Beware Haters at ComMA@ICON: Sequence and ensemble classifiers for Aggression, Gender bias and Communal bias identification in Indian languages

N Deepakindresh
Vellore Institute of Technology, Chennai, India
deepakindresh.n2019@vitstudent.ac.in

Aakash Ambalavanan
Vellore Institute of Technology, Chennai, India
aakash.ambalavanan2019@vitstudent.ac.in

Avireddy Rohan
Vellore Institute of Technology, Chennai, India
avireddynvsrk.rohan2019@vitstudent.ac.in

B. Radhika Selvamani
Center for Advanced Data Science,
Vellore Institute of Technology, Chennai, India
radhika.selvamani@vit.ac.in

Abstract

Aggressive and hate-filled messages are prevalent on the internet more than ever. These messages are being targeted against a person or an event online and making the internet a more hostile environment. Since this issue is widespread across many users and is not only limited to one language, there is a need for automated models with multilingual capabilities to detect such hostile messages on the online platform. In this paper, the performance of our classifiers is described in the Shared Task on Multilingual Gender Biased and Communal Language Identification at ICON 2021. Our team "Beware Haters" took part in Hindi, Bengali, Meitei, and Multilingual tasks. Our team used various models like Random Forest, Logistic Regression, Bidirectional Long Short Term Memory, and an ensemble model. Model interpretation tool LIME was used before integrating the models. The instance F1 score of our best performing models for Hindi, Bengali, Meitei, and Multilingual tasks are 0.289, 0.292, 0.322, 0.294 respectively.

1 Introduction

This project is a demonstration of the capabilities of sophisticated text based machine learning classifiers which contributed to identifying various hostile features of text data that are provided as a part of the competition for the ICON conference. Ever since social media has become mainstream and gained millions of active users everyday, it has played a pivotal role in various events of information exchange, like photos, comments, tweets etc. As social media platforms encourage free speech from all their users, people can express their opinions on everything they view. As the number of people and this interaction over the web has increased, incidents of aggression and related activities like trolling, cyberbullying, flaming, hate speech, etc. have also increased manifold across the globe(Kumar et al., 2018). Types of hate speech include biased comments against a specific gender, certain caste, race, community or just speech containing abusive language. So it became necessary to find automated solutions to identify hate and abusive text on these platforms to make them a safe place for everyone. In this paper, the performance of deep learning and ensemble classifiers are discussed and analyzed.

Our team "Beware Haters" participated in the shared task of building models to perform classification on multilingual data provided which contained text in three different Indian languages namely Hindi, Bengali, Meitei along with English. Each row in the data contains three different labels belonging to Communal bias, Aggressive, Gender bias. The task is to build and train models to perform classification over the data concerning these three labels. Our team also participated in the individual tasks where the training and the testing data given are purely in one of the particular languages mentioned previously.

The code for this project is available at url[1]

2 Background

This section gives a detailed description of the shared tasks along with the necessary datasets. This dataset (Kumar et al., 2021b) will also contain the description of the datasets for all the languages. Our team participated in Bengali, Hindi, Meitei, and the multilingual track of the competition. Hindi[2] is an Indo-Aryan language predominantly spoken in northern parts of India. Bengali[3] is the national language of Bangladesh and is also spoken in a few parts of India. The training dataset contains only texts in the Indian varieties of Bangla.

[1] https://github.com/Deepakindresh/ComMa-at-ICON-2021

[2] https://en.wikipedia.org/wiki/Hindi

[3] https://en.wikipedia.org/wiki/Bengali$_{language}$

Proceedings of the 18th International Conference on Natural Language Processing:
Shared Task on Multilingual Gender Biased and Communal Language Identification, pages 26–34
National Institute of Technology - Silchar, December 16-19, 2021.©2021 NLP Association of India (NLPAI)

Meitei[4] is a Tibeto-Burman language mainly spoken in the northeastern state of Manipur. Hindi and Bengali texts are written both in English and the respective language. The texts in Meitei are written in English script.

2.1 Task Description

Following is a detailed description of each subtask (Kumar et al., 2021a).

Sub-task A Aggression Identification (Singh et al., 2018). The task will be to develop a classifier that could make a 3-way classification in between 'Overtly Aggressive'(OAG), 'Covertly Aggressive'(CAG) and 'Non-Aggressive'(NAG) text data. [Fig. 1] illustrates the distribution of text classified as 'NAG', 'CAG' and 'OAG' for different languages.

