किताब के नाजायज औलाद है।’ is turned into ‘@askanshul, आसमानी किताब के नाजायज औलाद है।’

- ***Replace consecutive spaces with a single space:*** Twitter posts often contain multiple consecutive spaces. Those consecutive spaces are identified and replaced by a single space. For example, the English Twitter post ‘‘@realDonaldTrump technically that's still turning back the clock, you fathead https://t.co/jbkapjmpt1’’ is turned into ‘‘@realDonaldTrump technically that's still turning back the clock, you fathead https://t.co/jbkapjmpt1’’ and the Hindi Twitter post ‘@askanshul, आसमानी किताब के नाजायज औलाद है।’ is turned into ‘@askanshul, आसमानी किताब के नाजायज औलाद है।’

- ***Remove user mentions (by @):*** As our proposed methodology is sensitive towards HOF phrases only, the presence of any user mentions will not be helpful for the system. For that reason, any user mentions are removed. For example, the English Twitter post ‘‘@realDonaldTrump technically that's still turning back the clock, you fathead https://t.co/jbkapjmpt1’’ is turned into ‘‘technically that's still turning back the clock, you fathead https://t.co/jbkapjmpt1’’ and the Hindi Twitter post ‘@askanshul, आसमानी किताब के नाजायज औलाद है।’ is turned into ‘, आसमानी किताब के नाजायज औलाद है।’

- ***Replace emojis with corresponding text:*** Emojis, when used directly, are not useful in HOF detection as an individual emoji does not express any hate or offense. But when combined

with context, emojis can be expressive. For an instance ‘ ’ is neither hateful nor offensive content but ‘you piece of ’ is considered a derogatory comment. For that reason, all emojis present in the sentences are replaced with corresponding text. For example, the English Twitter post ‘‘technically that's still turning back the clock, you fathead https://t.co/jbkapjmpt1’’ is turned into ‘‘technically that's still turning back the clock, you fathead pile of poo https://t.co/jbkapjmpt1’’ and the Hindi Twitter post ‘, आसमानी किताब के नाजायज औलाद है।’ is turned into ‘,मल का ढेर आसमानी किताब के नाजायज औलाद है।’

- ***Remove URL:*** Often Twitter posts contain a link for supporting images or videos but as the proposed methodology only considers Twitter text for analysis, the present link is discarded. For example, the English Twitter post ‘‘technically that's still turning back the clock, you fathead pile of poo https://t.co/jbkapjmpt1’’ is turned into ‘‘technically that's still turning back the clock, you fathead pile of poo’’ and the Hindi Twitter post ‘, मल का ढेर आसमानी किताब के नाजायज औलाद है।’ remains unaltered as it does not contain any link.

- ***Expand contracted words:*** Contracted words are replaced with the equivalent phrase for better understanding. For example, the English Twitter post ‘‘technically that's still turning back the clock, you fathead pile of poo’’ is turned into ‘technically that is still turning back the clock, you fathead pile of poo’ and the Hindi Twitter post ‘, मल का ढेर आसमानी किताब के नाजायज औलाद है।’ remains unaltered as it does not contain any contracted word.

- ***Remove punctuation marks:*** Any punctuation marks present in the sentence are removed as punctuation marks are neither part of the hatebase nor contribute to the detection of HOF phrases. For example, the English Twitter post ‘‘technically that is still turning back the clock, you fathead pile of poo’’ is turned into ‘technically that is still turning back the clock you fathead pile of poo’ and the Hindi Twitter post ‘, मल का ढेर आसमानी किताब के नाजायज औलाद है।’ is turned into ‘मल का ढेर आसमानी किताब के नाजायज औलाद है’.

- ***Remove stop words:*** Any stop word present in the sentence is detected[3] and removed as stop words are neither part of the hatebase nor contribute to the detection of HOF phrases. For example, the English Twitter post ‘technically that is still turning back the clock you fathead pile of poo’ is turned into ‘technically still turning back clock fathead pile poo’ and the Hindi Twitter post ‘मल का ढेर आसमानी किताब के नाजायज औलाद है’ is turned into ‘मल ढेर आसमानी किताब नाजायज औलाद’.

Each word in the semi-processed sentence is lemmatized for the English dataset. For the Hindi dataset, lemmatization is not used due to the unavailability of a suitable lemmatizer.

