MUM at ComMA@ICON: Multilingual Gender Biased and Communal Language Identification using Supervised Learning Approaches

Asha Hegde[1 a], **M. D. Anusha**[1 b], **Sharal Coelho**[1 c], **H. L. Shashirekha**[1 d]
[1]Department of Computer Science, Mangalore University, Mangalore, India
{[a]hegdekasha, [b]anugowda251, [c]sharalmucs, [d]hlsrekha}@gmail.com

Abstract

Due to the rapid rise of social networks and micro-blogging websites, communication between people from different religion, caste, creed, cultural and psychological backgrounds has become more direct leading to the increase in cyber conflicts between people. This in turn has given rise to more and more hate speech and usage of abusive words to the point that it has become a serious problem creating negative impacts on the society. As a result, it is imperative to identify and filter such content on social media to prevent its further spread and the damage it is going to cause. Further, filtering such huge data requires automated tools since doing it manually is labor intensive and error prone. Added to this is the complex code-mixed and multi-scripted nature of social media text. To address the challenges of abusive content detection on social media, in this paper, we, team MUM, propose Machine Learning (ML) and Deep Learning (DL) models submitted to Multilingual Gender Biased and Communal Language Identification (ComMA@ICON) shared task at International Conference on Natural Language Processing (ICON) 2021. Word uni-grams, char n-grams, and emoji vectors are combined as features to train a ML model with Elastic-net penalty and multi-lingual Bidirectional Encoder Representations from Transformers (mBERT) is fine-tuned for a DL model. Out of the two, fine-tuned mBERT model performed better with an instance-F1 score of 0.326, 0.390, 0.343, 0.359 for Meitei, Bangla, Hindi, Multilingual texts respectively.

1 Introduction

The advancement in internet technology and social media websites have made the information reach the wider audience within no time. These characteristics of social media sites are attracting more and more people towards them leading to exponential rise in the amount of user-generated content on social media. In addition to the exchange of constructive and useful information, few miscreants are taking advantage of the anonymity of users and spreading the abusive and potentially harmful content over the web. While the act of bullying and hate speech kind of things existed very much before the internet, the reach and influence of the internet have given these acts unprecedented power and influence to affect the lives of many people. According to the report by Hinduja and Patchin (2010), these incidents have caused not only mental and psychological agony to the social media users, but have also forced some of them for suicidal attempts in the extreme cases. Abusive, aggressive, communal hate speech and any other forms of potentially harmful content getting generated on social media needs to be filtered out almost instantaneously in order to stop the further spread and the damage it is going to create. Filtering this harmful content manually is almost impossible due to the voluminous amount of data getting generated and also due to the increasing number of social media users. This has received the attention of the research community in recent years to automatically detect such content on social media Waseem et al. (2017).

Identifying the harmful content in social media data automatically is challenging as the social media data which is usually code-mixed do not adhere to the rules and regulations of any language. Further, in a multilingual country like India people tend to pen comments using words from multiple languages making the analysis of social media data more challenging.

To address some of the challenges in identifying gender biased and communal language in code-mixed, multi-scripted, multilingual content on social media, this paper describes the models submitted to ComMA@ICON[1] shared task at ICON

[1] https://competitions.codalab.org/competitions/35482

Proceedings of the 18th International Conference on Natural Language Processing:
Shared Task on Multilingual Gender Biased and Communal Language Identification, pages 64–69
National Institute of Technology - Silchar, December 16-19, 2021.©2021 NLP Association of India (NLPAI)

2021[2]. The shared task is a multi-label three level (Level A, B and C) Text Classification (TC) task in code-mixed and multi-scripted texts in Meitei, Bangla, Hindi, and also in Multilingual (a combination of Meitei, Bangla and Hindi). This shared task is addressed by constructing i) ML classifier with Elastic-net penalty which is trained using word unigrams and char n-grams combined with emoji vectors and ii) fine-tuning a pre-trained multi-lingual Bidirectional Encoder Representations from Transformers (mBERT) as a DL model.