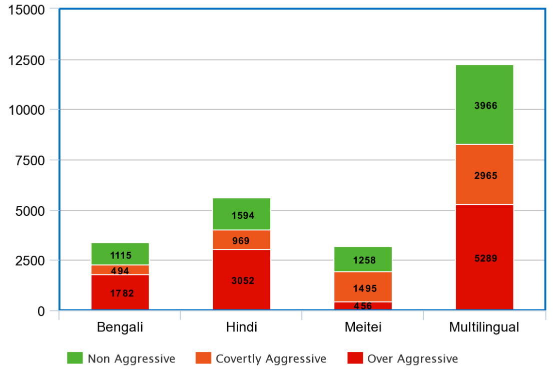


Figure 1: Distribution of text classified as 'NAG', 'CAG' and 'OAG' for different languages.

Sub-task B Gender Bias Identification (Malik et al., 2021). This task will be to develop a binary classifier for classifying the text as 'Gendered'(GEN) or 'Non-Gendered'(NGEN). [Fig. 2] shows the distribution of text classified as 'GEN' and 'NGEN' for different languages.

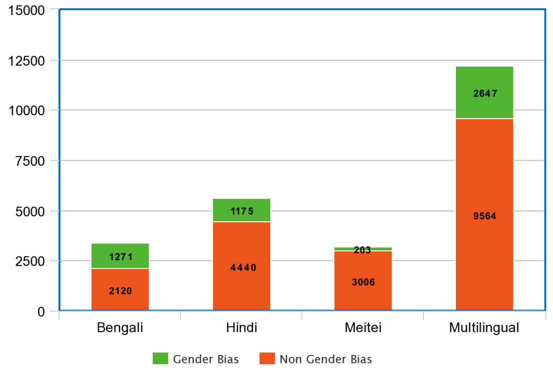


Figure 2: Distribution of text classified as 'GEN' and 'NGEN' for different languages.

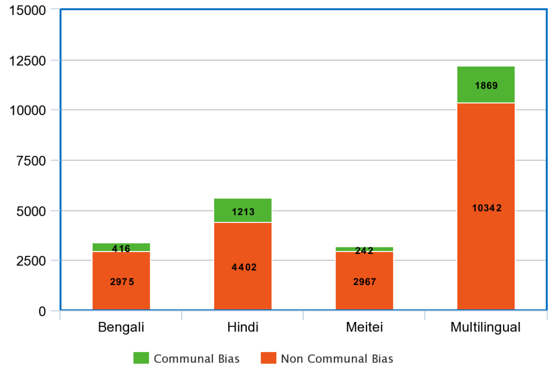


Figure 3: Distribution of text classified as 'COM' and 'NCOM' for different languages.

Sub-task C Communal Bias Identification: This task will be to develop a binary classifier for classifying the text as 'Communal' (COM) and 'Non-Communal'(NCOM).

[Fig. 3] shows how text written in various languages is divided into COM and NCOM categories.

2.2 Dataset

The size of the datasets used for training in various langages have been tabulated in [Table. 1]. This is a combination of both training and dev datasets provided by the organizers.

Language of the Dataset	**Size of dataset**
Multilingual	12211
Hindi	5615
Bengali	3391
Meitei	3209

Table 1: The dataset size for various languages

3 System Overview

The models that are involved in the classification of Gender, Communal and Aggressiveness bias were Random Forest, Logistic Regression and SVM. ensemble methods3.1 and Sequence classifiers are used3.2. Logistic Regression (Oriola and Kotzé, 2020) is a well-known simple regression model that serves as a basic model for binary classification. Random Forest (Liaw and Wiener, 2002) is a widely used meta estimator that fits a number of decision tree classifiers on various sub-samples

[4] https://en.wikipedia.org/wiki/Meitei$_{l}anguage$

of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The Support Vector Machine (Robinson et al., 2018) is a state-of-the-art machine learning model with proven performance in countless machine learning applications with sparse high dimensional data. It uses different kernels, namely Linear, polynomial, Radial basis function, and Sigmoid to transform the data to a lower dimension, which enables the application of a maximum margin classifier for obtaining the decision plane.