4.2. Creation of MagTIDS

[3]Stopwords of English are collected using the NLTK toolkit available at https://www.nltk.org/ and stopwords of Hindi are collected from the open-source repository available at https://github.com/Alir3z4/stop-words/blob/master/hindi.txt.

MagTIDS contains magnified tf-idf difference scores between HOF and NOT classes for each word. To generate the MagTIDS score of a word word_i, initially, we have calculated the tf-idf score of word_i w.r.t class HOF and NOT from the training dataset. Then the magnified difference score is obtained by multiplying a selected magnification factor with the absolute difference between the tf-idf score of word_i w.r.t. class HOF and NOT. The required formula to calculate the MagTIDS score for word_i is given below.

*MagTIDS(word_i) = magnification_factor * | tf_idf[word_i][HOF]-tf_idf[word_i][NOT] |*

where,

tf_idf[word_i][HOF]: tf-idf score of word_i w.r.t. class HOF
tf_idf[word_i][NOT]: tf-idf score of word_i w.r.t. class NOT

4.3. Creation of NonMagTIDS

NonMagTIDS contains non-magnified differences of tf-idf scores between HOF and NOT classes for each word. To generate the NonMagTIDS score of a word word_i, the absolute difference between the tf-idf score of word_i w.r.t class HOF and NOT is taken. The required formula to calculate the NonMagTIDS score for word_i is given below,

NonMagTIDS(word_i) = | tf_idf[word_i][HOF]-tf_idf[word_i][NOT] |

where,

tf_idf[word_i][HOF]: tf-idf score of word_i w.r.t. class HOF
tf_idf[word_i][NOT]: tf-idf score of word_i w.r.t. class NOT

4.4. MagTIDS based Binary Classification

Only detecting the important words of respective classes using the tf-idf score is not sufficient for the classification task as many non-offensive words like 'people', 'india', 'significant', 'like', 'trade', 'face' hold top scores in HOF class. To build the classification model sensitive to HOF words, we created two parsing modules. Each of them returns a cumulative score after parsing the preprocessed input string. First, the module parse_HOF uses HOF_knowledge_base and MagTIDS to calculate the cumulative score when any offensive, hate, or profane word is encountered, and second, the module parse_NOT uses only NonMagTIDS scores to count the cumulative score over consecutive words. After parsing with both parse_HOF and parsed_NOT modules, a normalized distribution of cumulative scores is obtained. Finally, the class corresponding to the module with the highest score is considered as output.

4.4.1. Parsing with module parse_HOF for HOF phrases

Module parse_HOF is sensitive towards HOF words. It takes a preprocessed string as input and returns a score based on the presence of HOF words. To identify the HOF words, parse_HOF mainly uses HOF_knowledge_base and MagTIDS scores. For an input sentence, initially parse_HOF sets the cumulative score to 0 and iterates over each word of the received preprocessed text. It tries to sense if any HOF keyword is present in the current text token. To reduce the detection of false-positive HOF words, a text token ‘word_i’ is considered as HOF if and only if at least one recognized HOF keyword kw_i, from HOF_knowledge_base, is a substring of word_i and the absolute difference between the

length of word_i and kw_i, not more than two. For example, token ‘banged’ is considered as HOF text w.r.t. keyword ‘bang’ but ‘bangalore’ is not. Later, If the word is a HOF, the MagTIDS score corresponding to the matched keyword kw_i is added to the cumulative score, else the NonMagTIDS score of word_i is added. Figure 1 represents the algorithm of the parse_HOF module below.