The rest of the paper is arranged as follows: the recent literature related to detection of abusive content in social media data is summarized in Section 2 and the proposed methodology is described in Section 3. Experiments and results are presented in Section 4 followed by conclusion and future work in Section 5.

2 Related work

Several models have been developed by researchers to detect offensive and abusive content in social media text Kumar et al. (2018). The description of some of the recent works are mentioned below:

Li and Fleyeh (2018) have proposed ML approaches using Logistic Regression (LR) with Elastic-net penalty and without penalty, Naive Bayes (NB), Support Vector Machines (SVM), and Random Forest (RF), all trained with word unigrams and bi-grams nd evaluated on the Swedish Twitter dataset containing the public opinion about IKEA - a popular store. LR model with Elastic-net penalty trained using word bi-grams outperformed other models with an accuracy of 0.81 and F1-score of 0.79.

Song et al. (2021) proposed a multilingual toxic text classifier integrating multiple pre-trained models and different loss functions and evaluated its performance on Jigsaw Multilingual Toxic Comment dataset. The proposed learning pipeline begins with a series of preprocessing steps, including translation, word segmentation, purification, digitization, and vectorization to convert word tokens into vectors suitable for TC. mBERT and Cross-lingual Language Model-Robustly Optimized BERT (XLM-RoBERTa) are employed for pre-training through Masking Language Modeling and Translation Language Modeling to incorporate the semantic and contextual information. The results of experiments show that fusion of loss functions and fusion of multilingual models outperform the mbERT and XLM-R models by obtaining F1-scores of 0.505 and 0.76 respectively, justifying the effectiveness and robustness of fusion strategy.

Anusha and Shashirekha (2020) have described the work submitted to subtasks (A and B) of Hate Speech and Offensive Content Identification (HASOC) shared task in Forum for Information Retrieval Evaluation (FIRE) 2020 to identify hate and offensive content in Indo-European Languages. They combine Term Frequency - Inverse Document Frequency (TF-IDF) vectors with additional text-based features to build an ensemble of Gradient Boosting, RF, and XGBoost classifiers, with soft voting. For subtasks A and B, they obtained macro-averaged F1-score of 0.5046 and 0.2596 for English, 0.5106 and 0.2595 for German, and 0.5033 and 0.2488 for Hindi respectively.

Tiţa and Zubiaga (2021) aims to classify English and French text into ”hateful” (hate speech data) and ”non-hateful” (clean/neutral data) categories by fine-tuning mBERT and XLM-RoBERTa on task-specific datasets. The mBERT and XLM-RoBERTa models achieved weighted averages of 0.71 and 0.41 for English and 0.52 and 0.55 for French, respectively.

Tanase et al. (2020) fine-tuned BERT, mBERT and XLM-RoBERTa - the pre-trained Transformer-based architectures using different combinations of task-specific datasets for tackling the problem of aggressiveness detection in MEX-A3T@IberLEF2020 shared task. XLM-RoBERTa model achieved an F1-score of 0.7969, the third-best score in the competition which proves that Transformer-based models can be successfully used to detect aggressiveness in Mexican Spanish tweets.

Davidson et al. (2017) trained a set of multi-class classifiers such as LR, NB, Decision Trees, RF, and Linear SVM to categorize tweets into one of ’hate speech’, ’offensive but not hate speech’, and ’neither offensive nor hate speech’ categories. Their best-performing model achieved a precision of 0.91, recall of 0.90, and an F1-score of 0.90.

Gómez-Adorno et al. (2018) trained a LR algorithm with linguistically motivated features and different types of n-grams to identify if a tweet is aggressive or not in the aggressive detection track at MEX-A3T 2018. They applied several pre-processing steps to standardize tweets in order to capture relevant information and achieved

[2] http://icon2021.nits.ac.in/shared_tasks.html

0.4285 F1-score.

Even though several techniques have been developed to detect abusive content in code-mixed script, very few attempts have been made for Indian languages. This opens up lots of possibilities for experiments on Indian languages, including those with low resources, as well as multilingual text and scripts.

