ICON 2021: 18th International Conference on Natural Language Processing
National Institute of Technology - Silchar
December 16-19, 2021

Shared Task on Multilingual Gender Biased and Communal Language Identification

PROCEEDINGS

Editors:
Ritesh Kumar, Siddharth Singh, Enakshi Nandi, Shyam Ratan, Laishram Niranjana Devi, Bornini Lahiri, Akanksha Bansal, Akash Bhagat, Yogesh Dawer

Introduction

Aggression and its manifestations in different forms have taken unprecedented proportions with the tremendous growth of Internet and social media. The research community, especially within the fields of Linguistics and Natural Language Processing, has responded by understanding the pragmatic and structural aspects of such forms of language usage and several other) and developing systems that could automatically detect and handle these.

In the ComMA project, we are working on these different aspects of aggressive and offensive language usage online and its automatic identification. As part of our efforts in the project, in this task, we present a novel multi-label classification task to the research community, in which each sample will be required to be classified as aggressive, gender biased or communally charged in 4 languages - Meitei, Bangla, Hindi and English. The motivation behind this is to understand the intersectionality across these three categories and also explore if using intersectional data could be helpful in the task or not.

We attracted a total of 54 registrations in the task, out of which 11 teams submitted their test runs. In this proceedings, we have included system description papers of 8 teams along with a task overview paper. We hope that the task and its findings will be interesting for researchers working in the different related areas of hate speech, offensive language, abusive language as well more generally in text classification.

Shared task page: https://sites.google.com/view/comma-at-icon2021/home

Main conference page: http://icon2021.nits.ac.in/index.html

ComMA@ICON Shared Task Organisers

Organising Committee

Ritesh Kumar, Dr. Bhimrao Ambedkar University, Agra
Bornini Lahiri, Indian Institute of Technology-Kharagpur
Akanksha Bansal, Panlingua Language Processing LLP, New Delhi
Akash Bhagat, Indian Institute of Technology-Kharagpur
Enakshi Nandi, Panlingua Language Processing LLP, New Delhi
Laishram Niranjana Devi, Panlingua Language Processing LLP, New Delhi
Shyam Ratan, Dr. Bhimrao Ambedkar University, Agra
Siddharth Singh, Dr. Bhimrao Ambedkar University, Agra
Yogesh Dawer, Dr. Bhimrao Ambedkar University, Agra

Editors

Ritesh Kumar, Dr. Bhimrao Ambedkar University, Agra
Siddharth Singh, Dr. Bhimrao Ambedkar University, Agra
Enakshi Nandi, Panlingua Language Processing LLP, New Delhi
Shyam Ratan, Dr. Bhimrao Ambedkar University, Agra
Laishram Niranjana Devi, Panlingua Language Processing LLP, New Delhi
Bornini Lahiri, Indian Institute of Technology Kharagpur
Akanksha Bansal, Panlingua Language Processing LLP, New Delhi
Akash Bhagat, Indian Institute of Technology-Kharagpur
Yogesh Dawer, Dr. Bhimrao Ambedkar University, Agra

Table of Contents

Shared Task Program

Sunday December 19, 2021

10:00–11:18 **Paper Session I**
Chair: TBD

10:00–10:13 *ComMA@ICON: Multilingual Gender Biased and Communal Language Identification Task at ICON-2021*
Ritesh Kumar, Shyam Ratan, Siddharth Singh, Enakshi Nandi, Laishram Niranjana Devi, Akash Bhagat, Yogesh Dawer, bornini lahiri and Akanksha Bansal

10:13–10:26 *Team_BUDDI at ComMA@ICON: Exploring Individual and Joint Modelling Approaches for Detecting Aggression, Communal Bias and Gender Bias*
Anand Subramanian, Mukesh Reghu and Sriram Rajkumar

10:26–10:39 *Hypers at ComMA@ICON: Modelling Aggressive, Gender Bias and Communal Bias Identification*
Sean Benhur, Roshan Nayak, Kanchana Sivanraju, Adeep Hande, CN Subalalitha, Ruba Priyadharshini and Bharathi Raja Chakravarthi

10:39–10:52 *Beware Haters at ComMA@ICON: Sequence and Ensemble Classifiers for Aggression, Gender Bias and Communal Bias Identification in Indian Languages*
Deepakindresh Gandhi, Aakash Ambalavanan, Avireddy Rohan and Radhika Selvamani

10:52–11:05 *DELab@IIITSM at ICON-2021 Shared Task: Identification of Aggression and Biasness Using Decision Tree*
Maibam Debina and Navanath Saharia

11:05–11:18 *LUC at ComMA-2021 Shared Task: Multilingual Gender Biased and Communal Language Identification without Using Linguistic Features*
Rodrigo Cuéllar-Hidalgo, Julio de Jesús Guerrero-Zambrano, Dominic Forest, Gerardo Reyes-Salgado and Juan-Manuel Torres-Moreno

11:18–11:28 ***Break***

Sunday December 19, 2021 (continued)

11:28–12:33 Paper Session II
Chair: TBD

11:28–11:41 *ARGUABLY at ComMA@ICON: Detection of Multilingual Aggressive, Gender Biased, and Communally Charged Tweets Using Ensemble and Fine-Tuned IndicBERT*
Guneet Kohli, Prabsimran Kaur and Jatin Bedi

11:41–11:54 *Sdutta at ComMA@ICON: A CNN-LSTM Model for Hate Detection*
Sandip Dutta, Utso Majumder and Sudip Naskar

11:54–12:07 *MUCIC at ComMA@ICON: Multilingual Gender Biased and Communal Language Identification Using N-grams and Multilingual Sentence Encoders*
Fazlourrahman Balouchzahi, Oxana Vitman, Hosahalli Lakshmaiah Shashirekha, Grigori Sidorov and Alexander Gelbukh

12:07–12:20 *MUM at ComMA@ICON: Multilingual Gender Biased and Communal Language Identification Using Supervised Learning Approaches*
Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho and Hosahalli Lakshmaiah Shashirekha

12:20–12:33 *BFCAI at ComMA@ICON 2021: Support Vector Machines for Multilingual Gender Biased and Communal Language Identification*
Fathy Elkazzaz, Fatma Sakr, Rasha Orban and Hamada Nayel

ComMA@ICON: Multilingual Gender Biased and Communal Language Identification Task at ICON-2021

Ritesh Kumar[æ], Shyam Ratan[æ], Siddharth Singh[æ], Enakshi Nandi[œ], Laishram Niranjana Devi[œ], Akash Bhagat[ø], Yogesh Dawer[æ], Bornini Lahiri[ø], Akanksha Bansal[œ]

[æ]Dr. Bhimrao Ambedkar University, Agra [ø]Indian Institute of Technology - Kharagpur
[œ]Panlingua Language Processing LLP, New Delhi
comma.kmi@gmail.com

Abstract

This paper presents the findings of the ICON-2021 shared task on Multilingual Gender Biased and Communal Language Identification, which aims to identify aggression, gender bias, and communal bias in data presented in four languages: Meitei, Bangla, Hindi and English. The participants were presented the option of approaching the task as three separate classification tasks or a multi-label classification task or a structured classification task. If approached as three separate classification tasks, the task includes three sub-tasks: aggression identification (sub-task A), gender bias identification (sub-task B), and communal bias identification (sub-task C).

For this task, the participating teams were provided with a total dataset of approximately 12,000, with 3,000 comments across each of the four languages, sourced from popular social media sites such as YouTube, Twitter, Facebook and Telegram and the the three labels presented as a single tuple. For the test systems, approximately 1,000 comments were provided in each language for every sub-task. We attracted a total of 54 registrations in the task, out of which 11 teams submitted their test runs.

The best system obtained an overall instance-F1 of 0.371 in the multilingual test set (it was simply a combined test set of the instances in each individual language). In the individual sub-tasks, the best micro f1 scores are 0.539, 0.767 and 0.834 respectively for each of the sub-task A, B and C. The best overall, averaged micro f1 is 0.713.

The results show that while systems have managed to perform reasonably well in individual sub-tasks, especially gender bias and communal bias tasks, it is substantially more difficult to do a 3-class classification of aggression level and even more difficult to build a system that correctly classifies everything right. It is only in slightly over 1/3 of the instances that most of the systems predicted the correct class across the board, despite the fact that there was a significant overlap across the three sub-tasks.

1 Introduction

The global reach of digital technology has resulted in the spread of social media applications to every section of society, making it a major medium of interaction for all kinds of people across the globe. Social media sites have, as a result, become significant documents of human discourse for the digital age. Social media discourse covers a broad spectrum and can be culturally and socio-politically specific to the region and people who engage in it, while also having a common grammar of form and content which have adapted to suit the platforms they appear in. A prime feature of social media discourse that has gained a lot of traction in the last few years is hate speech and aggression rooted in bias and prejudice. It manifests in the form of trolling, cyberbullying, flaming, and so on, and can have real-life consequences that are harmful, dangerous, and sometimes even fatal (Kumar et al., 2018b).

The ComMA project aims to limit the negative effects of such comments on social media sites by developing a system that is trained to identify and isolate comments from social media platforms that display aggression and bias towards the target's gender and religious identities and beliefs. As part of our efforts in the project, we present this novel multi-label classification task to the research community, in which each sample will be required to be classified as aggressive, gender biased or communally charged. We expect that the task will be interesting for researchers working in the different related areas of hate speech, offensive language, abusive language as well more generally in text classification.

Proceedings of the 18th International Conference on Natural Language Processing: Shared Task on Multilingual Gender Biased and Communal Language Identification, pages 1–12
National Institute of Technology - Silchar, December 16-19, 2021.

2 Related Work

Automatically identifying the various forms of abusive language online has been studied from different angles. Examples include trolling (Cambria et al., 2010; Kumar et al., 2014; Mojica, 2016; Mihaylov et al., 2015), flaming/insults (Sax, 2016; Nitin et al., 2012), radicalization (Agarwal and Sureka, 2015, 2017), racism (Greevy and Smeaton, 2004; Greevy, 2004), misogyny (Menczer et al., 2015; Frenda et al., 2019; Hewitt et al., 2016; Fersini et al., 2018; Anzovino et al., 2018; Sharifirad and Matwin, 2019), online aggression (Kumar et al., 2018a), cyberbullying (Xu et al., 2012; Dadvar et al., 2013), hate speech (Kwok and Wang, 2013; Djuric et al., 2015; Burnap and Williams, 2015; Davidson et al., 2017; Malmasi and Zampieri, 2017, 2018) and offensive language (Wiegand et al., 2018; Zampieri et al., 2019a). The terms used in the literature have overlapping properties as discussed in (Waseem et al., 2017) and (Zampieri et al., 2019a).

Most related studies focus on English, but a significant amount of work has been carried out for other languages too. This includes languages such as Arabic (Mubarak et al., 2020), German (Struß et al., 2019), Greek (Pitenis et al., 2020), Hindi (Mandl et al., 2019a), and Spanish (Basile et al., 2019).

The field has also seen a rapid development and availability of multiple datasets in multiple languages via various shared tasks and competitions, This shared task is one of many shared tasks that are being organised in this area, which include (Kumar et al., 2020, 2018a; Zampieri et al., 2019a,b; Mandl et al., 2019b, 2020a,b, 2021; Modha et al., 2021).

Among these, one of the most popular tasks, OffensEval series of tasks (Zampieri et al., 2019b, 2020), focused on offensive language identification and featured three sub-tasks: offensive language identification, offensive type identification, and offense target identification building on the annotation model introduced in the OLID dataset (Zampieri et al., 2019a) for English. This multiple sub-task model has been adopted by other shared tasks such as GermEval for German (Struß et al., 2019), HASOC for English, German, and Hindi (Mandl et al., 2019a), and HatEval for English and Spanish (Basile et al., 2019).

The tasks most similar to the current one were the TRAC - 1 and TRAC - 2 shared tasks. TRAC - 1 shared task on Aggression Identification (Kumar et al., 2018a) was hosted at the TRAC workshop at COLING 2018. It included English and Hindi data from Facebook and Twitter. It consisted of a three-way classification task with posts labelled as overtly aggressive, covertly aggressive, and non-aggressive. TRAC - 2 (Kumar et al., 2020) featured data from 3 languages - Bangla, Hindi and Ebglish - and included an additional sub-task of misogyny identification. The present task has been conceptualised as an extension of the TRAC-2 shared task, with more languages and an addition sub-task. Moreover, it is now also reformulated as a structured prediction task, along with three separate text classification tasks, to encourage teams towards leveraging the benefits of a multi-task setup in a largely overlapping setup.

3 Task Schedule and Setup

Participants for the present shared task were allowed to participate in one of the four languages - Meitei, Bangla, Hindi, or Multilingual - or all of them but they were required to submit predictions for all three subtasks (A, B and C). The English data is not provided separately and is included in the data of all the languages. Registered participants got dataset (training, development and test set) for training and evaluation in all languages through the Codalab platform [1].

