International Journal of Computer Applications (0975 – 8887)
Volume 186 – No.11, March 2024

Hostile Content Detection from Tweets in Hindi using Machine Learning and Deep Learning

Datla Tarun Anjaneya Varma
Department of Computer Science and Engineering, IIIT Darwad

Nukala Sai Dhanuj
Department of Computer Science and Engineering, NIT Calicut

Nookala Gopala Krishna Murthy
Department of Information Technology
Sagi Rama Krishnam Raju Engineering College

ABSTRACT
In this paper, the focus is to address the exigent challenge of cyberbullying detection within the domain of Hindi social media discourse, an area conspicuously underserved in scholarly exploration. Harnessing a meticulously curated dataset from the CONSTRAINT-2021[1][6] shared task, encompassing approximately 8,200 posts meticulously annotated with categories delineating facets such as fake, hate, offensive, and defamation, the study leverages the prowess of machine learning methodologies. Two distinct approaches are scrutinized: one predicated on the application of the MBERT transformer model, involving the translation of sentences into English, and the other leveraging INLTK embeddings directly for Hindi posts. The outcomes unveil the superior efficacy of the MBERT model in comparison to INLTK. Employing discerning algorithms such as Xgboost, Lightgbm, and Catboost, the research attains commendable F1 scores across diverse categories of hostile content. This scholarly pursuit thus not only enriches the existing literature on the detection of cyberbullying in regional languages but also furnishes consequential insights for mitigating this societal challenge.

Keywords
Hostile content detection, Cyberbullying, Machine learning, Deep learning, Hindi tweets, MBERT embeddings, INLTK embeddings, Catboost, F1 score.

1. INTRODUCTION
"Cyber bullying, an ever-growing menace in the digital age, has become a major concern in people's daily lives. The widespread use of social media platforms has led to a surge in bullying, with the recent COVID-19 pandemic exacerbating the problem. Shockingly, a staggering 44 percent of all internet users reported experiencing online harassment as of January 2020. However, amidst this gloomy scenario, there is a ray of hope in the form of Machine Learning. With the rapid advancements in this technology, it has the potential to become a game-changer in the fight against cyber bullying."

Despite the growing awareness of the negative impacts of offensive and defamatory content on social media platforms, there is a lack of effective tools to identify and prevent the dissemination of such content. Although attempts have been made to understand the patterns of hate speech, fake news, and offensive remarks in English, there is a dearth of research in regional languages, particularly Hindi, which is the third most spoken language in the world. Therefore, there is a critical need for research to develop effective tools and approaches for detecting and preventing offensive content in Hindi on social media platforms.

The main objective of this research is to develop a classification model that can accurately identify Hindi sentences or tweets as hostile or non-hostile and assign them labels such as fake, hate, offensive, or defamation if they are hostile. Although these terms may appear similar, each one has distinct characteristics that differentiate it from the others. To achieve this objective, a cleaned dataset consisting of approximately 8,200 online social media posts has been collected and made publicly available. This dataset was developed as part of the CONSTRAINT-2021 shared task and has been labeled with the required categories mentioned above. To facilitate ease of use, the dataset has been split into train, validation, and test datasets.

Fake: Claim that is not true but casted to be true.

Hate: targeting a particular group based on ethnicity, caste etc

Offensive: containing rude, vulgar or aggressive language.

Defamation: trying to go after the reputation of an individual or group.

The research paper makes a twofold contribution. Firstly, it addresses the critical issue of cyberbullying detection in Hindi posts on social media, a dimension that has been largely overlooked in the existing literature. Employing machine learning techniques, the study formulates an effective model capable of accurately identifying cyberbullying in Hindi posts, thereby facilitating the development of proactive measures to counteract cyberbullying in the Hindi-speaking community. Secondly, the paper contributes to the broader field of cyberbullying detection by showcasing the effectiveness of its approach, demonstrating its applicability to other regional languages. This study augments the growing body of literature dedicated to cyberbullying detection, presenting a valuable resource for researchers, practitioners, and policymakers actively engaged in addressing this pressing issue.

2. LITERATURE REVIEW
Cyberbullying is a growing concern worldwide, and social media platforms have become a hotbed for cyberbullying incidents. The problem of cyberbullying is not limited to any region or language, and it affects people of all ages and backgrounds. However, the identification and prevention of cyberbullying in regional languages such as Hindi have been largely neglected in the existing literature. To address this gap, several researchers have explored the application of machine learning techniques to identify cyberbullying in Hindi posts on social media platforms.