3 Methodology

The proposed methodology consists of Preprocessing, Feature Extraction and Model Construction as explained below:

3.1 Pre-processing

This step aims at removing the noise from the text and preparing the textual content in a format that the learning model can understand. As punctuation, digits, unrelated characters and stopwords are not pertinent to the TC task, they are removed. The stopwords list of Bangla[3] and Hindi[4] [5] are fine-tuned using English stopwords list provided by the Natural Language Tool Kit (NLTK)[6].

3.2 Feature Extraction

As combining word uni-grams, char n-grams and emoji vectors (obtained from pre-trained embeddings) features have shown reasonably good performance Vogel and Jiang (2019), this combination is used as features in this work too. The feature extraction steps are given below:

- **TF-IDF** measures the importance of a word in the corpus. To accomplish this task, several experiments were conducted and based on the results of those experiments, 5,000 frequent char n-grams in range (2, 3) and all words unigrams are extracted and transformed to vectors using TFidfVectorizer[7].

- **emo2Vec** is a word-level representation of emojis in Unicode that encodes them into real-valued, fixed-size vector representations. In emo2Vec[8], emojis are represented in a 300-dimensional space, similar to the Google News word2Vec embeddings. Since there are many emojis, instead of removing them leading to loss of information they are extracted from the text and vectorized using emo2Vec.

Table 1: Details of features used in ML model with Elastic-net penalty

Train set		
Languages	**#Emojis**	**#word uni-grams**
Meitei	236	13,377
Bangla	577	16,478
Hindi	287	6,230
Multilingual	1,100	32,578
Test set		
Meitei	102	13,377
Bangla	286	16,478
Hindi	185	6,230
Multilingual	573	32,578

Table 1 lists the number of emojis and word unigrams extracted from Train and Test sets. Classifier with Elastic-net penalty is trained with a combination of all the extracted features.

3.3 Model Construction

The multi-label classification task is modeled as three separate classification tasks, one for each level and the labels of each of the three levels are concatenated to form a single predicted tuple. A ML classifier using Elastic-net model and a DL classifier using fine-tuned mBERT are proposed to identify gender biased and communal language content. Description of the proposed models are give below:

- **Elastic-net** is a popular type of regularized linear regression that combines two popular penalties, Lasso (L1) penalty and Ridge (L2) penalty Marafino et al. (2015). While Lasso penalty uses shrinkage (for eg., data values are shrunk towards a central point, like the mean) to determine the regression coefficients, Ridge penalty acts to ”average out” estimates of correlated features, which imposes a grouping effect. The elastic-net model produces a significant advantage over the Lasso and Ridge penalties considered individually and gives decent results even with the basic features like word uni-grams and char n-grams. Figure 1

[3] https://github.com/stopwords-iso/stopwords-bn
[4] https://github.com/stopwords-iso/stopwords-hi
[5] https://github.com/TrigonaMinima/HinglishNLP
[6] https://www.nltk.org/
[7] https://scikit-learn.org/stable/modules/
[8] https://github.com/glnmario/emo2vec

depicts the structure of the ML classifier using Elastic-net.

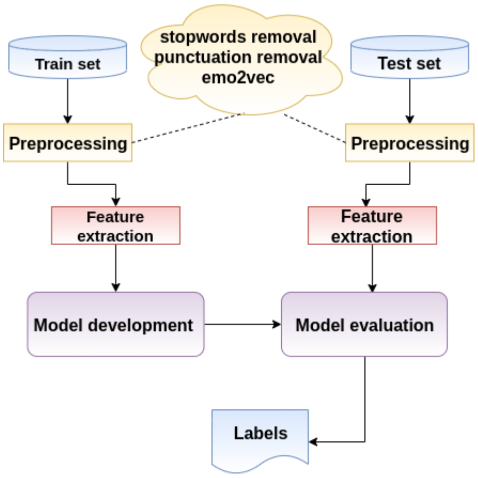

Figure 1: Structure of the ML classifier using Elastic-net

- **mBERT**: is a DL based model pre-trained on a large corpus of multilingual data in a self-supervised manner. mBERT model is fine-tuned using the task specific dataset. The classifier using fine-tuned mBERT model uses "bert" architecture and "bert-base-multilingual-cased" pre-trained model.