For the task, the participants were given around 4 weeks to experiment and develop the systems. After 4 weeks of releasing the train and development sets, the test set was released, after which the participants had 6 days to test and upload their systems. The entire timeline and schedule of the shared task is given in Table 1.

Date	Event
October 2, 2021	Training set release
November 3, 2021	Test set release
November 8, 2021	System submissions
November 14, 2021	Result announcement
November 24, 2021	System description paper
November 29, 2021	Reviews for papers
December 2, 2021	Camera-ready versions

Table 1: Timeline and schedule of the Multilingual Gender Biased and Communal Language Identification Shared Task at ICON - 18, 2021

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

In the evaluation phase, each team was permitted to submit up to 5 systems and their best run was included in the final ranking presented in this paper.

4 Dataset

We provided a multilingual dataset with a total of over 15,000 samples for training, development and testing in four languages: Meitei, Bangla, Hindi, and English. The dataset was marked at three levels: aggression, gender bias, and communal bias. Each level was represented in the form of an individual sub-task:

1. **Sub-task A: Aggression Identification**
   The task here was to develop a classifier that could make a 3-way classification between 'Overtly Aggressive' (OAG), 'Covertly Aggressive' (CAG), and 'Non-aggressive' (NAG) text data.

2. **Sub-task B: Gender Bias Identification**
   This task required the participants to develop a binary classifier to classify the text as 'gendered' (GEN) or 'non-gendered' (NGEN).

3. **Sub-task C: Communal Bias Identification**
   This task required the participants to develop a binary classifier to classify the text as 'communal' (COM) or 'non-communal'(NCOM).

The participants were allowed to approach the task either as three separate classification tasks, or a multi-label classification task, or one structured classification task.

The process of developing dataset used for the task has been discussed in detail in (Kumar et al., 2021).

4.1 Training Set

The training dataset contains a total of 12,211 comments from YouTube, Twitter, and Facebook in four languages: Meitei (Mni), Bangla (Ban), Hindi (Hi), and English (En) apart from Multilingual. A class-wise distribution of the test dataset is represented in Table 2.

4.2 Test Set

The test set consisted of a total of 2,989 comments from YouTube, Telegram, and Twitter in four languages: Meitei (Mni), Bangla (Ban), Hindi (Hi), and English (En) aprat from Multilingual. A class-wise distribution of the test dataset is represented in Table 3.

	Aggression			
	TOTAL	**OAG**	**CAG**	**NAG**
Mni	**3,209**	456	1,495	1,258
Ban	**3,391**	1,782	494	1,115
Hi	**5,615**	3,052	969	1,594
Multi	**12,211**	5,289	2,956	3,966
	Gendered			
	TOTAL	**GEN**	**NGEN**	
Mni	**3,209**	203	3,006	
Ban	**3,391**	1,271	2,120	
Hi	**5,615**	1,175	4,440	
Multi	**12,211**	2,647	9,564	
	Communal			
	TOTAL	**COM**	**NCOM**	
Mni	**3,209**	242	2,967	
Ban	**3,391**	416	2,975	
Hi	**5,615**	1,213	4,402	
Multi	**12,211**	1,869	10,342	

Table 2: Classwise Distribution of The ICON Training Dataset

	Aggression			
	TOTAL	**OAG**	**CAG**	**NAG**
Mni	**1,020**	315	391	314
Ban	**967**	465	244	258
Hi	**1,002**	440	85	477
Multi	**2,989**	1,220	720	1,049
	Gendered			
	TOTAL	**GEN**	**NGEN**	
Mni	**1,020**	317	703	
Ban	**967**	303	664	
Hi	**1,002**	204	798	
Multi	**2,989**	824	2,165	
	Communal			
	TOTAL	**COM**	**NCOM**	
Mni	**1,020**	141	879	
Ban	**967**	106	861	
Hi	**1,002**	362	640	
Multi	**2,989**	609	2,380	

Table 3: Classwise Distribution of The ICON Test Dataset

5 Participants and Approaches

A total of 54 teams registered for this shared task, with most of the teams registering to participate in all the languages. By design, all the teams were required to participate in all the three tracks. Finally a total of 11 teams submitted their systems

- out of these, 8 teams have been included in the official rankings while the other 3 are not because of delayed submission on their part - however they were also evaluated and are discussed here. All the 11 teams that submitted their system were invited to submit the system description paper, describing the their models and experiments conducted by them. The name of the participating teams and the language they participated in are given in Table 4. We give a brief description of the approaches used by each team for building their system. A detailed description of the approaches could be found in the paper submitted by each team. We give a brief summary of each team's system below -

- **Team_BUDDI** utilises two BERT-based models - one that was fine-tuned using Hindi-English code-mixed tweets for a language modelling task (for the Hindi dataset) and an XLM-RoBERTa model for the multilingual dataset. They fine-tuned the two models for individual sub-tasks as well as jointly for all the sub-tasks and demonstrate that joint modelling of the different sub-tasks perform better than the individual modelling.

- **Hypers** fine-tuned MURIL for Hindi, Meitei and Multilingual datasets and BanglaBERT for Bangla dataset. They used two custom poolers - attention pooler and mean-pooler. Except for Hindi data, in all other instances, attention-pooler has outperformed the mean-pooler.

- Team **Beware Haters** experimented with various kinds of models including Random Forest, Logistic Regression, SVM, Bi-LSTM and an ensemble of Random Forest, Logistic Regression and SVM. While Bi-LSTM worked well for the two binary classification tasks using multilingual dataset, Logistic Regression and the ensemble worked well for different monolingual test sets - this is expected given the fact that multilingual dataset is large enough for Bi-LSTM to generalise well.

- **DE_Lab@IIITSM** experimented with an enriched pre-processing step followed by using Decision Tree classifiers for the task.

- Team **LUC** experimented with multiple linear classifiers incl KNN, Naive Bayes, SVM, Random Forest, GBM, Adaboost and Neural networks. KNN with K = 1 was their best-performing model.

- Team **Arguably** experimented with two approaches - (a) Boosted Voting Ensembler of XGBOOST, LightGBM and Naive Bayes and (b) a fine-tuned IndicBERT model (which is an ALBERT model pre-trained on Indian languages). Among these the Ensembler outperformed or performed comparably to the IndiBERT model across all sub-tasks and languages.

- **sdutta** used a CNN-LSTM based model for prediction.

- **MUCIC** trained three classifiers: SVM, Random Forest and Logistic Regression using a combination of word and character n-grams, along with vectors from multilingual sentence encoder. They used two techniques of pre- and post-aggregation of labels.

- **MUM** uses two models - (a) Elastic-net trained on combination of word unigram character ngrams TF-IDF values, combined with the pre-trained Emo2Vec vector embeddings and (b) a multilingual BERT (mBERT) fine-tuned for the task. The mBERT model has given better results for all languages and all the sub-tasks.

- **BFCAI** has experimented with 4 different classifiers - SVM, simple linear classifier, Multilayer perceptron, Multinomial Naive Bayes and an ensemble of these classifiers.

6 Evaluation and Results

The systems have been evaluated on the basis of the following metrics -

- **instance F1:** It is the F-measure averaging on each instance in the test set i.e. the classification was considered right only when all the labels in a given instance are predicted correctly. It was the primary evaluation metric for the task and used for ranking the systems.

- **micro F1**: It gives a weighted average score of each class and is generally considered a good metric in cases of class-imbalance. Also it shows the performance of each system on individual sub-tasks.

Team	Meitei	Bangla	Hindi	Multilingual	System Description Paper
Team_BUDDI			✓	✓	(Subramanian et al., 2021)
Hypers	✓	✓	✓	✓	(Benhur et al., 2021)
Beware Haters	✓	✓	✓	✓	(Gandhi et al., 2021)
DE_Lab@IIITSM	✓		✓	✓	(Debina and Saharia, 2021)
LUC		✓		✓	(Cuéllar-Hidalgo et al., 2021)
Arguably			✓	✓	(Kohli et al., 2021)
sdutta	✓	✓	✓	✓	(Dutta et al., 2021)
MUCIC	✓	✓	✓	✓	(Balouchzahi et al., 2021)
MUCS	✓	✓	✓	✓	
MUM	✓	✓	✓	✓	(Hegde et al., 2021)
BFCAI	✓	✓	✓	✓	(Elkazzaz et al., 2021)
Total	**8**	**8**	**10**	**11**	**10**

Table 4: Teams participated in the Multilingual Gender Biased and Communal Language Identification Shared Task at ICON-2021.

The system results of each team for Meitei, Bangla, Hindi and Multilingual have been considered in two ways: system submissions within the deadline of the shared task and submissions after the deadline. The results of both have been presented in Tables 5[2] and 6. Language-wise, the best system obtained a weighted instance F1-score of approximately 0.322 for Meitei, 0.292 for Bangla, 0.398 for Hindi and 0.371 for multilingual. Overall, the highest instance F1-score is obtained for Bangla i.e. 0.398. For the score evaluation, apart from the instance F1-score, the overall micro-F1 is also calculated. It is also calculated of each system for all languages.

7 Error Analysis

We carried out an overall analysis of the errors generated by all the systems submitted for the task. This was done with an aim to understand the most difficult instances to classify. In this error analysis, we have analysed only those instances which have been classified wrongly by 'all' the models for sub-task A and those which have been clasified wrongly by at least '$\frac{3}{4}$' of all models in case of sub-task B and C [3] in all languages. A summary of the errors generated by the systems on the test data in all the languages have been presented below under "error types". Language wise error counts and error type counts in all sub-tasks are given in Tables 7 and 8 and Figure 1. We identified the recurring patterns that generate these errors and classified them as follows:

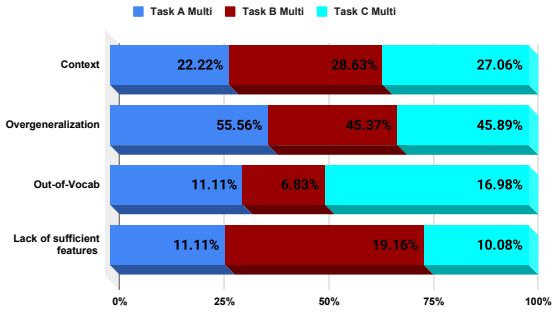


Figure 1: Error types proportion in each sub-task

- **Context:** Contextual errors occur when there is a mismatch between the gold and predicted labels of a comment based on whether or not the annotator or the system has taken into account the discursive context in which the comment exists. Such a context can include the contents of the video or post under which the comments are written, the other comments that are in conversation with or appear alongside the given comment, and the socio-political context in which certain content and comments find expression. The comments that have generated context based errors in this shared task include sarcastic or satirical comments, ambiguous comments (that can be legitimately labelled with more than one tag), and replies to previous comments (in the sense that they could be correctly classified only by

[2]These teams submitted systems after the deadline of shared task, which is why they have not been considered in the final ranking.

[3]this is so because we did not find any instance in these two sub-tasks which have been wrongly classified by all the models submitted for the task

Team	Meitei			Bangla			Hindi			Multilingual		
	Rank	Inst F1	Micro F1	Rank	Inst F1	Micro F1	Rank	Inst F1	Micro F1	Rank	Inst F1	Micro F1
Team_BUDDI	-	-	-	-	-	-	1	**0.398**	**0.709**	1	**0.371**	**0.713**
Hypers	3	0.129	0.472	2	0.223	0.579	2	0.336	0.683	2	0.322	0.685
Beware Haters	1	**0.322**	**0.672**	1	**0.292**	**0.704**	3	0.289	0.689	3	0.294	0.665
DE_Lab@IIITSM	2	0.267	0.625	-	-	-	4	0.263	0.629	4	0.258	0.632
LUC	-	-	-	3	0.17	0.597	-	-	-	5	0.234	0.615
Arguably	-	-	-		-	-	5	0.161	0.582	6	0.156	0.583
sdutta	4	0.007	0.279	4	0.006	0.294	6	0.047	0.335	7	0.02	0.288
MUCIC	5	0	0.69	5	0	0.723	7	0	0.697	8	0	0.701
MUCS	NA	0.35	0.681	NA	0.412	0.718	NA	0.341	0.706	NA	0.38	0.705
MUM	NA	0.326	0.661	NA	0.39	0.708	NA	0.343	0.691	NA	0.359	0.691
BFCAI	NA	0.317	0.664	NA	0.391	0.695	NA	0.304	0.678	NA	0.342	0.671