Models have been experimented with in every language for all the subtasks and are compared based on their accuracy and F1 score during the testing phase. It is observed that Logistic Regression performed comparatively well in Gender bias and Aggressiveness identification tasks and Random Forest did best in the Communal bias task for every language and SVM was the second-best for all tasks, hence it was decided to combine the three models for better performance and hence an ensemble of these classifiers was built.

3.1 Ensemble Model

Ensemble learning (Rahman and Tasnim, 2014) is a process where multiple diverse models are integrated in a way to obtain better predictive performance than what could be achieved by the models independently. Ensemble classifier of Random Forest, Logistic Regression and SVM with the soft voting method (Pedregosa et al., 2011) were experimented upon. The specific set of models used for the ensemble; a Random Forest classifier with 5000 estimators, a Logistic Regression model with max iterations up to 2000 and a Support Vector Machine using a radial bias function. In soft voting, the class labels were predicted based on the predicted probabilities pp for the classifier – this approach is only recommended if the classifiers are well-calibrated.

$$\hat{y} = \arg\max_i \sum_{j=1}^{m} w_j p_{ij} \tag{1}$$

The equation 1 is used for soft voting method, where wj is the weight that can be assigned to the jth classifier.

The ensemble method outperformed during the testing phase for the Gender bias and Aggressiveness task for all languages but underperformed in comparison to Random Forest and SVM for the Communal bias task. Hence Random Forest model was for this particular task in all the languages. Model interpretation has been done via Local Agnostic Model Interpretation approach to understanding the performance of the models before building the ensemble.

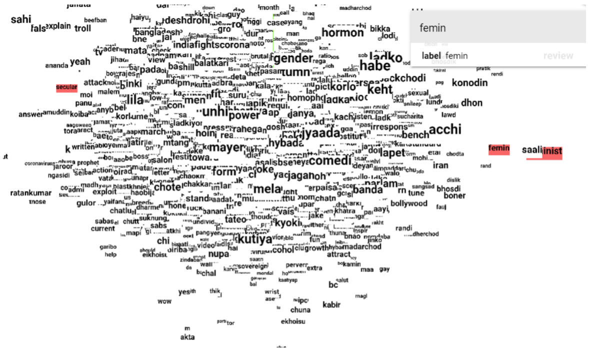


Figure 4: Word embedding for Gender bias identification in multilingual data

3.2 Sequence Model

Sequence classifiers using Bidirectional LSTM (Sundermeyer et al., 2012) were explored. Bidirectional LSTMs train two instead of one LSTMs on the input sequence of the text data. The first LSTM is trained on the forward input sequence while the latter is trained on the backward input sequence. This provides more context to the network and results in a fast and effective learning process. Our LSTM model learned embeddings for the top 10000 words from the whole corpus and these were plotted for further analysis on data and how our model works. The LSTM model made an exceptional performance and impacted even better than ensemble models when used for multilingual dataset because the size of the dataset was almost equal to all the three standalone languages combined together refer Table 1. Since Deep learning models require a huge amount of data for training the Bidirectional LSTM model was able to the surpass ensemble method's performance for multilingual tasks with ease after training for up to 20 epochs. Although the model performed well for binary text classification i.e. for the Gender and Communal bias task when it came to classifying Aggressiveness which was a multi-classification task it could not surpass the performance of the Logistic Regression model and hence Logistic Regression was used for this task. The word embeddings learned by our model was plotted using Principal component analysis from TensorFlow's word embedding projector[5] which is a dimensionality reduction algorithm (Maćkiewicz and Ratajczak, 1993) used to convert vectors with higher dimensions to 3 dimensional vectors for plotting purposes. As you can see to the right of the [Fig. 4] words like 'feminist', 'femin' and 'saali' are together denoting gender biased terms while words like 'secular' are on the opposite side. This denotes our model has performed well in understanding the context behind gender bias and also some small errors in the plot are tolerated since the plot is not with all 32 dimensions rather a reduced version of 3 dimensional vectors.

4 Experimental setup

Only the dataset provided by the organizers were used for all the tasks that we participated in. The training and development datasets are both used for training the models. Detailed description of datasets is provided in [Fig. 5].