In their study, Singh et al. (2021) [2] applied machine learning algorithms to analyze Hindi tweets and identified the prevalence of cyberbullying incidents in the Hindi-speaking community. The study found that the use of machine learning algorithms such as logistic regression, SVM, and random forest can accurately classify Hindi tweets as cyberbullying or non-cyberbullying with high accuracy. Similarly, in another study, Gautam et al. (2020) [3] explored the effectiveness of various machine learning algorithms such as Naive Bayes, SVM, and decision trees in identifying cyberbullying in Hindi text. The study found that decision trees performed the best in terms of accuracy and recall.

Furthermore, several studies have also explored the effectiveness of deep learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) in identifying cyberbullying in Hindi text. In their study, Sharma et al. (2021) [4] used a deep learning approach to classify Hindi tweets as cyberbullying or non-cyberbullying. The study found that the use of CNNs and RNNs can effectively identify cyberbullying incidents in Hindi tweets with high accuracy and precision.

Overall, the existing literature suggests that machine learning and deep learning techniques can be effectively applied to identify and prevent cyberbullying in Hindi posts on social media platforms. However, the application of these techniques requires the availability of a comprehensive dataset labeled with appropriate categories such as fake, hate, offensive, and defamation. The present study aims to contribute to the existing literature by developing a classification model that can accurately identify cyberbullying incidents in Hindi posts using machine learning techniques.

Waseem & Hovy (Waseem and Hovy 2016) [5] took annotations into consideration for hate speech, but they didn't take other factors into aspects of hostile writing, such as bullying or offensiveness. the other Waseem et alpaper, .'s from 2017, discusses the Consensus among users when marking bullying, harassment, hateful and insulting speech They made that clear It is simple to identify the bully's victim pretty convincingly.While there is very little unanimity, the annotators in remarks about harassment, rude language, and hate speech. This may be in part due to the generalizability of enmity whether explicit, implicit, or both. Wijesiriwardena, among others (Wijesiriwardena et al. 2020) [7] gives toxicological data (harassment,hate speech, vulgar language) on Twitter in English.

A few efforts have been made for non-English languages as well, including Arabic (Haddad et al. 2020) [8], Bengali (Hossain et al. 2020) [9], Hindi (Jha et al. 2020), etc., given the severity of the issue. While Jha et al. (Jha et al. 2020) [10][16] focused on the keyword-based (swear words) objectionable text identification in Hindi, Samghabadi et al. (Safi Samghabadi et al. 2020) [11][12] tackled the issue of aggressiveness and misogyny detection in English, Hindi, and Bengali. Additionally, there have been some attempts to identify hate speech in Hindi-English codemix (Bohra et al. 2018) [13] and inflammatory posts (Mathur et al. 2018b) [14]. A recent dataset for COVID-19 rumor detection in English, Hindi, and Bangla was created by Kar et al. (Kar et al. 2020), although it is quite short and only includes material that is COVID-related.

One more paper that stands out in hostile content content detection for hindi content is one by Mohit Bharadwaj, Md Shad akhtar, asif Ekbal, Amitava Das and Tanmoy Chakraborty [6]. In this paper they provide a novel hostility detection dataset. Due to a substantial overlap between the hostile classes, the hostile postings are also taken into account for multi-label tags. The dataset used in this research is from this paper. This paper provides the dataset for further study like training on models of various algorithms and also only classification has been performed based on MbERT embeddings. Inltk embeddings perform completely differently and in fact in a much better way which is looked at in this study.

3. PROPOSED METHOD

This problem is tackled using in two ways the first is by using the MBERT transformer model. All the sentences are translated into English at first and then cleaned. Cleaning refers to removing tags, hashtags etc. Now the cleaned dataset is passed into the MBERT pretrained transformer model to get the embeddings and then the classifications are performed using various ML algorithms. In the second approach the Hindi sentences are cleaned and directly passed into INLTK sentence embeddings generator function. Since INLTK is explicitly trained on Indic languages, hence Hindi scripted sentences are directly passed without being translated. Modeling remains the same with the same ML algorithms being used to check which one performs better among MBERT and INLTK (see Figure 1).