Table 5: Performance of teams on Meitei, Bangla, Hindi & Multilingual Dataset

Team	Meitei			Bangla			Hindi			Multilingual		
	Task A	Task B	Task C	Task A	Task B	Task C	Task A	Task B	Task C	Task A	Task B	Task C
Team_BUDDI	-	-	-	-	-	-	**0.628**	0.743	**0.757**	**0.539**	**0.767**	**0.834**
Hypers	0.372	0.609	0.435	0.434	0.674	0.63	0.555	0.784	0.709	0.519	0.715	0.822
Beware Haters	0.454	0.697	0.865	0.499	0.72	**0.895**	0.603	0.783	0.68	0.482	0.722	0.791
DE_Lab@IIITSM	0.344	0.682	0.849	-	-	-	0.479	0.726	0.682	0.413	0.694	0.791
LUC	-	-	-	0.368	0.561	0.861	-	-	-	0.446	0.675	0.726
Arguably	-	-	-	-	-	-	0.402	0.702	0.642	0.359	0.612	0.776
sdutta	0.388	0.311	0.138	0.438	0.339	0.107	0.44	0.204	0.361	0.376	0.281	0.208
MUCIC	**0.484**	**0.716**	**0.871**	**0.509**	**0.772**	0.89	0.606	**0.801**	0.683	0.534	0.764	0.806
MUCS	0.462	0.713	0.868	0.517	0.746	0.89	0.62	0.808	0.69	0.54	0.759	0.816
MUM	0.426	0.694	0.863	0.489	0.744	0.892	0.589	0.783	0.701	0.508	0.755	0.809
BFCAI	0.438	0.692	0.862	0.516	0.679	0.89	0.568	0.799	0.668	0.472	0.752	0.788

Table 6: Performance of teams in all sub-tasks on Meitei, Bangla, Hindi & Multilingual Dataset

	Task A	Task B	Task C
Mni	115	252	108
Ban	65	116	85
Hi	207	86	184
Multi	387	454	377

Table 7: Error counts in all sub-tasks by all teams

taking into account the previous comment(s)). Let us take a look the following examples of this kind of error -

1. Sahi baat hai iska 7 khoon to janm se maaf hai [Hindi]

   Translation: You're right, this person can get away with anything

   Gold label: GEN

   Predicted label: NGEN

   Explanation: This comment was made about a beautiful woman who had committed a mistake. The gold label is GEN because in the context of the conversation it is a gendered comment. However, the systems predict it as NGEN because they do not have access to or an understanding of that context, and the textual content itself does not indicate it is a gendered comment in any way.

2. #justiceforhindus #SaveBangladeshiHindus Boycott the budget speech [English]

   Gold label: COM

   Predicted label: NCOM

   Explanation: This comment was made in the context of some communally charged incidents that took place in Bangladesh in October 2021. The gold label is COM on the basis of that context, but the predicted label is NCOM because the systems do not have access to or an understanding of that context.

3. Ron Haokip oiram mani. Dance touba nupise thadou kuki ne. [Meitei]

   Translation: Ron might be Haokip. The girl dancing belongs to thadou kuki.

   Gold label: GEN

   Predicted label: NGEN

   Explanation: This comment was made in the context of a dance video. The gold label is GEN because in the context of the conversation it looks at girls as being a "property" of the boys of her own community. How-

	Task A			
	Context	**Overgeneralization**	**Out-of-Vocabulary**	**Lack of sufficient features**
Mni	10	95	4	6
Ban	28	15	-	22
Hi	48	105	39	15
Multi	86	215	43	43
	Task B			
	Context	**Overgeneralization**	**Out-of-Vocabulary**	**Lack of sufficient features**
Mni	52	156	20	24
Ban	55	19	1	41
Hi	23	31	10	22
Multi	130	206	31	87
	Task C			
	Context	**Overgeneralization**	**Out-of-Vocabulary**	**Lack of sufficient features**
Mni	54	29	16	9
Ban	26	47	-	12
Hi	22	97	48	17
Multi	102	173	64	38

Table 8: Language wise error type counts in each sub-task

ever, most of the systems predict it as NGEN because the sentence could be interpreted as a simple description of the identities out of the specific context.

- **Overgeneralization:** This kind of error occurs when the system overfits or overgeneralizes for certain linguistic features. In the bilingual Bangla-English Twitter data, the systems have frequently mispredicted the tags for communal and non-communal because they could not distinguish between political parties and religions, and region/nation and religion. Some other categories that the system could not distinguish between include caste vs religious identity, caste vs gender identity, religious vs gender identity, and personal vs group identity. Let us take a look at the following examples to understand this -

  1. mndir ko english mein bhi Mandir hi likhna chahiyada odd lagta hai temple [Hindi]

     Translation: Mandir (temple) should have been written as "Mandir" in English as well; temple sounds odd

     Gold label: NAG

     Predicted label: OAG

     Explanation: The error in this example arises from the mention of "mandir" or temple, which is a religious symbol. In this dataset, it has been noted that comments with words like 'temple' often are overtly or covertly aggressive in nature. As a result, the mere mention of temple in a comment has prompted the systems to overgeneralize and predict OAG as the aggression label for this comment.

- **Out-of-Vocabulary Error:** This error occurs because there are new words (often abusive, aggressive, sexist, or Islamophobic) that are coined by the commenters which are frequently mispredicted, because the systems do not recognize them from the training data and hence cannot label them as abuse, as they must.

  1. dadhivala topivala pancharputra katva suar ammichod betichod behanchod bakrichod haalaa ki aulaad Terrorists aur koi naam hai to btaao [Hindi]

     Translation: dadhivala topivala pancharputra katva[4], pig, motherfucker, daughterfucker, sisterfucker, goat-

[4]Islamophobic slurs

fucker, son of halala, terrorists - Are there any more names for them?
Gold label: COM
Predicted label: NCOM
Explanation: This comment contains some coined lexical items (pancharputra, topivala) that are Islamophobic in nature. However, since they were not part of the training set, the systems do not recognize them and are, hence, mispredicting the labels.

2. Gay jao yam yaoreye [Meitei]
Translation: many gay-jao (coined word meaning 'master of all gay') are present here.
Gold label: GEN
   (a) **Predicted label:** NGEN
   Explanation: The comment contains coined word 'gay-jao' which is sexist in nature but the system mispredicts it as NGEN.

- **Lack of sufficient features:** In certain cases the errors generated by the system are due to the fact that the comments are generic, incomplete, contain only emojis, or lack sufficient features that the system can identify to generate an accurate label. For instance, a comment as simple as "Hello" or "Thank you" or "Hm" has generated results for both gender bias and non-gendered bias. Such is also the case for religious or political slogans such as "Jai Shri Ram" or "Jai Hari bol", and emojis which may be labelled as CAG, NAG, or OAG by different systems based on different criteria. The systems also generate different results for specific lexical items in the data such as curse words or abusive words. This can be attributed to the fact that some systems take the etymology of the lexical items into account, which can be sexist at their core, while others treat them like words which have been bleached of their literal meaning or denotation.

1. @Sania Parvin oi je
   (a) **Translation:** @Sania Parvin that
   (b) **Gold label:** COM
   (c) **Predicted label:** NCOM
   (d) **Explanation:** This error is due to an incomplete comment which has been labelled COM based on its context in the gold set. However, many systems have labelled it NCOM because it does not contain sufficient features by which it could be assigned an appropriate label.
2. Allah madarchod hai yaar [Hindi]
Translation: Allah is motherfucker
Gold label: COM
Predicted label: NCOM
Explanation: This comment contains abuse that is aggressive, sexist, and Islamophobic. However, the systems have predicted the wrong labels for it, possibly, because there were not sufficient co-textual features to predict it correctly.
3. jaroj santan
Translation: Illegitimate child
Gold label: GEN
Predicted label: NGEN
   (a) **Explanation:** This comment contains a gendered abuse but many systems have labelled it as non-gendered, again, because the comment is too short to give a reliable judgement.
4. Porn film kumbi hek maladana [Meitei]
Translation: You definitely look like a porn actress
Gold label: GEN
   (a) **Predicted label:** NGEN
   Explanation: The comment targets character of a women by using such lexical items but most of the system mis-predicts it as NGEN - this could again be possibly because it is too short to provide sufficient features for correct prediction.

In all such cases of misprediction possibly because of there being too little features, some kind of data augmentation techniques or taking into consideration the sequence (of comments) or context might prove to be helpful.

8 Closing remarks

In this paper, we have presented the results of the shared task on automatic identification of aggressive language, gender bias and communal polarisation. The results show that while it is relatively

easier to get prediction on one of these categories right, it is still a very difficult task to predict all of these right for a single instance - the best team managed to get an instance F1 of only 0.371. However at the same time, we also see that the best result across all models and all teams is attained by a model that is jointly trained for all the sub-tasks and all the languages - this shows the value of multi-task and multilingual learning in low-resource situations. The second major takeaway related to the models is that ensemble of well-tuned linear classifiers are also useful for tasks like these and we see that one of the systems in top-3 is an ensemble system. In other instances as well, ensembles have proved to be better than or equivalent to the Transformers-based systems.

In terms of the model performance (and also reliability of the dataset), a comprehensive error analysis of the models submitted for the task show that a huge majority of the errors made by all the model relates to the generalisability of the models, manifested in terms of overfitting for certain linguistic features and inability of the models to perform well on data outside of the training set domain. This could be attributed to two possible reasons -

1. Lack of sufficient datapoints for system to generalise well - this could improved by augmenting the dataset with more instances.

2. Lack of sufficient diversity in the dataset - again this could be improved by augmenting the dataset with more instances. However, a more careful selection of the datapoints is essential such that the linguistic items which are not directly related to these classes (for example name of specific political parties or politicians) are proportionately distributed across different classes. This will also aid in building a dataset which is not biased towards specific entities and is representative of the phenomena under study.

In addition to this, the other most common source of error is the lack of contextual knowledge in the way dataset is presented and the way models are trained. This could be improved only by providing explicit contextual information in the dataset and also for models to take into consideration those information. We plan to make this available in the next version of the dataset.

Acknowledgments

We would like to thank all the teams for participating in the ICON-2021 shared task.

References

Swati Agarwal and Ashish Sureka. 2015. Using knn and svm based one-class classifier for detecting online radicalization on twitter.

Swati Agarwal and Ashish Sureka. 2017. Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems*.

Fazlourrahman Balouchzahi, Oxana Vitman, Hosahalli Lakshmaiah Shashirekha, Grigori Sidorov, and Alexander Gelbukh. 2021. Mucic at comma@icon: Multilingual gender biased and communal language identification using n-grams and multilingual sentence encoders. In *Proceedings of the 18th International Conference on Natural Language Processing: Shared Task on Multilingual Gender Biased and Communal Language Identification*, pages 58–63, NIT Silchar. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of SemEval*.

Sean Benhur, Roshan Nayak, Kanchana Sivanraju, Adeep Hande, CN Subalalitha, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021. Hypers at comma@icon: Modelling aggressive, gender bias and communal bias identification. In *Proceedings of the 18th International Conference on Natural Language Processing: Shared Task on Multilingual Gender Biased and Communal Language Identification*, pages 21–25, NIT Silchar. Association for Computational Linguistics.

Pete Burnap and Matthew L. Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2).

Erik Cambria, Praphul Chandra, Avinash Sharma, and Amir Hussain. 2010. *Do not feel the trolls*. ISWC, Shanghai.

Rodrigo Cuéllar-Hidalgo, Julio de Jesús Guerrero-Zambrano, Dominic Forest, Gerardo Reyes-Salgado, and Juan-Manuel Torres-Moreno. 2021. Luc at

comma-2021 shared task: Multilingual gender biased and communal language identification without using linguistic features. In *Proceedings of the 18th International Conference on Natural Language Processing: Shared Task on Multilingual Gender Biased and Communal Language Identification*, pages 41–45, NIT Silchar. Association for Computational Linguistics.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. pages pp 693–696.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language.

Maibam Debina and Navanath Saharia. 2021. Delab@iiitsm at icon-2021 shared task: Identification of aggression and biasness using decision tree. In *Proceedings of the 18th International Conference on Natural Language Processing: Shared Task on Multilingual Gender Biased and Communal Language Identification*, pages 35–40, NIT Silchar. Association for Computational Linguistics.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of WWW*.

Sandip Dutta, Utso Majumder, and Sudip Naskar. 2021. sdutta at comma@icon: A cnn-lstm model for hate detection. In *Proceedings of the 18th International Conference on Natural Language Processing: Shared Task on Multilingual Gender Biased and Communal Language Identification*, pages 53–57, NIT Silchar. Association for Computational Linguistics.

Fathy Elkazzaz, Fatma Sakr, Rasha Orban, and Hamada Nayel. 2021. Bfcai at comma@icon 2021: Support vector machines for multilingual gender biased and communal language identification. In *Proceedings of the 18th International Conference on Natural Language Processing: Shared Task on Multilingual Gender Biased and Communal Language Identification*, pages 70–74, NIT Silchar. Association for Computational Linguistics.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*.

Simona Frenda, Bilal Ghanem, Manuel Montes-y Gomez, and Paolo Rosso. 2019. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5).