The structure of DL classifier using fine-tuned mBERT model is shown in Figure 2.

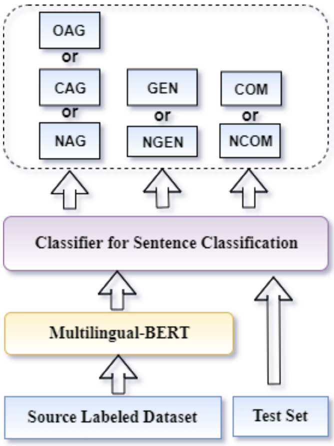

Figure 2: Structure of DL classifier using fine-tuned mBERT

4 Experiments and Results

The three level classification task for each language is as given below:

Table 2: Class-wise distribution of labels in the dataset

Task	Label	Meitei	Bangla	Hindi	Multi lingual
Training set					
Level A	OAG	297	1,274	2,526	4,096
	CAG	1,024	335	800	2,159
	NAG	888	782	1,289	2,959
Level B	GEN	148	902	950	1,999
	NGEN	2,061	1,489	3,665	7,215
Level C	COM	174	304	1,017	1,494
	NCOM	2,035	2,087	3,598	7,720
Development set					
Level A	OAG	159	508	526	1,193
	CAG	471	159	169	797
	NAG	370	333	305	1,007
Level B	GEN	55	369	225	648
	NGEN	945	631	775	2,349
Level C	COM	68	112	196	375
	NCOM	932	888	804	2,622

Table 3: Details of the datasets for the shared task

Language	Train set	Development set	Test set
Hindi	9,214	2,997	2,989
Bangla	2,391	1,000	967
Meitei	2,209	1,000	1,020
Multilingual	9,214	2,997	2,989

- **Level A - Aggression Level:** This is a multi-class classification problem consisting of three labels: 'Overtly Aggressive' (OAG), 'Covertly Aggressive' (CAG), and 'Non-Aggressive' (NAG)
- **Level B - Gender Bias:** This is a binary classification problem consisting of two labels, 'Gendered' (GEN) or 'Non-Gendered (NGEN)
- **Level C - Communal Bias:** This is a binary classification problem consisting of two labels, 'Communal' (COM) or 'Non-Communal' (NCOM)

Table 2 gives the class-wise distribution of labels in the dataset.

Several experiments were conducted using different range of word and char n-grams. Elastic-net penalty is used with 'saga' solver and 0.5 l1-ratio which is used for mixing the ratio of penalties from L1 and L2 regularization.

Training, Development and Test datasets provided by the organizers of the shared task Kumar et al. are shown in Table 3.

The models are evaluated based on instance-F1 and micro-F1 scores. instance-F1 gives an indication of the overall performance of the system while micro-F1 accounts for the partially correct predictions as well. Taken together they give an accurate evaluation of the classifier. Table 4 gives the performance of both the models.

The results clearly indicate that the fine-tuned mBERT model has performed better than Elastic-net model. In both the models, Communal and Gender Biased predictions are better compared to the predictions of Aggression. This problem is primarily due to the high degree of imbalance in the dataset which may lead to overfitting.

The results of the shared task are displayed in the task website[9]. Figure 3 shows the comparison of micro-F1 scores of our models with that of the other top performing models. Our models have shown good performance and are among the top three models in the shared task.

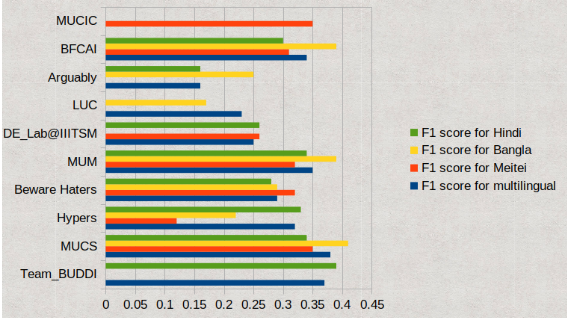

Figure 3: Comparison of instance-F1 scores of the proposed model with top performing models of the shared task

5 Conclusion

This paper describes the models proposed by our team, MUM, to ComMA@ICON shared task at ICON 2021 to identify multilingual gender biased and communal language in Bangla, Hindi, Meitei and multilingual text. By using fine-tuned mBERT and Elastic-net regularization, our team was able to achieve competitive results for all the languages and fine-tuned mBERT model outperformed the other. Future research will examine different sets of features and feature selection models, as well as different approaches for detecting the problematic content.