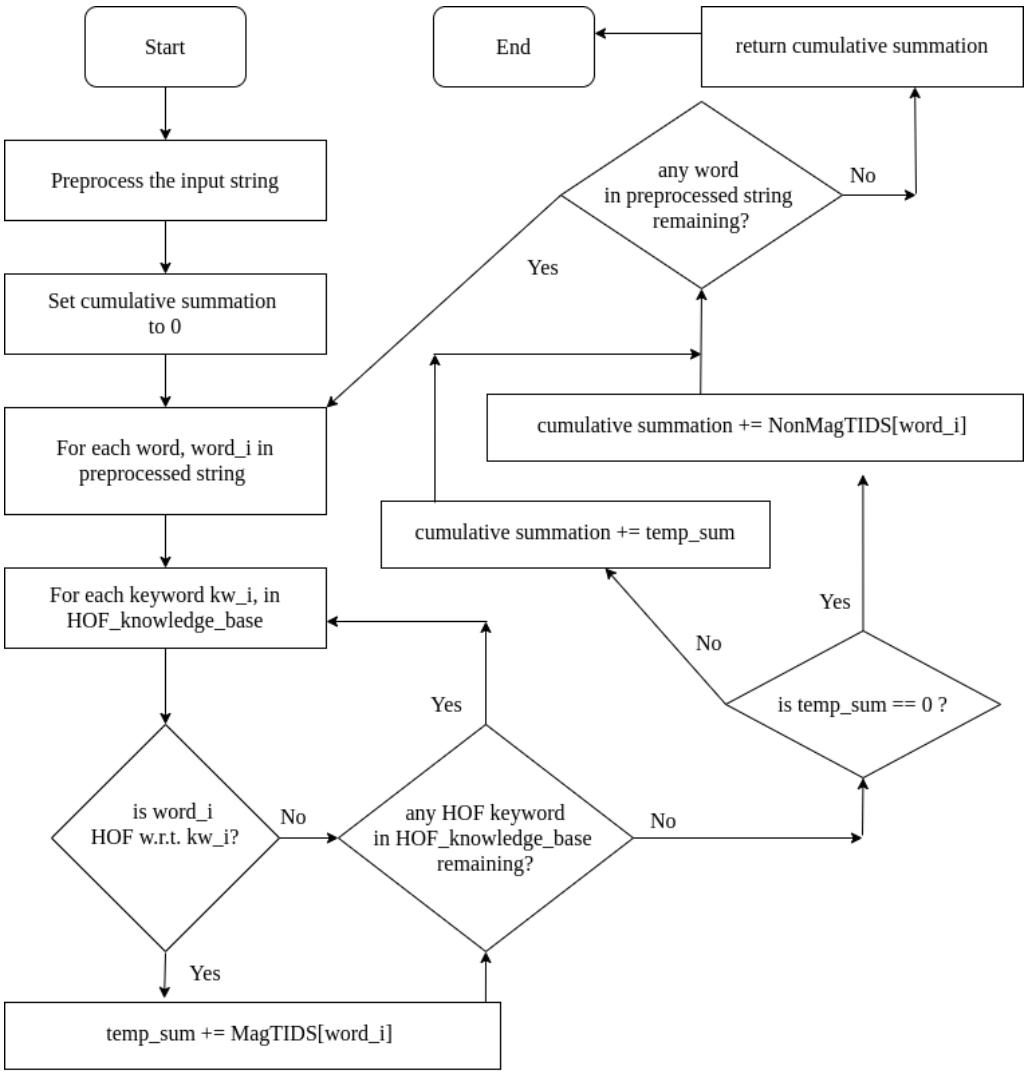

Figure 1: Algorithm of the parse_HOF module

4.4.2. Parsing with module parse_NOT

Module parse_NOT takes a preprocessed string as input and returns a score based on NonMagTIDS scores. It does not check sensitivity towards any HOF or NOT words. Initially parse_NOT sets the cumulative score to 0. Then iterates over each word of the received preprocessed text and increases the cumulative score by their NonMagTIDS scores. Figure 2 represents the algorithm of the parse_NOT module.

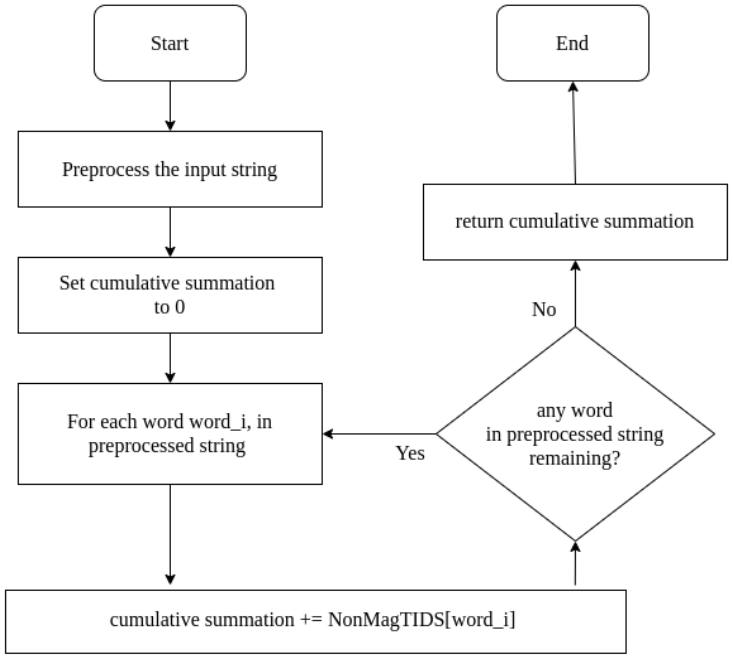

Figure 2: Algorithm of the parse_NOT module

5. Results

5.1. Results using the training dataset

The proposed classification model is applied to the training dataset for the English and the Hindi of subtask-1A of HASOC 2021 with the magnification factor from one to a thousand.