Deepakindresh Gandhi, Aakash Ambalavanan, Avireddy Rohan, and Radhika Selvamani. 2021. Beware haters at comma@icon: Sequence and ensemble classifiers for aggression, gender bias and communal bias identification in indian languages. In *Proceedings of the 18th International Conference on Natural Language Processing: Shared Task on Multilingual Gender Biased and Communal Language Identification*, pages 26–34, NIT Silchar. Association for Computational Linguistics.

Edel Greevy. 2004. *Automatic text categorisation of racist webpages*. Ph.D. thesis, Dublin City University.

Edel Greevy and Alan Smeaton. 2004. Classifying racist texts using a support vector machine. Sheffield, U.K. SIGIR 2004 - the 27th Annual International ACM SIGIR Conference.

Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, and Hosahalli Lakshmaiah Shashirekha. 2021. Mum at comma@icon: Multilingual gender biased and communal language identification using supervised learning approaches. In *Proceedings of the 18th International Conference on Natural Language Processing: Shared Task on Multilingual Gender Biased and Communal Language Identification*, pages 64–69, NIT Silchar. Association for Computational Linguistics.

Sarah Hewitt, Thanassis Tiropanis, and C. Bokhove. 2016. The problem of identifying misogynist language on twitter (and other online social spaces). In *Proceedings of the 58th ACM Conference on Web Science, WebSci '16*.

Guneet Kohli, Prabsimran Kaur, and Jatin Bedi. 2021. Arguably at comma@icon: Detection of multilingual aggressive, gender biased, and communally charged tweets using ensemble and fine-tuned indicbert. In *Proceedings of the 18th International Conference on Natural Language Processing: Shared Task on Multilingual Gender Biased and Communal Language Identification*, pages 46–52, NIT Silchar. Association for Computational Linguistics.

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

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018a. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating aggression identification in social media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 1–5, Marseille, France. European Language Resources Association (ELRA).

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018b. Aggression-annotated corpus of Hindi-English code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Srijan Kumar, Francesca Spezzano, and V.S. Subrahmanian. 2014. Accurately detecting trolls in slashdot zoo via decluttering.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Proceedings of AAAI*.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting hate speech in social media. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 467–472, Varna, Bulgaria. INCOMA Ltd.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech.

Thomas Mandl, Sandip Modha, M. AnandKumar, and Bharathi Raja Chakravarthi. 2020a. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohna Dave, Chintak Mandlia, and Aditya Patel. 2019a. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation (FIRE)*.

Thomas Mandl, Sandip Modha, Daksh Patel, Mohna Dave, Chintak Mandlia, and Aditya Patel. 2019b. Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages). In *Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation*.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Amit Jaiswal, Durgesh Nandini, Daksh Patel, Prasenjit Majumder, and Johannes Schäfer. 2020b. Overview of the hasoc track at fire 2020: Hate speech and offensive content identification in indo-european languages. In *FIRE 2020: Forum for Information Retrieval Evaluation, Virtual Event, 16th-20th December 2020*. ACM.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schäfer, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini, and Amit Kumar Jaiswal. 2021. 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.

Filippo Menczer, Rachael Fulper, Giovanni Luca Ciampaglia, Emilio Ferrara, Yong-Yeol Ahn, Alessandro Flammini, Bryce Lewis, and Kehontas Rowe. 2015. Misogynistic language on twitter and sexual violence. In *Proceedings of the ACM Web Science Workshop on Computational Approaches to Social Modeling (ChASM)*.

Todor Mihaylov, Georgi D. Georgiev, A. Ontotext, and Nakov Preslav. 2015. Finding opinion manipulation trolls in news community forums. CoNLL.

Sandip Modha, Thomas Mandl, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Tharindu Ranasinghe, and Marcos Zampieri. 2021. 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.

L. G. Mojica. 2016. Modeling trolling in social media conversations.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2020. Arabic offensive language on twitter: Analysis and experiments.

Nitin, Ankush Bansal, Siddhartha Mahadev Sharma, Kapil Kumar, Anuj Aggarwal, Sheenu Goyal, Kanika Choudhary, Kunal Chawla, Kunal Jain, and Manav Bhasinar. 2012. Classification of flames in computer mediated communications.

Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in greek. In *Proceedings of LREC*.

Sasha Sax. 2016. Flame wars: Automatic insult detection.

Sima Sharifirad and Stan Matwin. 2019. When a tweet is actually sexist. a more comprehensive classification of different online harassment categories and the challenges in nlp.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of germeval task 2, 2019 shared task on the identification of offensive language. In *Proceedings of KONVENS*.

Anand Subramanian, Mukesh Reghu, and Sriram Rajkumar. 2021. Team_buddi at comma@icon: Exploring individual and joint modelling approaches for detecting aggression, communal bias and gender bias. In *Proceedings of the 18th International Conference on Natural Language Processing: Shared Task on Multilingual Gender Biased and Communal Language Identification*, pages 13–20, NIT Silchar. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proceedings of GermEval*.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the NAACL*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology (NAACL-HLT)*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Team_BUDDI at ComMA@ICON: Exploring Individual and Joint Modelling Approaches for detecting Aggression, Communal Bias and Gender Bias

Anand Subramanian *
BUDDI AI
anands@buddi.ai

Mukesh Reghu *
BUDDI AI
mukeshr@buddi.ai

Sriram Rajkumar
BUDDI AI
sriramr@buddi.ai

Abstract

The ComMA@ICON 2021 Shared Task involved identifying the level of aggression and identifying gender bias and communal bias from texts in various languages from the domain of social media. In this paper, we present the description and analyses of systems we implemented towards these tasks. We built systems utilizing Transformer-based models, experimented by individually and jointly modelling these tasks, and investigated the performance of a feature engineering method in conjunction with a joint modelling approach. We demonstrate that the joint modelling approaches outperform the individual modelling approach in most cases.

1 Introduction

Social media has revolutionized how people communicate and engage in discourse and debate regarding various issues in society. In India, regional languages have influenced how content is generated on social media, with English not being the only language in which people interact with each other on forums. However, it is vital to ensure that discourse on social media is civil and respectful and does not serve as an outlet to malign or abuse users. Abusive or hostile content can manifest in various forms, including but not restricted to aggressive or hostile personal comments or posts, content that may be communal and malign religious sentiments or content that may be discriminatory based on gender. Therefore it is imperative to handle these types of content in a time-bound and sensitive manner. Modelling such text could help build automated or human-in-the-loop systems that can assist manual content moderators in reviewing and flagging such objectionable content.

Natural Language Processing can be extremely integral in this regard. However, modelling text from the social-media domain comes with challenges that need to be addressed. First of all, text generated on social media is significantly different from the text in books or newspapers. One such difference is the comparatively short length of texts in social media, such as tweets. Another difference would be the informal nature of discourse on social media forums, which includes the usage of slang, emojis and hashtags. Thirdly, social media text may be code-mixed, further complicating the process. An example of this is Hinglish, a combination of Hindi and English. These factors must be considered when modelling such text.

2 About the Task

The ComMA Project's Shared Task on Multilingual Gender Biased and Communal Language Identification (Kumar et al., 2021a) [1] [2] provided datasets spanning Hindi, Bangla (Indian variety), Meitei and English (Kumar et al., 2021b). The shared task comprised 3 sub-tasks which involved detecting the level of aggression, the identification of gender bias, and the identification of communal bias in a given text.

3 Challenges

Since the task involved data from informal domains of discourse like social media, some factors were to be considered while building systems for these tasks. Some of those considerations were:

1) **The dataset comprises code-mixed text for each language.** For instance, the text as part of the Hindi corpus may contain Hindi words written in English script (Hinglish), purely Hindi and purely English words as part of it. Thus, it is essential to

*Equal Contribution

[1] https://sites.google.com/view/comma-at-icon2021/overview

[2] 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 13–20
National Institute of Technology - Silchar, December 16-19, 2021.

ensure that models trained toward this task adapt to the code-mixed nature of the text.

2) **The level of aggression, presence of gender bias and communal bias are annotated for each text in the dataset.** The sub-tasks revolve around identifying these labels given a text. Multiple approaches are possible for solving these problems. The sub-tasks can be modeled independently or modeled jointly.

4 System overview

We built systems towards solving the three sub-tasks, for the Hindi corpus and the Multilingual Corpus, considering the factors mentioned above. To this end, for tackling the Hindi corpus, we utilize a BERT (Devlin et al., 2019) model, which was finetuned on Hinglish tweets with the Language modelling (LM) task (Bhange and Kasliwal, 2020) (Kasliwal and Bhange) (meghanabhange/Hinglish-Bert), hereafter referred to as **Hinglish-BERT**, as our starting point for all systems we submitted for the Hindi Task.

We utilize XLM-Roberta (**XLM-R**) (Conneau et al.) (Hugging Face - XLM-Roberta-Base) as our starting point for the system built as part of our submission towards the Multilingual Corpus as Hindi, Bengali, and English are part of the list of languages used for training the XLM-R model.

5 Methods

Each of these three tasks of aggression prediction **(AG)**, gender bias prediction **(GEN)** and communal bias prediction **(COM)** in the dataset are multi-class problems where the set of possible classes for each of the tasks are given by $\mathcal{Y}_{AG} = \{NAG, CAG, OAG\}$, $\mathcal{Y}_{GEN} = \{NGEN, GEN\}$ and $\mathcal{Y}_{COM} = \{NCOM, COM\}$. Refer to Table 1 for legend of the classes.

We are given a multi-task text classification dataset given by $\mathcal{D} = \left\{(x^{(i)}, y^{(i)})\right\}_{i=1}^{N} \subset \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the set of all preprocessed texts in the dataset and $\mathcal{Y} = \mathcal{Y}_{\text{AG}} \times \mathcal{Y}_{\text{GEN}} \times \mathcal{Y}_{\text{COM}}$ is the set of all possible annotations or predictions for the text.

Also, for a given sample $(x, y) \in \mathcal{D}$ and task $t \in \{AG, GEN, COM\}$, we define y_t as the true class annotation corresponding to the task t.

We first preprocess each of the text in the raw dataset using the preprocessing steps from (Bhange and Kasliwal, 2020) to obtain the preprocessed texts $x \in \mathcal{X}$.

Short Form	Description
Aggression (AG)	
NAG	Non-aggressive
CAG	Covertly aggressive
OAG	Overtly aggressive
Gender Bias (GEN)	
NGEN	Non-gendered
GEN	Gendered
Communal Bias (COM)	
NCOM	Non-communal
COM	Communal

Table 1: Legend of short forms and descriptions for each of the classes for each of the tasks.

We then embed each of these preprocessed texts $x \in \mathcal{X}$ into a hidden representation $\mathbf{h}_x$ by feeding x to the Hinglish-BERT backbone and extracting its hidden representation corresponding to the **[CLS]** (classification) token (which is used as the representation for text **x**). For the **XLM-R** model, we use the representation of the $\langle$**s**$\rangle$ token as the representation of the text.

We now define the function **Hinglish-BERT** which embeds a given preprocessed text x into a hidden representation $\mathbf{h}_x$ as described above.

$$\mathbf{h}_x = \textbf{Hinglish-BERT}(x) \tag{1}$$

For each of the tasks, $t \in \{AG, GEN, COM\}$, we then use *task-specific head layers* $\mathbf{H}_t$ to obtain the **prediction probabilities** $\hat{\mathbf{y}}_t$ for each of the classes in the task from the hidden representations of the texts as given by:

$$\hat{\mathbf{y}}_t = \mathbf{H}_t(\mathbf{h}_x) \in \{0, 1\}^{|\mathcal{Y}_t|} \tag{2}$$

where, the task specific head $\mathbf{H}_t$ is two fully connected layers stacked on one another with **ReLU** activation in between and **softmax** at the output as graphically represented in Fig 1.

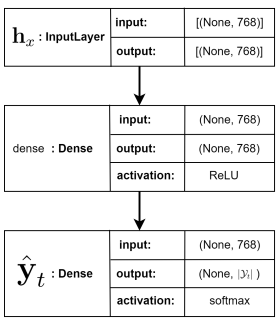


Figure 1: Architecture of the task-specific head $\mathbf{H}_t$

We train all our models end-to-end. We use the **Cross Entropy** loss for each of the individual tasks t, and define the task-specific loss $\mathcal{L}_t$ as:

$$\mathcal{L}_t = \textbf{CrossEntropyLoss}(\textbf{y'}_t, \hat{\textbf{y}}_t) \quad (3)$$

where, $\textbf{y'}_t$ is the one-hot probability vector corresponding to the true class annotation for the task t, y_t.

5.1 Three Individual Task-Specific Models

In this approach, we fine-tune three independent **Hinglish-BERT** models with *task-specific head* for each of the tasks and optimize them for their corresponding task-specific loss $\mathcal{L}_t$ (Equation 3).