Table 4: Results of the proposed modesl

mBERT Model				
Evaluation Metrics	Languages			
	Meitei	Bangla	Hindi	Multi lingual
instance-F1	0.326	0.390	0.343	0.359
Overall micro-F1	0.661	0.708	0.691	0.691
Aggression micro-F1	0.426	0.489	0.589	0.508
Gender Bias micro-F1	0.694	0.744	0.783	0.755
Communal Bias micro-F1	0.863	0.892	0.701	0.809
Elastic-net Model				
instance-F1	0.319	0.357	0.312	0.339
Overall micro-F1	0.671	0.708	0.694	0.696
Aggression micro-F1	0.439	0.475	0.587	0.522
Gender Bias micro-F1	0.707	0.762	0.794	0.754
Communal Bias micro-F1	0.866	0.886	0.700	0.812

[9] https://sites.google.com/view/comma-at-icon2021/results?authuser=0

References

M D Anusha and H L Shashirekha. 2020. An Ensemble Model for Hate Speech and Offensive Content Identification in Indo-European Languages. In *FIRE (Working Notes)*, pages 253–259.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Helena Gómez-Adorno, Gemma Bel Enguix, Gerardo Sierra, Octavio Sánchez, and Daniela Quezada. 2018. A machine learning approach for detecting aggressive tweets in spanish. In *IberEval@ SEPLN*, pages 102–107.

Sameer Hinduja and Justin W Patchin. 2010. Bullying, Cyberbullying, and Suicide. volume 14, pages 206–221. Taylor & Francis.

Ritesh Kumar, Guggilla Bhanodai, Rajendra Pamula, and Maheshwar Reddy Chennuru. 2018. Trac-1 shared task on aggression identification: Iit (ism)@ coling'18. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pages 58–65.

Ritesh Kumar, Lahiri Bornini, Bansal Akanksha, Nandi Enakshi, Niranjana Devi Laishram, Ratan Shyam, Singh Siddharth, Bhagat Akash, and Dawer Yogesh. Comma@icon: Multilingual Gender Biased and Communal Language Identification Task at ICON-2021. In *In Proceedings of the 18th Interna- tional Conference on Natural Language Processing (ICON): COMMA@ICON 2021 Shared Task.*

Yujiao Li and Hasan Fleyeh. 2018. Twitter Sentiment Analysis of New Ikea Stores Using Machine Learning. In *2018 International Conference on Computer and Applications (ICCA)*, pages 4–11. IEEE.

Ben J. Marafino, W. John Boscardin, and R. Adams Dudley. 2015. Efficient and Sparse Feature Selection for Biomedical Text Classification via the Elastic Net: Application to ICU Risk Stratification from Nursing Notes. volume 54, pages 114–120.

Guizhe Song, Degen Huang, and Zhifeng Xiao. 2021. A Study of Multilingual Toxic Text Detection Approaches Under Imbalanced Sample Distribution. volume 12, page 205. Multidisciplinary Digital Publishing Institute.

Mircea-Adrian Tanase, George-Eduard Zaharia, Dumitru-Clementin Cercel, and Mihai Dascalu. 2020. Detecting aggressiveness in mexican spanish social media content by fine-tuning transformer-based models. In *IberLEF@ SEPLN*, pages 236–245.

Teodor Tiţa and Arkaitz Zubiaga. 2021. Cross-Lingual Hate Speech Detection using Transformer Models.

Inna Vogel and Peter Jiang. 2019. Bot and Gender Identification in Twitter using Word and Character N-Grams. In *CLEF (Working Notes)*.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A Typology of Abusive Language Detection Subtasks.