5.1.1. For the English dataset

Datasets HASOC_EN_2019, HASOC_EN_2020, HASOC_EN_2021, HASOC_EN_COMBINED are used to evaluate the proposed classification framework on the English training dataset. The evaluation scores on the English training datasets are represented in Table 2.

Table 2
Evaluation of the proposed binary classification model on the training data for English.

Name of the dataset	Macro F1	Macro Precision	Macro Recall	Accuracy	Magnifi-cation Factor
HASOC_EN_2019	0.6135	0.6442	0.6138	0.6643	36
HASOC_EN_2020	0.7459	0.7369	0.7585	0.7988	6
HASOC_EN_2021	0.6570	0.6547	0.6622	0.6802	75
HASOC_EN_COMBINED	0.5276	0.5287	0.5282	0.5333	5

The column ‘Magnification Factor’ in Table 2 indicates the optimal value for which the proposed classification model performs best. The performance of the model for each magnification factor (in the range of one to a thousand) on each dataset in English is represented in Figure 3.

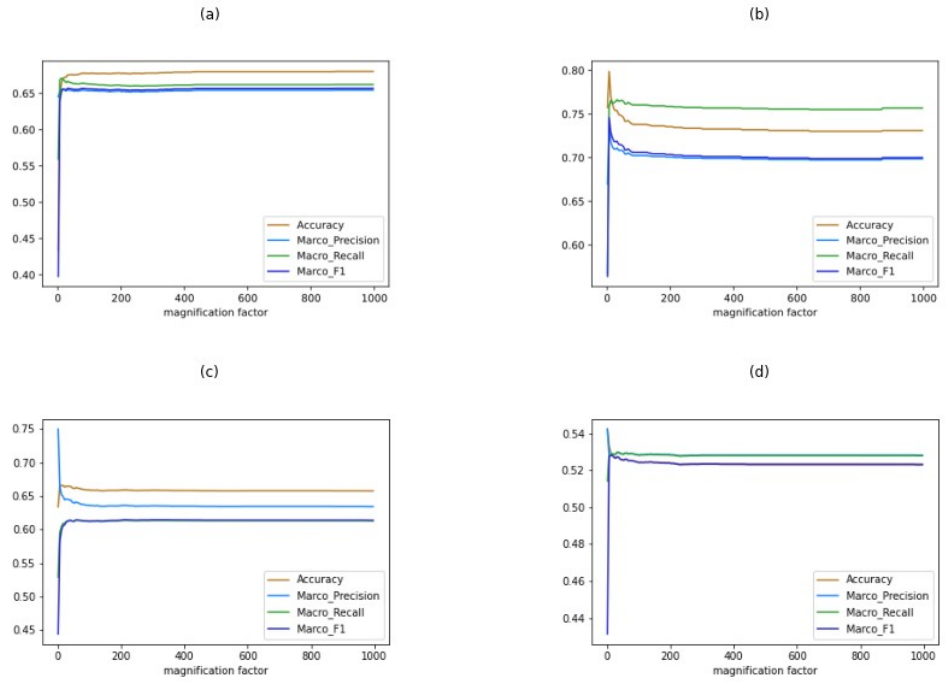

Figure 3: Performance of the proposed classification model for magnification value from 1 to 1000.
Figure 3 (a) indicates performance on dataset HASOC_EN_2021
Figure 3 (b) indicates performance on dataset HASOC_EN_2020
Figure 3 (c) indicates performance on dataset HASOC_EN_2019
Figure 3 (d) indicates performance on dataset HASOC_EN_COMBINED

5.1.2. For the Hindi dataset

Datasets HASOC_HI_2019, HASOC_HI_2020, HASOC_HI_2021, HASOC_HI_COMBINED are used to evaluate the proposed classification framework on the Hindi training dataset. The evaluation scores on the Hindi training dataset are represented in Table 3.