The task-specific model $\textbf{model}_t$ and its *prediction probabilities* $\hat{\textbf{y}}_t$ corresponding to each of the tasks, $t \in \{AG, GEN, COM\}$ are given by:

$$\hat{\textbf{y}}_t = \textbf{model}_t(x) = \textbf{H}_t\big(\textbf{Hinglish-BERT}_\textbf{t}(x)\big) \quad (4)$$

5.2 Joint Modelling Approaches

We also build systems that jointly model the tasks using a single model architecture to investigate if performance improvements are possible due to joint modelling. A different method of jointly modelling such tasks was attempted in (Mishra et al., 2020).

The tasks are significantly intersectional, i.e., a text with *communal bias* present in it, may have *aggressive content* present, or a text may have *aggressive content* with an *overt gender bias*, etc. It is possible for the model to potentially learn better representations when these tasks are modeled jointly. These approaches also have significantly fewer parameters than training individual task-specific models due to a **shared Hinglish-BERT Backbone**.

5.2.1 Joint Model for Tasks (Three Heads)

In this approach, we jointly model the three tasks using a single model architecture that has a **common Hinglish-BERT backbone** with three task-specific heads (each corresponding to one of the tasks).

For each of the tasks, $t \in \{AG, GEN, COM\}$, the prediction probabilities for the classes in task are given by:

$$\hat{\textbf{y}}_t = \textbf{model}_t(x) = \textbf{H}_t\big(\textbf{Hinglish-BERT}_\textbf{common}(x)\big) \quad (5)$$

Method	Notation
Three Individual Models - Hindi Data	HIN-3-IND
Joint Model with Three Heads - Hindi Data	HIN-JNT-3H
Joint Model with Three Heads and Feature Engg - Hindi Data	HIN-JNT-3H+FE
Joint Model with Hierarchical Heads for Aggression Task - Hindi Data	HIN-JNT-4H
Joint Model with Three Heads - Multilingual Data	MULTI-JNT-3H

Table 2: Notations for denoting each of the systems

We then use the true class annotations and predicted class probabilities for each of the tasks to compute the task specific losses $\mathcal{L}_{AG}$, $\mathcal{L}_{GEN}$ and $\mathcal{L}_{COM}$. We then combine these losses by averaging them to get our overall loss $\mathcal{L}$, which we optimize for in our model.

$$\mathcal{L} = \frac{\mathcal{L}_{AG} + \mathcal{L}_{GEN} + \mathcal{L}_{COM}}{3} \quad (6)$$

In the multilingual case, we fine-tune an **XLM-R model** instead of a *Hinglish-BERT model* and jointly model the tasks using the same approach.

5.2.2 Joint Model for Tasks with Feature Engineering (Three Heads)

In this approach, we build upon our previous Joint Model for Tasks (Three Heads) approach. However, in the preprocessing step, we introduce a special token (i.e., an unused token from the BERT vocabulary) to act as a marker to surround words that could be informative to the model while learning the three tasks. These words could be obtained through a curated lookup.

For example: We used "ye sabse bdi **[unused1] chutiya [unused1]** aurat h **[unused1] bc [unused1]**" to replace our preprocessed text "ye sabse bdi **chutiya** aurat h **bc**" in which the words "chutiya" and "bc" are present in our lookup of curated words.

The usage of marker tokens has been widely explored for tasks like Relation Extraction (RE) in NLP (Wu and He, 2019) (Baldini Soares et al., 2019) (Shen and Huang, 2016).

This approach could also potentially reduce the necessity to retrain the model to address failure cases in unseen data by adding those words that may be informative to the model from these failed cases to our lookup. Since the marker token is used

System	Instance-F_1	Overall micro-F_1	Aggression micro-F_1	Gender Bias micro-F_1	Communal Bias micro-F_1
HIN-3-IND	0.3461 ± 0.0136	0.7004 ± 0.0119	0.6074 ± 0.0276	0.781 ± 0.0111	0.7128 ± 0.0287
HIN-JNT-3H	0.3832 ± 0.0317	0.7151 ± 0.0144	0.6142 ± 0.0203	0.7868 ± 0.0199	0.7443 ± 0.0211
HIN-JNT-3H+FE	0.3749 ± 0.0355	0.7131 ± 0.0133	0.6084 ± 0.0232	0.7894 ± 0.0161	0.7415 ± 0.0174
HIN-JNT-4H	0.383 ± 0.0337	0.7083 ± 0.0159	0.611 ± 0.0178	0.7591 ± 0.0258	0.7549 ± 0.0129

Table 3: Summary of results on the test set of the **Hindi** dataset averaged across runs on 5 random seeds for various approaches with **Hinglish-BERT**

System	Instance-F_1	Overall micro-F_1	Aggression micro-F_1	Gender Bias micro-F_1	Communal Bias micro-F_1
MULTI-JNT-3H	0.371	0.713	0.539	0.767	0.834

Table 4: Results on the test set of the **Multilingual** dataset with **XLM Roberta** as reported by the organizers on the leaderboard

to highlight a word from the lookup, it could act as a predictive signal in the data.

5.2.3 Joint Model for Tasks with Hierarchical Heads for Aggression Prediction (Four Heads)

In this approach, we break the task of aggression prediction into a *hierarchy of two binary classification sub-tasks*. We first predict whether the text is *aggressive* or *non-aggressive* (this sub-task is referred to as **A1**). Then we further predict each of the texts predicted as aggressive as being either *covertly* or *overtly aggressive* (this sub-task is referred to as **A2**).

Similar to Joint Model for Tasks (Three Heads), we jointly model each of the four tasks/sub-tasks (GEN, COM, A1, A2) using a single model architecture with a **common Hinglish-BERT backbone** and task-specific (or sub-task specific) heads (each corresponding to one of the tasks). Therefore, for each of the tasks, the prediction probabilities for the classes in the task are as given by Equation 5.

For the tasks **GEN** and **COM**, we compute class prediction probabilities $\hat{\mathbf{y}}_{GEN}$ and $\hat{\mathbf{y}}_{COM}$, and thereby compute the task-specific losses $\mathcal{L}_{GEN}$ and $\mathcal{L}_{COM}$ as in Joint Model for Tasks (Three Heads).

Further, for each of the samples $(x, y) \in \mathcal{D}$, we define the true annotation for the **A1** subtask, y_{A1}, as follows:

$$y_{A1} = \begin{cases} AG & \text{if } y_{AG} \in \{CAG, OAG\} \\ NAG & \text{if } y_{AG} = NAG \end{cases} \tag{7}$$

We then compute the prediction probabilities for the classes $\{AG, NAG\}$ in the sub-task **A1** using the common Hinglish-BERT with sub-task specific head $\mathbf{H}_{A1}$, and thereby compute the sub-task specific loss $\mathcal{L}_{A1}$ as the Cross Entropy loss of the prediction probabilities ($\hat{\mathbf{y}}_{A1}$) and true annotations (y_{A1}).

For the fourth sub-task **A2**, given a mini-batch (or dataset) $\mathcal{D}_{train}$ for training, we filter the samples for which the true aggression annotation y_{AG} belongs to $\{CAG, OAG\}$ as given by:

$$\mathcal{D}_{train,A2} = \big\{(x, y) \big| (x, y) \in \mathcal{D}_{train} \wedge y_{AG} \in \{CAG, OAG\}\big\} \tag{8}$$

We then compute the sub-task specific loss $\mathcal{L}_{A2}$ for the mini-batch (or dataset) $\mathcal{D}_{train}$ using only the samples in $\mathcal{D}_{train,A2}$ (Eq. 8).

During training on mini-batch D_{train}, we compute the prediction probabilities for the classes $\{CAG, OAG\}$ in the sub-task **A2** for only the samples in $D_{train,A2}$ using the common Hinglish-BERT with sub-task specific head $\mathbf{H}_{A2}$, and thereby compute the sub-task specific loss $\mathcal{L}_{A2}$ as the *Cross Entropy* loss of the prediction probabilities ($\hat{\mathbf{y}}_{A2}$) and true annotations (y_{A2} which equals y_{AG})[3].

During training, given a mini-batch of samples D_{train}, we define the corresponding overall loss $\mathcal{L}$ which we optimize for in our model as:

$$\mathcal{L} = \left[\frac{|\mathcal{D}_{train}| \times (\mathcal{L}_{A1} + \mathcal{L}_{GEN} + \mathcal{L}_{COM}) + |\mathcal{D}_{train,A2}| \times \mathcal{L}_{A2}}{3 \times |\mathcal{D}_{train}| + |\mathcal{D}_{train,A2}|} \right] \tag{9}$$

[3] y_{A2} equals y_{AG} for samples in $D_{train,A2}$ as we only have samples with true class annotations **CAG** and **OAG** in it

where, for mini-batch D_{train}, $D_{train,A2}$ is as given in Equation 8.

6 Experimental Setup

We utilize PyTorch[4](Paszke et al., 2019) and the Transformers Library[4] from Hugging Face (Wolf et al., 2020) for implementing our systems. The code and resources used are made available on GitHub[5]. The systems we built and their respective notations are summarized in Table 2.

6.1 Preprocessing of texts

We preprocess each of the texts before feeding them to the models for both training/inferencing.

For the **Hindi** dataset, we preprocess the texts along the lines of (Bhange and Kasliwal, 2020) by performing the following transformations on them:

- Replace "@" with "mention", "#" with "hashtag" and retweet related information in texts with the word "Retweet"; remove http(s) URLs
- Convert emojis to their text equivalent using the emoji packages (Kim et al.)

For the **Multilingual** dataset, the preprocessing involves the removal of retweet related information, mentions of users, http(s) URLs and emojis.

6.2 Hyperparameters

The default hyperparameters we used while training all the systems unless mentioned otherwise below are summarized in the Table 5.

Hyperparameter	Value
Tokenizer max sequence length	128
Training batch Size	32
Learning rate	5e-5
Number of training epochs	10

Table 5: Default hyperparameters for the systems which are to considered unless specified otherwise

We used the AdamW (Loshchilov and Hutter, 2017) optimizer for training all our systems.

For the **HIN-JNT-3H+FE** system, we use a learning rate of 3e-4 for the parameters in the task-specific heads and a lower learning rate of 5e-5 for the parameters in the common Hinglish-BERT backbone.

For the **HIN-JNT-3H+FE** system, we train the models for 20 epochs.

We evaluate the model checkpoints for each of the systems after each epoch using the validation set and pick the checkpoint with the *best instance-F_1* for joint modelling (**JNT**) systems and the checkpoint with the *best accuracy* for each of the individual models for individual modelling (**IND**) systems. We further evaluate this best model checkpoint which was picked on the test set, and report the scores.

For the **HIN-JNT-3H+FE** system, we used a publicly available lookup of profanity words from (pmathur5k10), (Mathur et al., 2018) in combination with a set of words that could be indicative of profanity or used in a profane manner, (tabulated in Table 7) which were manually curated by analyzing some of the samples from the corresponding train and validation splits.

6.3 Evaluation Metrics

The shared task uses instance-F_1 as the primary evaluation metric and overall micro-F_1 as the secondary evaluation metric for the systems[6].

7 Results

For the **Hindi set**, we initially performed one run of each of the systems and submitted the results to the leaderboard. The scores for these runs are present in the first row of Tables 9, 10, 11 and 12 [7]. In these runs, we observed that the **HIN-JNT-4H** system performed the best, followed by the **HIN-JNT-3H** system.

We further re-ran the systems four more times, with different seeds for each run to account for the impact of randomness in our systems' performances. In terms of instance-F_1, we observe that the joint modelling approaches often outperform the system of individually trained models across the runs. This is also evident in the mean scores reported in Table 3, and it highlights the potential benefits of jointly modelling the tasks.

However, when we further compare the performance within the different joint modelling approaches, we observe no clear winner under all circumstances, as the performance often varies with

[4]We use PyTorch library version 1.10.0+cu111 and Transformers library version 4.12.5

[5]https://github.com/BUDDI-AI/ComMA-ICON-2021

[6]https://competitions.codalab.org/competitions/35482#learn_the_details-evaluation

[7]The scores corresponding to the run submitted to the leaderboard are marked with an "(L)" in the "Run" column in each of these tables.

True \ Predicted	NAG	CAG	OAG
NAG	309	48	120
CAG	31	13	41
OAG	73	53	314

(a) Aggression

True \ Predicted	NGEN	GEN
NGEN	740	58
GEN	134	70

(b) Gender Bias

True \ Predicted	NCOM	COM
NCOM	593	47
COM	199	163

(c) Communal Bias

Table 6: **Confusion matrices** for **test set** predictions by the **HIN-JNT-3H** system's model corresponding to the **5**th **run** for each of the three tasks of Aggression, Gender Bias and Communal Bias Identification

the random seed used as part of the run.

For the **Multilingual set**, we submitted only one system, **MULTI-JNT-3H**, whose results are presented in Table 4. This system jointly modeled all the three tasks using the approach from Joint Model for Tasks (Three Heads), and we observe that the model performs quite competitively.