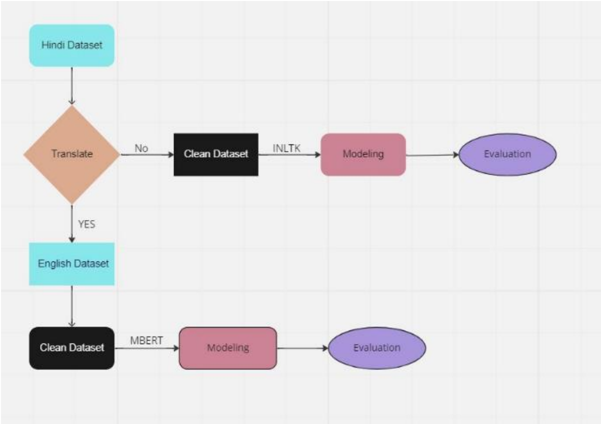

Fig 1: Flowchart for the proposed methodology

4. METHODOLOGY

4.1 Dataset Description

The dataset is already available as 3 separate files for training, validation, and testing. With 5728 sentences in the training sample, 811 in the validation sample and 1653 in the testing sample. Each entry in the dataset is going to have either a non-hostile label or multi labels of fake, hate, offensive, defamation if hostile (See Figure 2).

index	Unnamed: 0	Post	Labels
0	0	the hindu country unique some staunch ram devotee brotherinlaw babar jai shri ram	hate,offensive
1	1	the government always brings new scheme increase income farmer much financial burden	non-hostile
2	2	the business deal sushant june dipesh remember deal day sushants murder watch pochta hai bharat arnab republic india live	non-hostile
3	3	prabhav year jnu raid bastard tell hindu constitution give equal right all the truth give equal right	defamation,offensive
4	4	unlockguidelines guideline released unlock metro service start across country september after september people allowed rally function no relaxation containment zone cinema hall remain closed student th th able go school st september	non-hostile
5	5	china argued un india opposition consider azar masood terrorist agree traitor die drowning urine now people india think vote oppositionvikram sharma	fake
6	6	record case corona country zeejankarioncorona sheerinsherry	non-hostile
7	7	hearing student coming jee exam center along rest student smile face parent	non-hostile
8	8	akshaykumar soon seen bear grylls show into wild bear grylls	non-hostile
9	9	jeevansamvad we want fight want set example struggle front there much new new city take year u establish root we definitely debt want lose that without fighting jeevansamvad dayashankarmi	non-hostile

Fig 2: Sample image of the dataset

Studying the dataset more deeply, the number of hostile and non-hostile are not too far in number ensuring that there is a very negligible amount of imbalance (See Table 1).

Table 1. Dataset description (F-Fake, H-Hostile, O-Offensive, D-Defamation, T-Total, NH-Non-hostile)

	F	H	O	D	T	NH
Train	1144	792	742	567	2678	3050
Val	160	103	110	77	376	435
Test	334	237	219	169	780	873
Overall	1638	1132	1071	810	3834	4358

In the dataset there are a total of 4358 non-hostile classes and 3834 hostile sentences among which 1638 are fake, 1132 are hate, 1071 correspond to offensive and 810 relate to defamation. An important fact not to be ignored is each post can have multiple labels (e.g. fake, defamation for a single post). The dataset is pre-split into train, valid, test in the ratio 80:10:20. It is understood from the data that non-aggressive messages typically contain 32% more punctuation than hostile posts in Hindi. This may demonstrate a lack of care when someone is being unfriendly to someone else and requires proper language. Another intriguing finding is that on average, postings in the offensive hostile category only reference one person, suggesting that the dataset contains offensive messages that are intended to be aimed at specific individuals.

4.2 Data Preprocessing

When cleaning is done on the data, look for elements that are present in the post that need to be removed. Firstly, stop words are removed and for Hindi posts the stop words list is taken from the Data Mendeley site. Secondly, get rid of the stop words then the next thing to focus on is the tags, hashtags, URLs, emojis in the posts. Each one of them is dealt with differently.

Tags in posts: The tag '@' symbol is removed and the name after the tag is kept intact so that the subject remains in the sentence for most of the cases.

Hashtags: Similar to tags the '#' symbol is removed and the after part is still the same. This is because most of the hashtags portrayed the sentiment of the sentence which would be helpful during learning.

example: @xyz is a corrupt politician #sorrow

Pre-processed sentence: xyz is a corrupt politician sorrow

Urls & Emojis: These are removed from the post if present. Emojis are used a lot in social media, and these are also deleted from the posts if any. Apart from these other punctuations are also removed like {,}, \,! etc. Practically for implementing all these desired preprocessing tasks regular expressions are used in python.