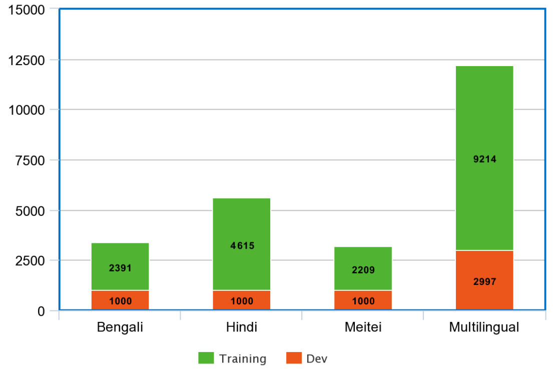


Figure 5: Distribution of Training and Dev datasets used for Training the model in all languages.

4.1 Methods for Preprocessing

The text was either removed or was transformed using pattern matching techniques to deem them fit the classification models under consideration. Non-informative features from the text like URLs, white spaces, non-word characters, RT tags were filtered. Other features like emojis have been filtered out. Stop words are those that appear very frequently in the text but don't help in conveying any meaning, for example, words like 'the', 'a', 'an', 'are' are removed in their corresponding languages as they would mislead algorithms like Tf-idf as it works based on word count and could lead to misclassifications. Hindi and Bangla stop words have been removed from both the multilingual dataset and datasets in the respective languages. In addition, stemming, tokenization and lemmatization of the preprocessed text were performed.

4.2 Vectorization

Tf-idf vectorization was used for ensemble, Logistic Regression and Random Forest models from sklearn and set 'min df' to 3 that ignore terms that have a document frequency strictly lower than the given threshold. [Fig. 6] shows the word cloud in meitei after removing stopwords and applying Tf-idf vectorization.

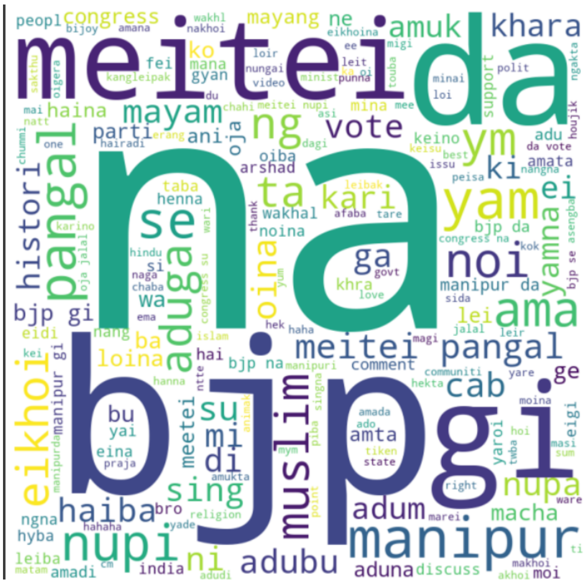


Figure 6: The Word Cloud of the TF-IDF Vector Space after Removing the Stop Words in Meitei

Word embeddings from keras[6] were used for vectorization in Bidirectional LSTM and the input length set as 130 which is the maximum character length of a sentence and set the embedding vector length to 32 as data was limited and also set the size of the input dimensions are 10001. A plot of word embedding vectors used for Communal bias is shown at [Fig. 7] and [Fig. 8] using T-SNE algorithm (van der Maaten and Hinton, 2008) for plotting from TensorFlow's embedding projector[7] after 100 iterations with the default parameters. From the figures, it can be clearly inferred that words with communal bias such as 'muslimvirus', 'boycottmuslim', 'hinduphob' are placed on the top of the plot clearly differentiating from non communal biased words like 'jayhind', 'hindi' which are placed at the bottom.