7.1 Analysis

From Table 3, we observe that from among all the systems on the *Hindi dataset*, the **HIN-JNT-3H** system has the *best mean Instance-F_1 score across the 5 runs of the system on the test set*. Therefore, we pick the **HIN-JNT-3H** system and select the system's model corresponding to the run with the highest instance-F_1 (i.e., **Run 5** from Table 10). We then analyze the ***confusion matrices*** *of the selected model on the test set for each of the three tasks* (which are tabulated in Table 6).

7.1.1 Aggression Level Identification Task

For this task, we observe that out of the **85** samples with true class annotation **CAG**, only 13 samples **(15.3%)** are correctly predicted by the model as belonging to class **CAG**, whereas 31 samples **(36.47%)** are predicted as **NAG** and 41 samples **(48.24%)** are predicted as **OAG**. This indicates that the model may be struggling to sufficiently identify texts with subtle characteristics of aggression, and instead classifies them into one of the two extremes (**NAG** or **OAG**).

7.1.2 Gender Bias Identification Task

For this task, we observe that the model has a precision of **54.69%** and a recall of **34.31%** for the **GEN** class whereas it has a precision of **84.67%** and a recall of **92.73%** for the **NGEN** class. It indicates that the model performs better in accurately recalling and identifying non-gendered texts than recognizing gendered text.

7.1.3 Communal Bias Identification Task

For this task, the model has a precision of **77.62%** and a recall of **45.03%** on the **COM** class. It indicates that, while the model may face issues in retrieving all the communally-biased text samples (as indicated by its recall), the samples predicted as **COM** by the model are quite likely to be communally biased (as indicated by its precision).

7.1.4 Class Imbalance

As indicated by Table 8, it is clear that there is an imbalance in class distribution across the tasks in the train set of the Hindi data, which could account for some of the problems discussed previously. Techniques from the imbalanced learning literature, such as sampling or weighted loss functions, could be explored.

Conclusion

Thus, we present the description and analyses of the systems we submitted towards these tasks. Future extensions to this work could include assessing the performance of our systems across different folds of the data for more robust evaluation. The performance of other transformer-based models on the corpora could also be analyzed.

Acknowledgments

We would like to thank our colleagues in the Research and Development division and the management of BUDDI AI for their constant support and encouragement.

References

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Meghana Bhange and Nirant Kasliwal. 2020. HinglishNLP at SemEval-2020 task 9: Fine-tuned language models for Hinglish sentiment detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 934–939, Barcelona (online). International Committee for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal Vishrav Chaudhary Guillaume Wenzek, Francisco Guzmán, Edouard Grave Myle Ott Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hugging Face - XLM-Roberta-Base. xlm-roberta-base.

Nirant Kasliwal and Meghana Bhange. Nirantk/hinglish: Hinglish text classification.

Taehoon Kim, Kevin Wurster, and Tahir Jalilov. Emoji for python.

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.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. 2018. Did you offend me? classification of offensive tweets in Hinglish language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 138–148, Brussels, Belgium. Association for Computational Linguistics.

meghanabhange/Hinglish-Bert. meghanabhange/hinglish-bert.

Sudhanshu Mishra, Shivangi Prasad, and Shubhanshu Mishra. 2020. Multilingual joint fine-tuning of transformer models for identifying trolling, aggression and cyberbullying at TRAC 2020. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 120–125, Marseille, France. European Language Resources Association (ELRA).

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

pmathur5k10. pmathur5k10/hinglish-offensive-text-classification: Hinglish offensive text classification.

Yatian Shen and Xuanjing Huang. 2016. Attention-based convolutional neural network for semantic relation extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2526–2536, Osaka, Japan. The COLING 2016 Organizing Committee.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Perric Cistac, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. pages 38–45. Association for Computational Linguistics.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 23612364, New York, NY, USA. Association for Computing Machinery.

A Appendix

The words in Table 7 are added to the corpus in a purely research motivated manner since they are words that can potentially be used in a profane manner in text, and we investigate if they could aid systems in better learning to recognize instances of aggressive, communal or gender biased text.

Terms				
bc	mc			
RANDI	RANDY	hore	Dalla	bs
hijra	gay			
chod	chutiya	chutiyo		
pussy	Gand	G**d	Lawde	Lowde
OLAD	harami	bootlicker		
हरामी	पिछवाड़े	भड़वे	गांडू	

Table 7: Set of code mixed Hindi words which are potentially informative for the tasks, and which were manually curated by analyzing some samples from the train and validation splits

Class	# Samples
NAG	1289
CAG	800
OAG	2526
NGEN	3665
GEN	950
NCOM	3598
COM	1017

Table 8: Distribution of classes across tasks for the **train** split of **Hindi** dataset

Run	Instance-F_1	Overall micro-F_1	Aggression micro-F_1	Gender Bias micro-F_1	Communal Bias micro-F_1
1 (L)	0.3453	0.6979	0.6457	0.7695	0.6786
2	0.3603	0.7083	0.5988	0.7745	0.7515
3	0.3433	0.6929	0.5768	0.7884	0.7136
4	0.3253	0.6866	0.5908	0.7764	0.6926
5	0.3563	0.7162	0.6248	0.7964	0.7275

Table 9: Results on the test set of the **Hindi** dataset using the **HIN-3-IND system**

Run	Instance-F_1	Overall micro-F_1	Aggression micro-F_1	Gender Bias micro-F_1	Communal Bias micro-F_1
1 (L)	0.3603	0.6946	0.6188	0.7585	0.7066
2	0.3972	0.7242	0.6257	0.7924	0.7545
3	0.3473	0.7112	0.5818	0.7994	0.7525
4	0.3832	0.7129	0.6098	0.7754	0.7535
5	0.4281	0.7325	0.6347	0.8084	0.7545

Table 10: Results on the test set of the **Hindi** dataset using the **HIN-JNT-3H system**

Run	Instance-F_1	Overall micro-F_1	Aggression micro-F_1	Gender Bias micro-F_1	Communal Bias micro-F_1
1 (L)	0.3413	0.7006	0.5978	0.7794	0.7246
2	0.3683	0.7112	0.5998	0.7784	0.7555
3	0.3473	0.7006	0.5998	0.7754	0.7265
4	0.3882	0.7226	0.5948	0.8094	0.7635
5	0.4291	0.7305	0.6497	0.8044	0.7375

Table 11: Results on the test set of the **Hindi** dataset using the **HIN-JNT-3H+FE system**

Run	Instance-F_1	Overall micro-F_1	Aggression micro-F_1	Gender Bias micro-F_1	Communal Bias micro-F_1
1 (L)	0.3982	0.7092	0.6277	0.7425	0.7575
2	0.3932	0.7169	0.6038	0.7914	0.7555
3	0.4242	0.7295	0.6317	0.7824	0.7745
4	0.3373	0.6949	0.6008	0.7435	0.7405
5	0.3623	0.691	0.5908	0.7355	0.7465

Table 12: Results on the test set of the **Hindi** dataset using the **HIN-JNT-4H system**

Hypers at ComMA@ICON: Modelling Aggressiveness, Gender Bias and Communal Bias Identification

Sean Benhur[1]*, **Roshan Nayak**[2]*, **Kanchana Sivanraju**[1]
Adeep Hande[3], **Subalalitha Chinnaudayar Navaneethakrishnan**[4],
Ruba Priyadharshini[5], **Bharathi Raja Chakravarthi**[6]
[1]PSG College of Arts and Science [2]BMS College of Engineering
[3]Indian Institute of Information Technology Tiruchirappalli
[4] SRM Institute of Science and Technology, India [5]ULTRA Arts and Science College
[6]National University of Ireland Galway
seanbenhur@gmail.com

Abstract

Due to the exponentially increasing reach of social media, it is essential to focus on its negative aspects as it can potentially divide society and incite people into violence. In this paper, we present our system description of work on the shared task ComMA@ICON, where we have to classify how aggressive the sentence is and if the sentence is gender-biased or communal-biased.These three could be the primary reasons to cause significant problems in society. As team Hypers we have proposed an approach which utilizes different pretrained models with Attention and mean pooling methods. We were able to get Rank 3 with 0.223 Instance F1 score on Bengali, Rank 2 with 0.322 Instance F1 score on Multi-lingual set, Rank 4 with 0.129 Instance F1 score on Meitei and Rank 5 with 0.336 Instance F1 score on Hindi. The source code and the pretrained models of this work can be found here[1].

1 Introduction

The Internet is a vast network that connects devices all over the world. Due to mobile technology and affordable internet plans, users can access the Internet with ease, leading to the tremendous growth of the Internet, which is unprecedented. As of January 2021, there were 4.66 billion active internet users, 59.5% of the world's population. Users would undoubtedly want to increase their reach virtually, and hence the interaction among the people would increase. The people these days are more vocal and, at any cost, want their voices or opinions to be reached to a multitude of people. Hence, people search for a platform to share their views, and social media is an ideal place for that. This exact mindset of people has fueled the copious amounts of social media users globally.

Social media are the technologies that allow the creation, sharing, or exchange of information, interests, ideas, and other forms of expression. Its use is an ever-increasing phenomenon of the 21st century (Livingstone and Brake, 2010). There are a plethora of social media platforms, each attracting people in unique ways. As of January 2021, there were 4.2 billion active social media users. Considering the reach of social media, they can spread people's opinions in a few minutes (Zhang and Vos, 2015). Hence it will have a significant effect on society which could be both positive as well as harmful (Harchekar, 2017). But there are instances in which the situation would go out of hand. For example, people could differ in their opinions, and people with similar views tend to form a group to denounce the group with ideas that are not the same as theirs. During the denouncement, there is a possibility that a user could show his improper behaviour, thus making offensive (Hande et al., 2021b), misogynistic (Shushkevich and Cardiff, 2019), hateful (Bhatia et al., 2021), or any kind of statements that has the potential to create controversy (Coletto et al., 2017). Such statements may be intended towards an individual or a group and are not considered to be good or acceptable in the society. As such behaviour would influence others wrongly and instigate violence or affect the mental health, leading to unpleasant situations. Hence it is necessary to flag such posts and its advisable to take them down from the social media platform and also retribute the user responsible for such posts(Hande et al., 2021a). There could be several reasons that a post by the user is considered inappropriate. Considering how important it is to regulate toxic post, in this paper we will be presenting a system to identify if the user is being aggressive on some individual or a community, or being biased regard-

*Equal Contribution

[1]https://github.com/seanbenhur/multilingual_aggresive_gender_bias_communal_bias_identifcation

Proceedings of the 18th International Conference on Natural Language Processing: Shared Task on Multilingual Gender Biased and Communal Language Identification, pages 21–25
National Institute of Technology - Silchar, December 16-19, 2021.

Text	Language	Aggressive	Gender Bias	Communal Bias
angakpa nupini eiga unaradi fadoubi	Meitei	CAG	NGEN	NCOM
hehhh ym pkte nupi ng	Meitei	CAG	GEN	NCOM
tome hola bal chera tumi nijai jante	Bengali	OAG	GEN	NCOM
you know to whom im addressing, 'ye hindustan ke liye dimak h jo usko ander se khokla krre h' #muslimvirus	Hindi	OAG	NGEN	COM
mulle tere allah ki ma ka bhosda	Hindi	OAG	GEN	COM
gudmarani chale ke maaaar	Multi	OAG	GEN	NCOM
jay bheem namo buddhay	Multi	OAG	NGEN	NCOM

Table 1: Samples from the dataset and their corresponding class labels for each of the tasks.

Language	NAG	CAG	OAG	NGEN	GEN	NCOM	COM	Total
Meitei	1,258	1,495	456	3,006	203	2,967	242	3,209
Bengali	1,115	494	1,782	2,120	1,271	2,975	416	3,391
Hindi	1,594	969	3,052	4,440	1,175	4,402	1,213	5,615
Multi-lang	3,966	2,956	5,289	9,564	2,647	10,342	1,869	12,211

Table 2: Samples distribution in the training set.

ing the gender (Jatmiko et al., 2020; Hande et al., 2020), or is targeting a particular religion or caste (roy, 2016).

Undoubtedly, English is the widely spoken language in the world (Crystal, 2008). But as there are no hardbound rules that users must text in English, the text found on social media could be multilingual and lack grammatical rules (Yuliah et al., 2020). Also, there could be unwanted symbols in the text (Chakravarthi et al., 2021). Considering all such challenges, in this paper, we present a model to classify the multi-lingual sentence written by the user as to how aggressive it is and if it is gendered and communal oriented text. The dataset had multilingual texts with the code-mix of English and several other languages native to India. Meitei and Bangla are native to the Indian states of Manipur and West Bengal, respectively, whereas Hindi is predominant in Northern India.

The rest of the paper is structured as follows, section 2 describes about dataset used for the shared task. The section 3 describes the models and architectures that were used for the tasks. In section 4 we discuss about the results obtained during the study, and the last section 5 concludes the work.