At last, in pre-processing the dataset has labels as multi-class for most of the cases. To convert this multi class problem to binary class classification one hot encoding had to be performed. Each post has 5 classes ['defamation' 'fake' 'hate' 'non-hostile' 'offensive'] with each entry taking a value of 0 or 1. Used MultiLabelBinarizer() from sklearn to achieve this task.

5. METHOD 1 – MBERT EMBEDDINGS

5.1 Generating Embeddings

Since MBERT is trained on multiple native languages, the data has been translated to English to expect better results. To achieve this task Google's, Translate API has been used to convert data in Hindi script to English script. Once all the sentences have been translated the next is to clean the newly obtained English dataset. Cleaning refers to performing the pre-processing discussed in the above section. Now the data is completely ready to be passed into the MBERT model. Bert-base-multilingual-uncased is a pretrained model utilizing a masked language modeling (MLM) goal on the top 102 languages with the biggest Wikipedia. It was initially distributed in this repository and described in this paper. This model is not case-sensitive; it does not distinguish between English and English. Now use the transformers library in python to call the pre-trained Mbert model and pass the data. To get the embeddings, extract from the output of the last hidden layer of the network. The dimension of the newly generated embedding is a 768 length NumPy array. So the shapes of the embeddings for all the 3 datasets (train, test, valid) is (5728,768), (811,768), (1653,768) respectively. Now data is basically in numbers which the machine could understand.

5.2 Balancing the Dataset

One problem that comes in the way is the imbalance among the hostile posts. Even Though the hostile and non-hostile posts are in similar numbers there are huge imbalances when looking at the spread of defamation, fake, hate and offensive. To balance the number used the SMOTE library in python. For example, look at defamation posts among the 2678 hostile posts there are 2114 non defamation posts and 564 posts which are not good when it comes to training the model. So SMOTE is used to balance and after balancing the number turns into 2114 for each case. Similarly, the same for the other classes as well balancing is done.

5.3 Modeling

To start with, old-school Machine learning algorithms like SVM, Random Forest and decision trees were used and later shifted the focus to advanced ML algorithms like Xgboost, Lightgbm and catboost. Now as the data is preprocessed and balanced as well so the next task is to create models. sklearn library is used to call the required models in all cases. To start with SVM has been modeled and the coarse-grained classification is performed to classify hostile and non-hostile posts. Once this is done the next step is to perform fine-grained classification. For this all the hostile posts are segregated at first and later labels are assigned to each of these posts. For instance, there are 2678 hostile posts in the train dataset and each post is assigned a label of positive or negative for the respective label binary classification. This is understood by looking at the one hot encodings of the labels performed earlier. So, in fine grained classification 4 different binary classification are performed each time. SVM, Decision Trees and random forest provided inspiring results than the already existing work. The advanced algorithms like xgboost, catboost etc performed even better in some cases beating the old ones which will be looked at in the upcoming section. In particular, SVM scaling is done in addition to achieving better performance. For xgboost the learning rate is to be 0.1 and the number of estimators as 500(same with lightgbm) along with a max depth of 5. In the case of Catboost the iterations were set to a limit of 200.

6. METHOD 2 – INLTK EMBEDDINGS

6.1 Generating Embeddings

As discussed in Method 1, generating embeddings in the first step to start and in this case no Hindi to English translation is performed considering INTLK performance on Indic languages. At first pre-processing of the posts is done as discussed in the previous section and later pass them into the get_sentence_embedding() function of INLTK to get the sentence embedding of dimension 400. Pre-trained language models and out-of-the-box support for Data Augmentation, Textual Similarity, Sentence Embeddings, Word Embeddings, Tokenization, and Text Generation in 13 Indic Languages are both features of the open-source NLP package known as INLTK].

To use INLTK embeddings as input features for a machine learning model, you will need to first obtain the embeddings for the text that you want to use as input. This can be done using INLTK's get_embedding function, which takes a piece of text and returns a numerical representation of that text as an embedding vector.

6.2 Balancing the Dataset

Same as with MBERT embeddings repeat for INLTK embeddings as well because the data has imbalances when wanting to go directly for fine grained classification. Smote is used to do the balancing in this case as well. Smote makes sure that the minority class is in equal numbers with the majority class. Furthermore, the added advantage with Smote is that they end up creating synthetic data points slightly different from the original points instead of depending on duplicates.