[5] https://projector.tensorflow.org/

[6] https://www.tensorflow.org/text/guide/word$_{e}mbeddings$

[7] https://projector.tensorflow.org/

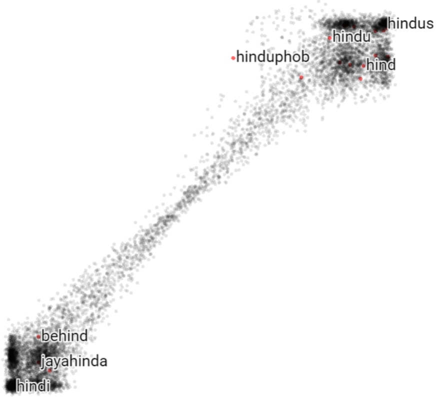


Figure 7: Plot for Communal bias in multilingual data with highlighted word as 'hindu' using T-SNE plot

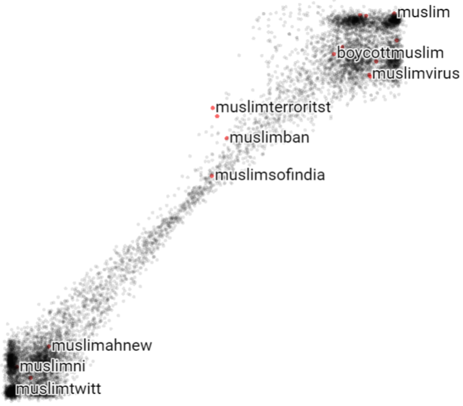


Figure 8: Plot for Communal bias in multilingual data with highlighted word as 'muslim' using T-SNE plot

4.3 Models and Hyperparameter Tuning

The Logistic Regression model from sklearn that was used for classification of Aggressiveness for multilingual data was done setting the multiclass parameter as 'multinomial' and 'max iter' as 1000.

The Random Forest model from sklearn that was used for classification of Communal bias for Meitei, Bengali and Hindi datasets had the 'n estimators' parameter set to 5000.

The ensemble model used for classification of Gender bias and Aggressiveness for Meitei, Bengali and Hindi datasets was done using VotingClassifier from sklearn with three estimators namely LogisticRegression with 'max iter' set to 2000, RandomForest with 'n estimators' set to 5000, and SVC(SVM) with kernel set to 'rbf'. The weights for the models were 2,1,1 respectively and the voting method used was 'soft'.

The Bidirectional LSTM that was used for multilingual language classification of Gender and Communal bias was done using the keras library. Embedding layer followed by 2 Bidirectional layers with 64 and 32 units each and 2 hidden layers with 64 and 1 unit each with the activation as 'relu' (Agarap, 2018) and 'sigmoid' (Elfwing et al., 2017) respectively were added. The loss function used was 'binary crossentropy' (Mannor et al., 2005) and the optimizer was set to 'adam' (Kingma and Ba, 2014). The model ran for 20 epochs as the training accuracy and validation accuracy were highest during training for the testing phase.

Task	Instance F1 Score	Micro F1 Score
Bengali	0.292	0.704
Hindi	0.289	0.689
Meitei	0.322	0.672
Multilingual	0.294	0.665

Table 2: Performance of our highest ranked models for various languages

5 Results

We primarily made 2 submissions where the first submission comprised of ensemble models for Gender bias and Aggressiveness task in all languages and the Random Forest model for the Communal bias task. The second submission predominantly used Bidirectional LSTM for the Gender and Communal bias task and also used Logistic regression for the Aggressiveness identification task for all languages. The instance and micro F1 scores of our best performing models can be found in [Table. 2]. The task wise performance of our highest ranked models for various languages is shown in [Fig. 14].[Fig. 9] shows the performance of various models in subtasks on the multilingual data.For gender bias identification(GEN),the ensemble model gave a slightly higher micro-f1 score compared to Bi-LSTM.The ensemble performed equally well with logistic regression for aggression identification.Random Forest model didnt show a very significant performance compared Bi-LSTM for Communal bias identification.[Fig. 10] shows the performance of various models in subtasks on the Hindi data.The ensemble again performs better for Gender bias identification compared to Bi-LSTM. It also performed slightly better compared logistic regression for aggression detection.The Random Forest gave slightly better micro-F1 compared to

Bi-LSTM.[Fig. 11] shows the performance of various models in subtasks on the Bangla data.Both the ensemble and Bi-LSTM gave a similar micro-f1 score for gender bias identification.However, the ensemble performed slightly better compared to Logistic regression for aggression detection.Both the Random Forest and Bi-LSTM performed equally well for communal bias identification.[Fig. 12] shows the performance of various models in subtasks on the Meitei data.Both the ensemble and the Bi-LSTM performed similarly for the gender bias identification. The ensemble and the logistic regression model performed similarly for aggression detection. The Random Forest performed better compared to Bi-LSTM for communal bias identification.