2 Dataset

The ComMA dataset was provided in this task (Kumar et al., 2021a; B et al., 2021). The dataset had annotations for aggression, gender bias, and communal bias identification for multi-lingual social media sentences(Kumar et al., 2021b). The dataset comprises of code-mixed sentences has 15,000 code-mixed sentences. It is divided into 12,000 sentences for development and 3,000 sentences for the test. The data is divided into four sets, namely Meitei, Bengali, Hindi, and Multi-lingual. The Multi-lingual set comprises sentences of all three languages. The Table 1 gives an idea of how the dataset could look like. The sentences in every set are classified into one of the classes for each of the three tasks. The tasks and their classes include,

- **Aggression Classification**: The text is divided into Overtly Aggressive (OAG), Covertly Aggressive (CAG) or Non-aggressive (NAG)
- **Gender Bias Classification**: The text is divided into gendered (GEN) or non-gendered (NGEN).
- **Communal Bias Classification**: The text is divided into communal (COM) or non-communal (NCOM)

The samples count of classes is far from equal.

Language	NAG	CAG	OAG	NGEN	GEN	NCOM	COM	Total
Meitei	370	471	159	945	55	932	68	1,000
Bengali	333	157	501	624	367	879	112	991
Hindi	305	167	526	775	225	804	196	998
Multi	1,007	797	1,193	2,349	648	2,622	375	2,997

Table 3: Distribution of samples in the dev set.

		Language	AGG			GB			CB			Overall		
Model	**Arch**		P	R	F1	P	R	F1	P	R	F1	P	R	F1
MURIL	AP	Meitei	0.470	0.470	0.470	0.599	0.599	0.599	0.493	0.493	0.493	**0.521**	**0.521**	**0.521**
MURIL	MP	Meitei	0.471	0.471	0.471	0.603	0.603	0.603	0.356	0.356	0.356	0.477	0.477	0.477
csebuetnlp/banglabert	AP	Bengali	0.642	0.642	0.642	0.755	0.755	0.755	0.692	0.692	0.692	**0.696**	**0.696**	**0.696**
csebuetnlp/banglabert	MP	Bengali	0.635	0.635	0.635	0.762	0.762	0.762	0.612	0.612	0.612	0.670	0.670	0.670
MURIL	AP	Hindi	0.594	0.594	0.594	0.816	0.816	0.816	0.909	0.909	0.909	0.773	0.773	0.773
MURIL	MP	Hindi	0.683	0.683	0.683	0.827	0.827	0.827	0.902	0.902	0.902	**0.804**	**0.804**	**0.804**
MURIL	AP	Multi-Lang	0.618	0.618	0.618	0.839	0.839	0.839	0.661	0.661	0.661	0.706	0.706	0.706
MURIL	MP	Multi-Lang	0.612	0.612	0.612	0.823	0.823	0.823	0.891	0.891	0.891	**0.791**	**0.791**	**0.791**

Table 4: Results on the dev set. AGG: Aggressive, GB: Gender Bias, CB: Communal Bias, Arch: Architecture, AP: Attention-pooling, MP: Mean-pooling. Metrics, Micro average scores of P: Precision, R: Recall, F1: F1-score calculated on the dev set. Overall scores are the average of the aggressive, gender bias, and communal bias.

Hence the dataset is quite imbalanced. The dataset distribution is displayed in the Table 2.

3 Methodology

In this section, we describe the methodology of our systems, including data preprocessing and Model architecture. We use mean pooled, and Attention pooled pretrained models, which was shown to provide better results (Benhur and Sivanraju, 2021). We trained all the models with the batch size of 8, dropout of 0.3, linear scheduler for learning rate scheduling with 2e-5 as an initial learning rate.

3.1 Data Preprocessing

The task dataset consists of both codemixed and native scripts; for the Bengali dataset, we converted the emojis into Bengali language using bnemo GitHub repository[2], we removed URLs and punctuations in the text for all the languages. Since the dataset is imbalanced, we sampled the dataset uniformly.

3.2 Pretrained Models

We finetuned pretrained transformers with custom poolers on hidden states on MURIL (Khanuja et al., 2021) for Hindi, Meitei and Multilingual datasets and BanglaBert (Bhattacharjee et al., 2021) for Bengali dataset. In this section, we describe our Pooling methods and pretrained models.

[2] https://github.com/faruk-ahmad/bnemo

3.2.1 MURIL

MURIL is a pretrained model, specifically made for Indian languages. MuRIL, the pretrained model, is trained in 16 different Indian Languages. Instead of the usual Masked Language Modelling approach, the model is trained on both Masked Language Modelling(MLM) objective and Translated Language Modelling(TLM) objective. In the TLM objective, both translated and transliterated sentence pairs are sent to the model for training. This model outperforms mBERT on all the tasks for Indian languages on the XTREME (Hu et al., 2020) benchmark.

3.2.2 BanglaBert

Banglabert is pretrained on more than 18.5 GB of a corpus in Bengali corpora. Banglabert achieves the state of the art performance on Bengali texts on five downstream tasks. It outperforms multilingual baselines with more than a 3.5 percentage score. Banglabert is pretrained using ELECTRA (Clark et al., 2020) with Replaced Token Detection(RTD) objective. In this setup, two networks, a generator network and discriminator network, are used, while training both the networks are trained jointly. The generator is trained on the Masked Language Modelling objective, where a portion of the tokens in the sentence is masked and is asked to predict the masked tokens using the rest of the input. The masked tickets are replaced by tokens sampled from the generator's output distribution for the corresponding masks for the discrimina-

Model	Architecture	Language	Overall Micro F1	Overall Instance F1
MURIL	Attention-pooler	Meitei	**0.472**	**0.129**
MURIL	Mean-pooler	Meitei	0.436	0.080
csebuetnlp/banglabert	Attention-pooler	Bengali	**0.579**	**0.223**
csebuetnlp/banglabert	Mean-pooler	Bengali	0.572	0.201
MURIL	Attention-pooler	Hindi	0.662	0.326
MURIL	Mean-pooler	Hindi	**0.683**	**0.336**
MURIL	Attention-pooler	Multi-Lang	**0.685**	**0.322**
MURIL	Mean-pooler	Multi-Lang	0.601	0.280

Table 5: Results on the test set. Overall, the Micro F1 score is calculated by the average of aggressive, gender, and communal biases. Instance F1 score is similar to the F1 score but when all the three labels are predicted correctly.

Language	Overall Micro F1	Overall Instance F1
Meitei	0.472	0.129
Bangla	0.579	0.223
Hindi	0.683	0.336
Multi-Lang	0.685	0.322

Table 6: Results obtained when submitted to the competition.

tor input. The discriminator then has to predict whether each token is from the original sequence or not. After pretraining, the discriminator is used for finetuning.

3.3 Attention Pooler

The attention operation described in equation 1 is applied to the CLS token in last hidden state of the pretrained transformer; we hypothesize that this helps the model learn the contribution of individual tokens. Finally, the returned pooled output from the transformer is further passed to a linear layer to predict the label.

$$o = W_h^T softmax(q h_{CLS}^T) h_{CLS} \quad (1)$$

where W_h^T and q are learnable weights and h_{CLS} is the CLS representation and o is the output.

$$y = softmax(W_o^T + bo) \quad (2)$$

3.4 Mean Pooler

In the mean-pooling method, the last hidden state of the tokens are averaged on each sentence, and it is passed onto the linear layer to output the final probabilities.

4 Results

Pretrained models with different pooling methods were trained on each language set and then validated on dev sets. For the competition submissions, we submitted the model with a higher Micro F1 score on the dev set. Table 4 shows the results of the dev set, and the Table 3 depicts the data distribution of the set used to validate the trained models. The training process was done on Tesla P100 GPU. In the test set submissions, We were able to get Rank 3 with 0.223 InstanceF1 score on Bengali, Rank 2 with 0.322 Instance F1 score on Multi-lingual set, Rank 4 with 0.129 Instance F1 score on Meitei and Rank 5 with 0.336 Instance F1 score on Hindi. The competition results are shown in Table 6. Table 5 shows the Overall Micro F1 score and Instance F1 score on the test set. The pretrained model MURIL was not trained on Meitei, but it still achieved comparable performance on the Test set; we hypothesize that since MURIL was trained both on transliterated pairs on TLM objective and the Meitei dataset also only consisted of code-mixed texts, we get a fair results on meitei test set.

5 Conclusion

In this paper, we experimented with different pooling methods, namely Attention Pooling and Mean Pooling and pretrained models, to classify sentences, how aggressive they are, and whether gender-oriented or communal. From Table 5 its evident that attention pooling worked better in most of the cases. We have also discussed the various essential reasons why the work on this is necessary. As for future work, we will consider improving our scores, especially on multilingual and meitei datasets, and experimenting with other pretrained models.

References

2016. Social media in the domain of communal violence: a study of assam riot 2012.

Senthil Kumar B, Aravindan Chandrabose, and Bharathi Raja Chakravarthi. 2021. An overview of fairness in data – illuminating the bias in data pipeline. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 34–45, Kyiv. Association for Computational Linguistics.

Sean Benhur and Kanchana Sivanraju. 2021. Psg@dravidian-codemix-hasoc2021: Pretrained transformers for offensive language identification in tanglish.

Mehar Bhatia, Tenzin Singhay Bhotia, Akshat Agarwal, Prakash Ramesh, Shubham Gupta, Kumar Shridhar, Felix Laumann, and Ayushman Dash. 2021. One to rule them all: Towards joint indic language hate speech detection.

Abhik Bhattacharjee, Tahmid Hasan, Kazi Samin Mubasshir, Mohammad Rahman, Anindya Iqbal, and Rifat Shahriyar. 2021. Banglabert: Combating embedding barrier for low-resource language understanding.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2021. Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators.

Mauro Coletto, Kiran Garimella, Aristides Gionis, and Claudio Lucchese. 2017. Automatic controversy detection in social media: A content-independent motif-based approach. *Online Social Networks and Media*, 3-4:22–31.

David Crystal. 2008. Two thousand million? *English Today*, 24(1):3–6.

Adeep Hande, Siddhanth U Hegde, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. 2021a. Benchmarking multi-task learning for sentiment analysis and offensive language identification in under-resourced dravidian languages.

Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2020. KanCMD: Kannada CodeMixed dataset for sentiment analysis and offensive language detection. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 54–63, Barcelona, Spain (Online). Association for Computational Linguistics.

Adeep Hande, Karthik Puranik, Konthala Yasaswini, Ruba Priyadharshini, Sajeetha Thavareesan, Anbukkarasi Sampath, Kogilavani Shanmugavadivel, Durairaj Thenmozhi, and Bharathi Raja Chakravarthi. 2021b. Offensive language identification in low-resourced code-mixed dravidian languages using pseudo-labeling.

Jyoti Suraj Harchekar. 2017. Impact of social media on society.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.

Mochamad Jatmiko, Muh Syukron, and Yesi Mekarsari. 2020. Covid-19, harassment and social media: A study of gender-based violence facilitated by technology during the pandemic. *The Journal of Society and Media*, 4:319–347.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

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.

Sonia Livingstone and David R Brake. 2010. On the rapid rise of social networking sites: New findings and policy implications.

Elena Shushkevich and John Cardiff. 2019. Automatic misogyny detection in social media: A survey. *Computación y Sistemas*, 23.

Siti Yuliah, Yessy Purnamasari, and Elsa Yunita. 2020. Grammatical errors in social media caption. *INTERNATIONAL JOURNAL OF LANGUAGE LITERATURE*, 8.

Boyang Zhang and Marita Vos. 2015. How and why some issues spread fast in social media. *Onlne Journal of Communication and Media Technologies*, 5:371–383.

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$_{l}anguage$

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.

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.

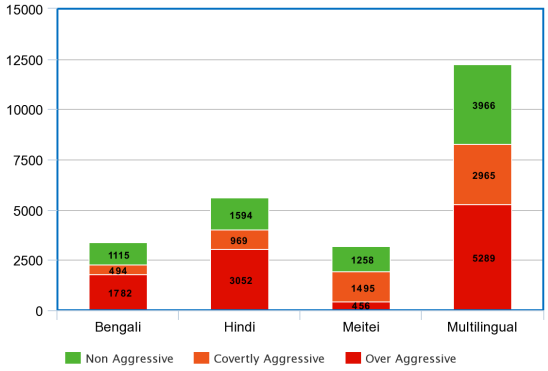


Figure 1: Distribution of text classified as 'NAG', 'CAG' and 'OAG' for different languages.

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.

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

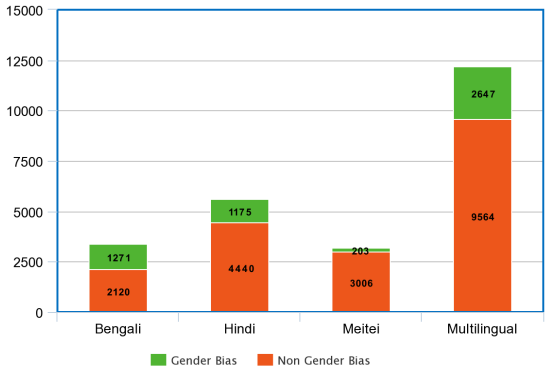


Figure 2: Distribution of text classified as 'GEN' and 'NGEN' for different languages.