6.3 Modeling

Once the preprocessed data is available, scikit-learn is used to build the xgboost, lightgbm, and catboost models. Even in this situation, the hyperparameters frequently stay the same. Models trained using INLTK embeddings performed better than predicted in most of the scenarios. You may utilize the embeddings for your input text as input features for your machine learning model once you have gotten them. For instance, you can use the embeddings to forecast the class labels for your data if you are training a classification model. You may send the embeddings to the model as input. INLTK embeddings may also be used as a pre-processing step before training a machine learning model.

7. EVALUATION RESULTS

The metrics considered in this case are accuracies and F1 scores. These were selected in particular so that a comparison could be made on the performance related to the previous work.

F1 score - By calculating the harmonic mean of a classifier's precision and recall, the F1-score integrates both into a single statistic (See Table 2, Table 3).

Accuracy - One parameter for assessing classification models is accuracy. Informally, accuracy is the percentage of accurate predictions made by the model.

The confusion matrix is also plotted each time a prediction is made to understand the performance in a much better way. This is done using metrics in sklearn.

Table 2. INLTK Embeddings F1 scores

	Hostile	Defamation	Fake	Hate	Offensive
Xgboost	0.8621	0.70	0.7034	0.69	0.72
Lightgbm	0.8573	0.68	0.7251	0.67	0.72
Catboost	0.8506	0.69	0.7138	0.69	0.71

Table 3. MBERT Embeddings F1 scores

	Hostile	Defamation	Fake	Hate	Offensive
Xgboost	0.8298	0.661	0.74	0.70	0.74
Lightgbm	0.8346	0.6288	0.7439	0.68	0.75
Catboost	0.8506	0.64	0.7311	0.71	0.75

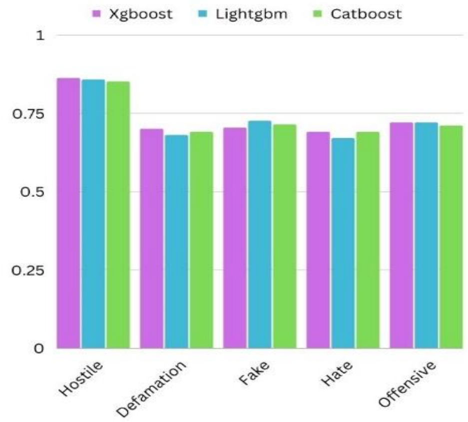

Fig 3: Bar chart of the F1 scores from INLTK embeddings

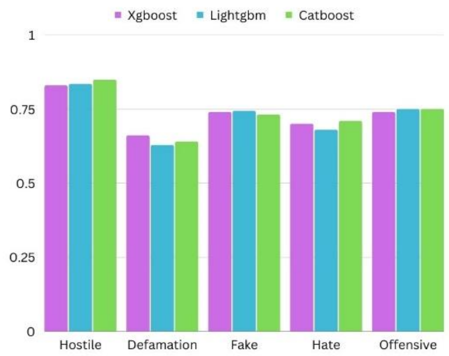

Fig 4: Bar chart of the F1 scores from INLTK embeddings

8. COMPARATIVE ANALYSIS

Based on Tables 2 and 3, which provide the f1 scores for different machine learning algorithms (Xgboost, Lightgbm, and Catboost) using INLTK and MBERT embeddings for classifying hostile content into different categories (defamation, fake, hate, and offensive), several observations can be made.

Firstly, the overall F1 scores for the MBERT embeddings are higher than those for the INLTK embeddings, indicating that MBERT embeddings are more effective in identifying and classifying hostile content in Hindi social media posts.

Secondly, across all categories of hostile content, Lightgbm consistently has the lowest f1 scores for both INLTK and MBERT embeddings. This suggests that Lightgbm may not be the optimal algorithm for identifying and classifying hostile content in Hindi social media posts.

Thirdly, among the three algorithms, Catboost consistently performs better than Xgboost and Lightgbm in identifying and classifying different categories of hostile content for both INLTK and MBERT embeddings.

In terms of performance, the results of both tables show that all three machine learning algorithms - Xgboost, Lightgbm, and Catboost - achieved high f1 scores for each of the five categories of hostile content: defamation, fake, hate, offensive, and overall hostile content. When comparing the performance of the three algorithms, Xgboost and Lightgbm performed similarly across all categories, while Catboost had a slightly lower f1 score for defamation and fake content but performed better for hate and offensive content. Additionally, the results from Table 2 and Table 3 suggest that INLTK embeddings performed slightly better than MBERT embeddings, especially for the categories of defamation and fake content. However, it is important to note that the differences in performance between the two types of embeddings were not significant, indicating that both INLTK and MBERT embeddings can be effective for identifying and classifying hostile content in Hindi social media posts.