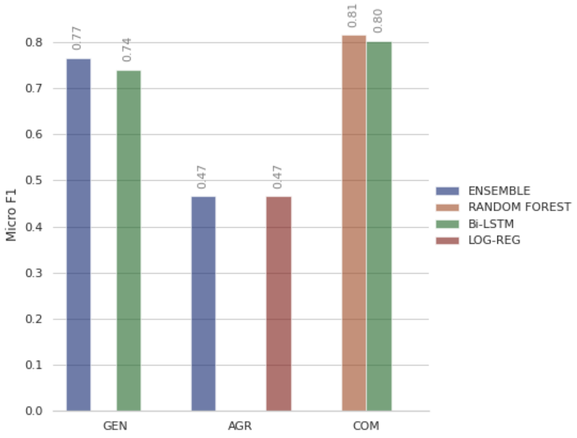


Figure 9: Comparison of Micro F1 scores for each sub task on Multilingual data.

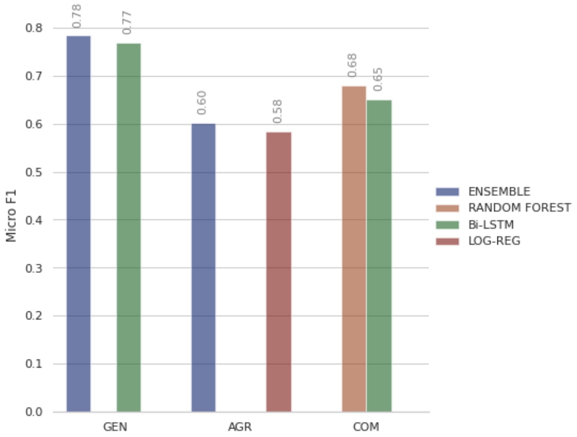


Figure 10: Comparison of Micro F1 scores for each sub task on Hindi data.

Our ensemble and Random forest model performed extremely well for the tasks in Meitei and Bangla and helped us achieve the 1st rank in both the subjects. We secured 3rd rank in Hindi and multilingual tasks where our Bidirectional LSTM contributed most for our rank in multilingual tasks along with Logistic Regression while ensemble technique and Random Forest were used for Hindi. Instance F1 Score Based Ranking Of Team Beware Haters is given in [Fig. 13].

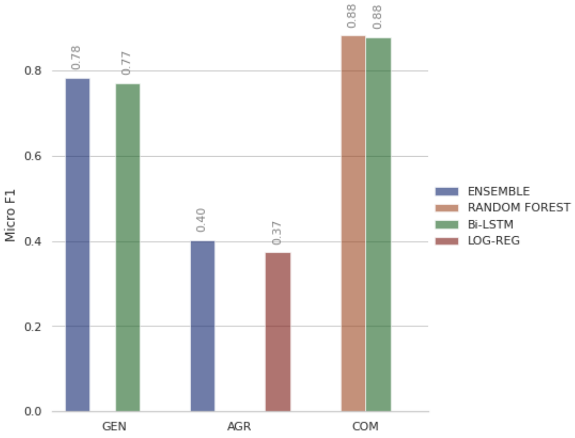


Figure 11: Comparison of Micro F1 scores for each sub task on Bangla data.

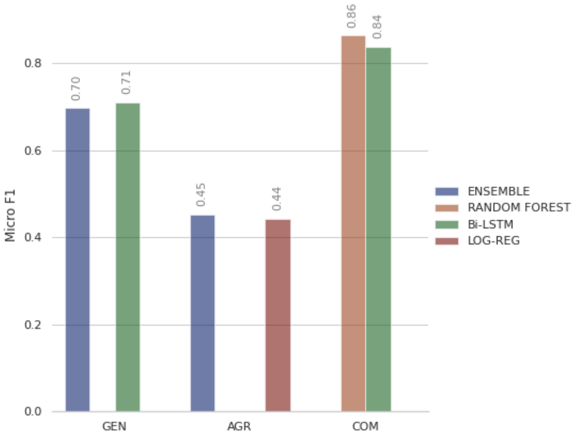


Figure 12: Comparison of Micro F1 scores for each sub task on Meitei data.