Table 3
Evaluation of the proposed binary classification model on the training data for Hindi.

Name of the dataset	Macro F1	Macro Precision	Macro Recall	Accuracy	Magnific-ation Factor
HASOC_HI_2019	0.775	0.801	0.789	0.776	95
HASOC_HI_2020	0.783	0.798	0.781	0.789	45
HASOC_HI_2021	0.650	0.780	0.6437	0.750	45
HASOC_HI_COMBINED	0.540	0.623	0.570	0.611	81

The column ‘Magnification Factor’ in Table 3 indicates the optimal value for which the proposed classification model performs best. The performance of the model for each magnification factor (in the range of one to a thousand) on each dataset in Hindi is represented in Figure 4.

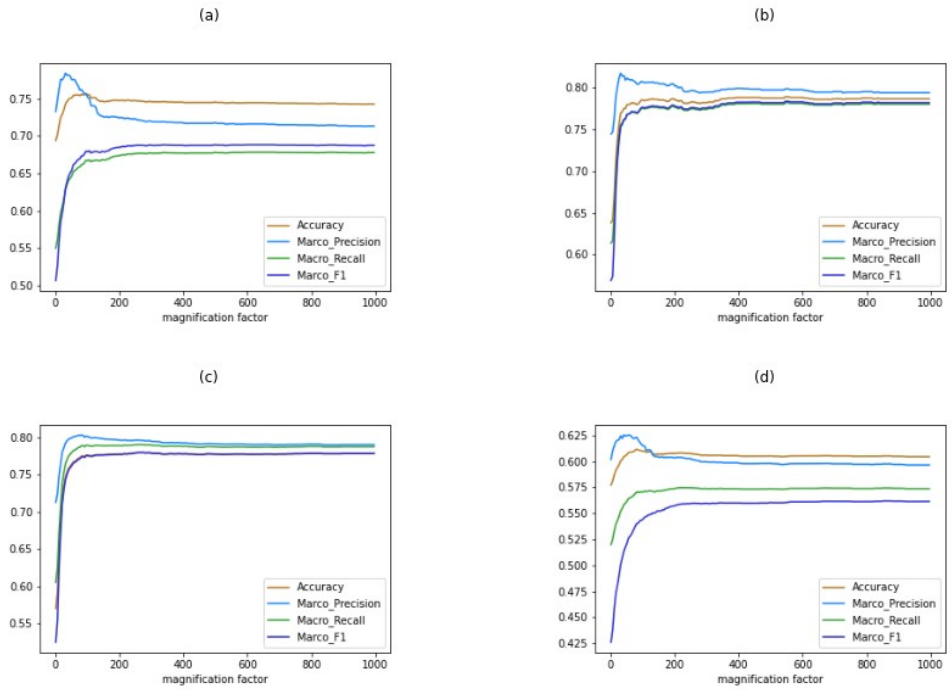

Figure 4: Performance of the proposed classification model for magnification value from 1 to 1000.
Figure 4 (a) indicates performance on dataset HASOC_HI_2021
Figure 4 (b) indicates performance on dataset HASOC_HI_2020
Figure 4 (c) indicates performance on dataset HASOC_HI_2019
Figure 4 (d) indicates performance on dataset HASOC_HI_COMBINED

5.2. Results using the test dataset

It is observable in Figure 3 and Figure 4, that performance of the proposed classification model starts to converge around magnification factors 100 and 150 respectively. Therefore a magnification factor of 100 and 150 is chosen for the evaluation of the proposed classification model on test data for English and Hindi respectively. The evaluation score for the test dataset is given in Table 4.

Table 4
Evaluation of the proposed binary classification model on the test data.

Name of the dataset	Macro F1	Macro Precision	Macro Recall	Accuracy	Magnification Factor
subtask-1A (English)	0.6813	0.6797	0.6882	0.69243	100
subtask-1A (Hindi)	0.6762	0.7126	0.6658	0.74151	150

6. Analysis

A few instances of misclassified Twitter posts for both the English and Hindi test datasets are mentioned in Table 5.