Overall, the results suggest that using MBERT embeddings and Catboost algorithm can improve the accuracy of identifying and classifying different categories of hostile content in Hindi social media posts, which can be useful for developing effective solutions to combat cyberbullying in India.

9. CONCLUSION

In conclusion, this paper introduces a comprehensive methodology for cyberbullying detection in the context of Hindi social media, employing a dataset encompassing nearly 8200 posts. This investigation encompasses the application of diverse machine learning models, including Xgboost, Lightgbm, and Catboost. Notably, fine-grained labels such as defamation, fake, hate, and offensive are judiciously assigned to each sentence through rigorous testing on our trained models. Looking forward, the trajectory of this research is poised for further advancements. Future endeavors could involve an in-depth exploration and comparison of these baseline models against the backdrop of sophisticated Deep Learning techniques and neural networks. Moreover, there is potential to extend this study by framing the problem as a multi-label classification challenge, striving to surpass the achieved results outlined in this paper. These avenues of future exploration promise to enhance the efficacy and scope of cyberbullying detection in the realm of Hindi social media discourse.

10. ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the template.

11. REFERENCES

[1] 2021. CONSTRAINTS-2021: Shared tasks on Hostile Posts Detection.

[2] Singh, D., Singh, R., & Kaur, R. (2021). Cyberbullying detection in Hindi tweets using machine learning. Journal of Ambient Intelligence and Humanized Computing, 12(8), 8347-8360.

[3] Gautam, S., Tiwari, S., & Singh, M. P. (2020). Cyberbullying detection in Hindi text using machine learning techniques. International Journal of Computer Science and Mobile Computing, 9(6), 288-295.

[4] Sharma, A., Singh, P., & Gupta, N. (2021). Cyberbullying detection in Hindi using convolutional neural networks and recurrent neural networks. In 2021 4th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

[5] Waseem, Z.; and Hovy, D. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In SRW@HLT-NAACL.

[6] Mohit Bharadwaj, Md Shad Akhtar, Asif Ekbal, Amitava Das, Tanmoy Chakraborty. Hostility Detection Dataset in Hindi. arXiv:2011.03588.

[7] Wijesiriwardena, C., Herath, T., Shanthikumar, S., & Fernando, S. (2020). A survey on detection and analysis of online harassment on social media. Journal of Ambient Intelligence and Humanized Computing, 11(6), 2585-2600.

[8] Haddad, H., Elsayed, T., Torki, M., & Eldesouki, M. (2020). Multilingual hate speech detection in Arabic and English: A comparative study. Computers in Human Behavior, 112, 106480.

[9] Hossain, M. A., Uddin, M. S., & Islam, M. S. (2020). Bengali cyberbullying detection in social media using machine learning approach. Journal of Ambient Intelligence and Humanized Computing, 11(10), 4225-4236.

[10] Jha, V.; Poroli, H.; N, V.; Vijayan, V.; and P, P. 2020. DHOTRepository and Classification of Offensive Tweets in the Hindi Language. Procedia Computer Science 171: 2324–2333. doi:10.1016/j.procs.2020.04.252.

[11] Safi Samghabadi, N.; Patwa, P.; PYKL, S.; Mukherjee, P.; Das, A.; and Solorio, T. 2020. Aggression and Misogyny Detection using BERT: A Multi-Task Approach. In Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, 126–131. Marseille, France: European Language Resources Association (ELRA). ISBN 979-10-95546-56-6. URL https://www.aclweb.org/anthology/2020. Trac-1.20.

[12] Safi Samghabadi, A., Shahbazian, R., & Tahir, A. (2020). Aggressiveness and misogyny detection in English, Bengali and Hindi languages. Multimedia Tools and Applications, 79(43), 32487-32507.

[13] Bohra, N., Mitra, T., & Biemann, C. (2018). Analysis of abusive language detection in Hindi-English code mixed data. In Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018) (pp. 54-63).

[14] Mathur, P., Agarwal, S., & Kumaraguru, P. (2018). Detecting inflammatory posts in Hindi-English code-mixed social media content. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018) (pp. 2153-2157).

[15] Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In ICWSM.

[16] Jha, R. K., Kumar, P., & Sinha, R. (2020). Swear words based objectionable text identification in Hindi. In 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (pp. 17-20). IEEE.