5.1 Metrics of Evaluation

Evaluation and ranking of the teams were based on the standard multi-label classification metrics.[8]

- Instance-F1: It is the F-measure averaging on each instance in the test set i.e. the classification will be considered right only when all the labels in a given instance is predicted correctly. It was the primary evaluation metric for all the tasks.

[8] shorturl.at/muHK2

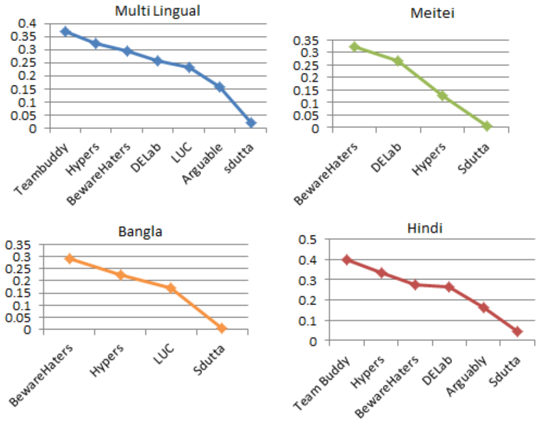


Figure 13: Instance F1 Score Based Ranking Of Team Beware Haters.

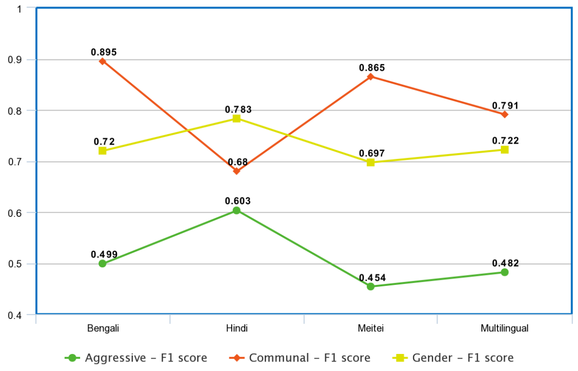


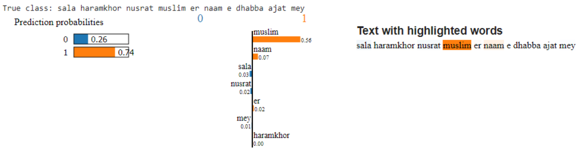
Figure 14: Task wise performance of our highest ranked models for various languages.

- Micro-F1: It gives a weighted average score of each class and is generally considered a good metric in cases of class-imbalance.

  The scores obtained by various models in the tasks in shown in [Table. 2].

5.2 LIME interpretation

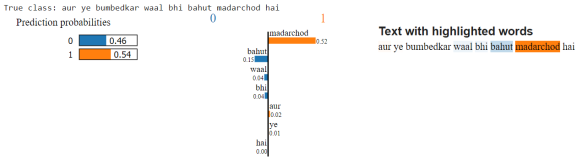
Model explanation strategies were used to better understand the models. LIME is a local agnostic model interpreter (Ribeiro et al., 2016) which provides uniform explanations, irrespective of the model as it is model agonistic. [Fig. 16] shows an example of Gender bias identification. The word "madarchod" is gender abusive word, which is identified by our model and classified as "GEN". A similar example can also be found in the case of identifying communal biases in Bengali. In [Fig. 15], the word "muslim", which denotes the Islamic community is identified, and the text is classified as "COM". Additionally, an example of aggressiveness identification in Meitei is shown in [Fig. 17].



True class: sala haramkhor nusrat muslim er naam e dhabba ajat mey

sala haramkhor nusrat muslim er naam e dhabba ajat mey

Figure 15: LIME explanations for Communal bias identification in Bengali



True class: aur ye bumbedkar waal bhi bahut madarchod hai

aur ye bumbedkar waal bhi bahut madarchod hai

Figure 16: LIME explanations for Gender bias identification in Hindi

5.3 Error Analysis

One of the sources of error is the class imbalance problem which occurs due to the unequal number of biased and unbiased examples in the training dataset, where solutions like undersampling and oversampling of the dataset could lead to major changes in document frequency of Tf-idf vectorization and overfitting issues thus they were left alone and trained. Furthermore, the misclassification made by our models was analyzed using LIME. In [Fig. 18], the text to be analyzed reads ”boy epual girl”. From this, it can be deduced that the word ”equal” is spelled incorrectly as ”epual”. This text is not intended to be gender biased, but due to a misspelled word, our model classifies it as gender biased.