Table 5
Examples of misclassified Twitter posts for both English and Hindi test dataset

Instance id	Original Text	Preprocessed Text	Expected Label	Predicted Label
from subtask-1A (English)				
1	the world suffers a lot. not because of the violent of the bad people but because of the silence of the good people." // relevant always #bengalburning #bjp	world suffers lot violent bad people silence good people relevant always	NOT	HOF
2	he fails india, he fails the world, hefails humanity. #vinashakvista #resignmodi https://t.co/3jluapqhuy	fails india fails world fails humanity	HOF	NOT
3	you have failed as #primeminister @narendramodi #modimadedisaster we want proper #democracy you are not that leader you were in 2013. #resignpmmodi https://t.co/nghswp9ea5	failed want proper leader 2013	HOF	NOT

from subtask-1A (Hindi)				
4	@hemantmkpandya @news24tvchannel @aloksharmaaicc @manakgupta गधा तू है,इसलिए एक ही बक रहा है।	गधा तू हैइसलिए एक ही बक रहा है।	HOF	NOT
5	फट्टू हैं bjp वाले #cruelmamata #bengalviolence #bengalburning https://t.co/13vmf806ht	फट्टू हैं bjp वाले	HOF	NOT
6	हमारी वाहवाही संपूर्ण संसार में है। पर बेशर्मों से शर्म की दुहाई क्यों? #prayforfarmersvictory #farmersprotest #resignmodi https://t.co/iwebqufwdw	हमारी वाहवाही संपूर्ण संसार में है। पर बेशर्मों से शर्म की दुहाई क्यों	NOT	HOF

It is noticeable that even though instances 1,6 were NOT statements, they were predicted as HOF. The reason behind this misclassification is that words like 'suffer', 'violent' of instance 1, and the word 'बेशर्मों' of instance 6, are part of HOF_knowledge_base of the English and the Hindi respectively. Although words like 'suffer', 'violent', 'बेशर्मों' are not HOF by nature but are highly associated with HOF posts in the training datasets. As a result, while parsing with module parse_HOF the cumulative score shoots high which in turn results in misclassification. Also instances 2-5 belong to HOF but were classified as NOT as they do not contain any foul, offense, or vulgar words in the statement.

7. Conclusions and Future Work

We have seen from instances of Table 5 that misclassification occurred when non-HOF words which are highly associated with HOF context are used in NOT statements or only non-HOF are used for HOF statements. So, including context information, while classifying a statement can improve the performance of the model. Although our proposed classification model is able to identify HOF statements when hate offensive phrases are present in the statement. In future, the usage of different transformer-based models along with external datasets will be considered for research work.

8. Acknowledgment

We are thankful to the organizers of HASOC 2021 for providing the opportunity. We acknowledge all the co-authors also for their efforts and contribution to this research work.

9. References

[1] K. Saha, E. Chandrasekharan, M. De Choudhury. 'Prevalence and Psychological Effects of Hateful Speech in Online College Communities. Proc ACM Web Sci Conf. 2019;2019:255-264. doi:10.1145/3292522.3326032

[2] B. Mathew, R. Dutt, P. Goyal, A. Mukherjee. 'Spread of Hate Speech in Online Social Media'. arXiv preprint arXiv:1812.01693v1.

[3] S.O. Sood, E. F. Churchill, and J. Antin. 2012b. Automatic Identification of Personal Insults on Social News Sites. J. Am. Soc. Inf. Sci. Technol., 63(2):270–285, February.

[4] M. Mondal, L. A. Silva, and F. Benevenuto. A Measurement Study of Hate Speech in Social Media. In HT, 2017.

[5] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive Language Detection in Online User Content. In WWW, pages 145–153, 2016.

[6] N. Djuric, Hate Speech Detection with Comment Embeddings. In: Proceedings of the 24th international conference on world wide web, New York; 2015. P. 29–30

[7] H. M. Saleem. A Web of Hate: Tackling Hateful Speech in Online Social Spaces. arXiv:1709.10159 [cs]. 2017

[8] J.H. Park, P. Fung. One-step and Two-step Classification for Abusive Language Detection on Twitter. arXiv preprint arXiv:1706.01206. 2017

[9] Z. Zhang, D. Robinson, and J. A. Tepper. Detecting Hate Speech on Twitter using a Convolution - GRU based Deep Neural Network. In Proceedings of ESWC 2018.

[10] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. Deep Learning for Hate Speech Detection in Tweets. In Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee 2017, 759–760.

[11] 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/.

[12] 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, CoRR abs/2108.05927(2021). URL: https://arxiv.org/abs/2108.05927, arXiv:2108.05927.

[13] 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.