6 Conclusion

We participated in the Shared Task on Multilingual Gender Biased and Communal Language Identification. The sub tasks required building and testing models for multiclass classification task on Aggression identification and binary classification tasks on Gender bias identification and Communal bias identification. The datasets have been provided in Bengali, Hindi, Meitei and finally on multilingual data. Ensemble classifier consisting of Logistic regression, Random Forest and SVM performed better compared to Bi LSTM for Hindi, Meitei, Bengali in both Subtask B and C. But for multilingual data, Bi-LSTM has performed better in these Subtasks. However, in the final submission for Subtask A, Logistic regression has performed better compared to the other models tested. Our team "Beware Haters" ranked 1st in the leaderboard

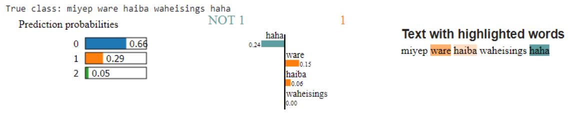


Figure 17: LIME explanations for aggression identification in Meitei

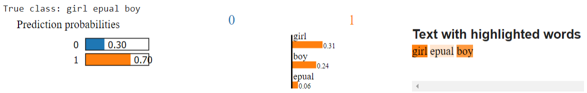


Figure 18: LIME explanations for misclassified example of Gender bias identification

for the Meitei and Bangla dataset and 3rd for both Hindi and multilingual datasets.

We further aim to improve the performance at subtasks by using transformer models like XLM Roberta which have been proven to perform better on multilingual datasets. We also aim to explore other deep learning models which might achieve better performance compared to what our models achieved.

Acknowledgments

We sincerely thank the organizing committee for recognizing our paper and accepting it to be presented at the conference. We would also like to thank the anonymous reviewers who gave useful suggestions, which were taken into account.

References

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. 2017. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Ritesh Kumar, Bornini Lahiri, Akanksha Bansal, Enakshi Nandi, Laishram Niranjana Devi, Shyam Ratan, Siddharth Singh, Akash Bhagat, and Yogesh Dawer. 2021a. Comma@icon: Multilingual gender biased and communal language identification task at icon-2021. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON): COMMA@ICON 2021 Shared Task*, Silchar, India. NLP Association of India (NLPAI).

Ritesh Kumar, Enakshi Nandi, Laishram Niranjana Devi, Shyam Ratan, Siddharth Singh, Akash Bhagat, Yogesh Dawer, and Akanksha Bansal. 2021b. The comma dataset v0.2: Annotating aggression and bias in multilingual social media discourse.

Ritesh Kumar, Atul Kr Ojha, Marcos Zampieri, and Shervin Malmasi. 2018. Proceedings of the first workshop on trolling, aggression and cyberbullying (trac-2018). In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*.

Andy Liaw and Matthew Wiener. 2002. Classification and regression by randomforest. *R News*, 2(3):18–22.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2021. Socially aware bias measurements for hindi language representations. *arXiv preprint arXiv:2110.07871*.

Shie Mannor, Dori Peleg, and Reuven Rubinstein. 2005. The cross entropy method for classification. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 561–568, New York, NY, USA. Association for Computing Machinery.

Andrzej Maćkiewicz and Waldemar Ratajczak. 1993. Principal components analysis (pca). *Computers Geosciences*, 19(3):303–342.

Oluwafemi Oriola and Eduan Kotzé. 2020. Evaluating machine learning techniques for detecting offensive and hate speech in south african tweets. *IEEE Access*, 8:21496–21509.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Akhlaqur Rahman and Sumaira Tasnim. 2014. Ensemble classifiers and their applications: A review. *International Journal of Computer Trends and Technology*, 10(1):31–35.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier.

David Robinson, Ziqi Zhang, and Jonathan Tepper. 2018. Hate speech detection on twitter: Feature engineering vs feature selection. In *European Semantic Web Conference*, pages 46–49. Springer.

Vinay Singh, Aman Varshney, Syed Sarfaraz Akhtar, Deepanshu Vijay, and Manish Shrivastava. 2018. Aggression detection on social media text using deep neural networks. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 43–50.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.