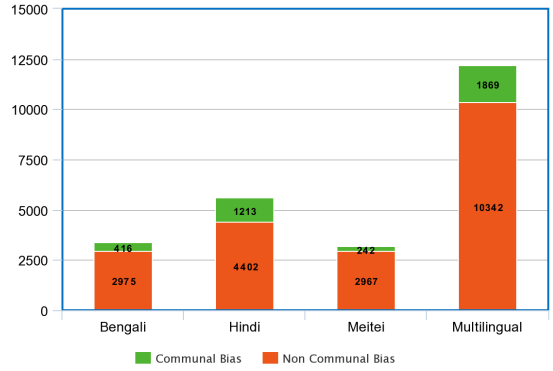


Figure 3: Distribution of text classified as 'COM' and 'NCOM' for different languages.

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

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.

$$\hat{y} = \arg\max_{i} \sum_{j=1}^{m} w_j p_{ij} \quad (1)$$

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.

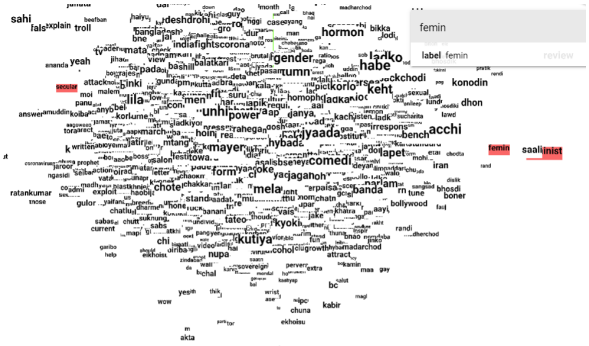


Figure 4: Word embedding for Gender bias identification in multilingual data

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 em-

bedding 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].

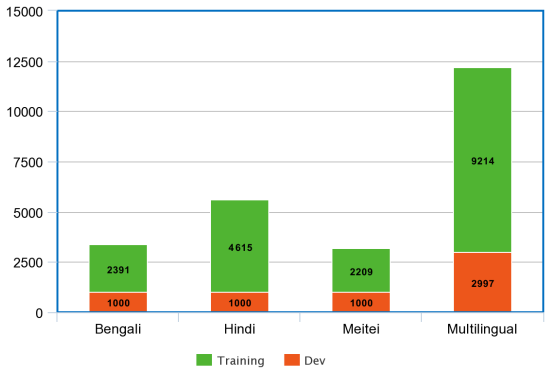


Figure 5: Distribution of Training and Dev datasets used for Training the model in all languages.

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/

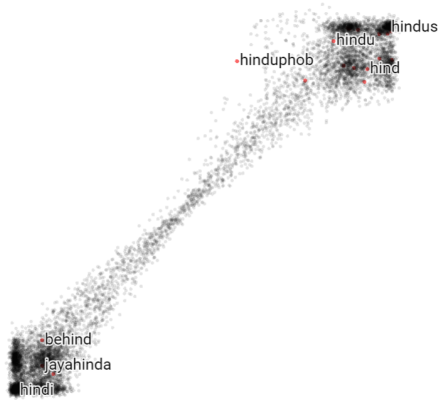


Figure 7: Plot for Communal bias in multilingual data with highlighted word as 'hindu' using T-SNE plot

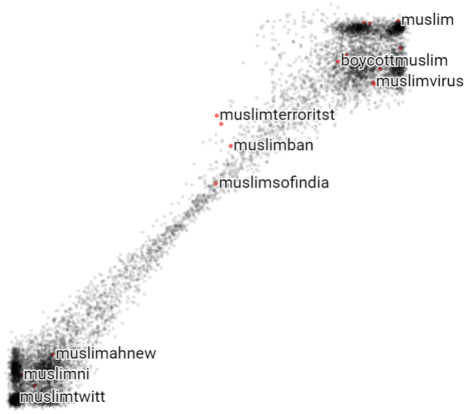


Figure 8: Plot for Communal bias in multilingual data with highlighted word as 'muslim' using T-SNE plot

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.

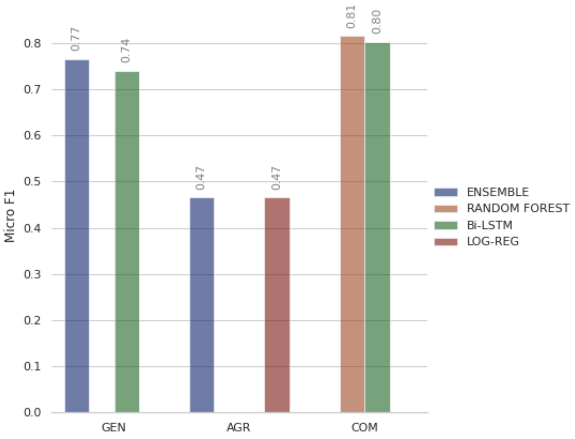


Figure 9: Comparison of Micro F1 scores for each sub task on Multilingual data.

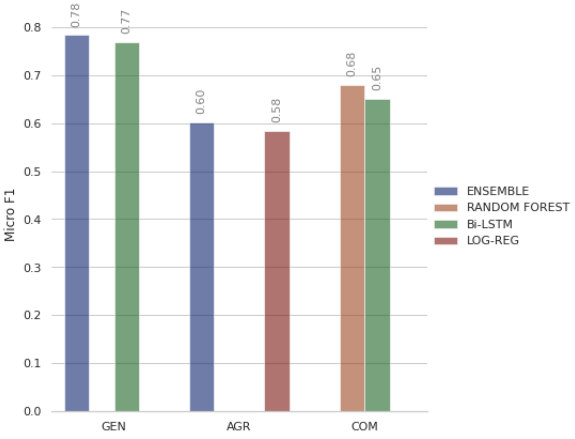


Figure 10: Comparison of Micro F1 scores for each sub task on Hindi data.

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

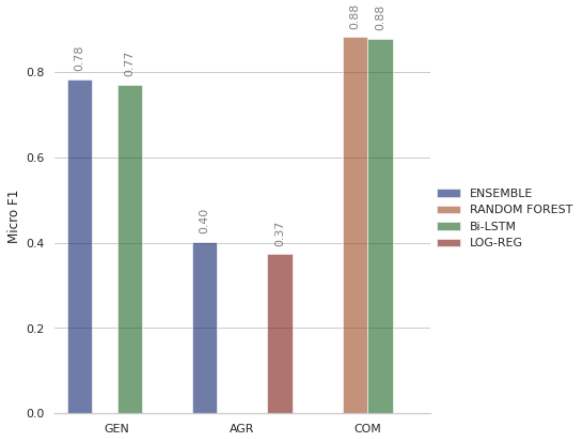


Figure 11: Comparison of Micro F1 scores for each sub task on Bangla data.

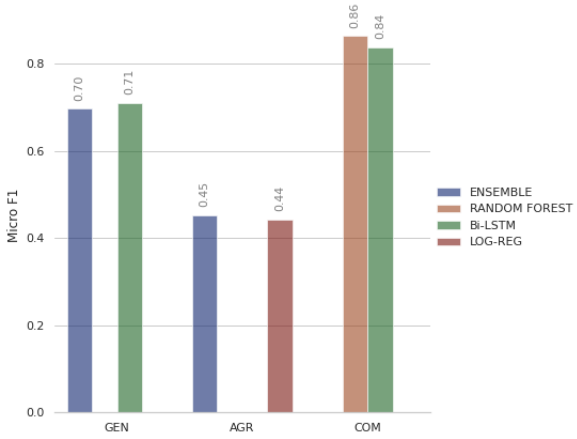


Figure 12: Comparison of Micro F1 scores for each sub task on Meitei data.

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].

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

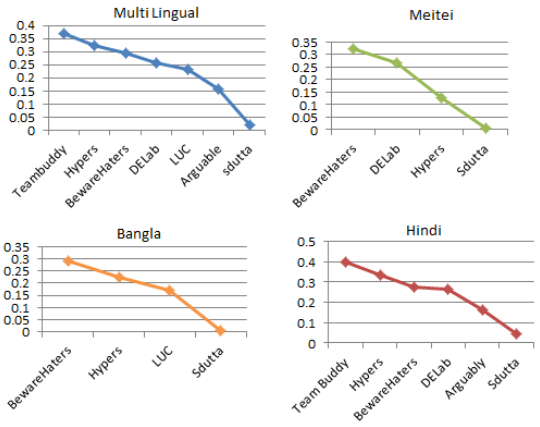


Figure 13: Instance F1 Score Based Ranking Of Team Beware Haters.

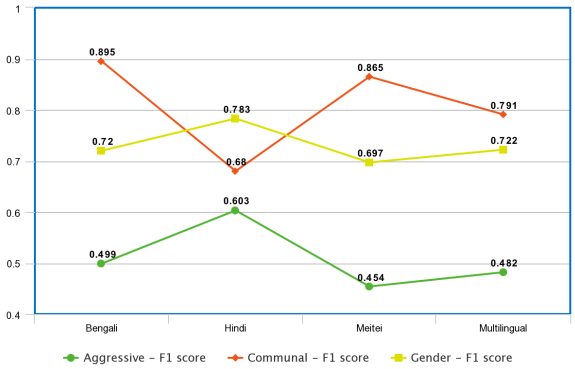


Figure 14: Task wise performance of our highest ranked models for various languages.

- 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].

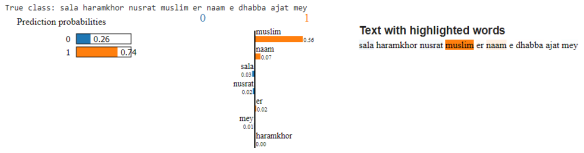


Figure 15: LIME explanations for Communal bias identification in Bengali

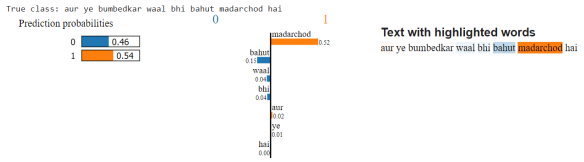


Figure 16: LIME explanations for Gender bias identification in Hindi

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

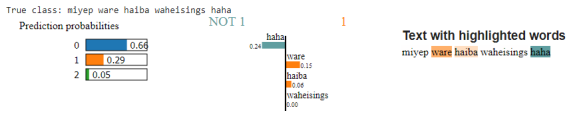


Figure 17: LIME explanations for aggression identification in Meitei

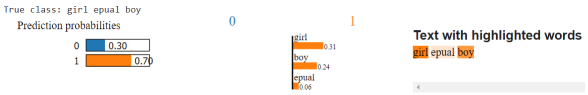


Figure 18: LIME explanations for misclassified example of Gender bias identification

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*.

DELab@IIITSM at ComMA@ICON-2021 Shared Task: Identification of Aggression and Biasness using Decision Tree

Maibam Debina Devi
IIIT Senapati, Manipur
debina@iiitmanipur.ac.in

Navanath Saharia
IIIT Senapati, Manipur
nsaharia@iiitmanipur.ac.in

Abstract

This paper presents our system description on participation in ICON-2021 Shared Task subtask 1 on multilingual gender-biased and communal language identification as team name: DELab@IIITSM. We have participated in two language specific *Meitei, Hindi* and one multilingual *Meitei, Hindi and Bangla* with English code-mixed languages identification task. Our method includes well design pre-processing phase based on the dataset, the frequency-based feature extraction technique TF-IDF which creates the feature vector for each instance using *(Decision Tree)*. We obtained weights are 0.629, 0.625 and 0.632 as the overall micro F1 score for the Hindi, Meitei and Multilingual datasets.

1 Introduction

The present scenario of social media has opened great opportunities (Liu et al., 2020) in natural language processing. Different social media platforms provide users to express/deliver its opinion exclusively. The post or comments made over it can be affectionate, sarcastic, aggressive, bias, etc (Datta et al., 2020; Baeza-Yates, 2018). Its impact is highly immense, which can lead to serious problem (Baccarella et al., 2018). Understanding and analyzing different topics has become an important area in today's world. It allows researchers with high exposure to various topics, which add growth in the field and to society.

Usage popularity of such platforms extensively implies growth in data availability. Machine learning approaches have gained their recognition (Liu et al., 2020) and played as back-boned in various experiments over social media content analysis.

This experiment is based on the ICON2021 shared task over-identification of aggression and bias related to gender and communal (particularly first subtask). It has provided separate *Hindi, Meitei and Bangla* and multilingual dataset *Combination of all the separate dataset* with English code-mixed for the task. Each dataset consists of 3 different classes, namely *aggressive, gender bias, and communal bias*. The experiment aims to identify classes and their intersectionality among them. Our model is based on frequency-based feature extraction technique (*TFIDF* (Aizawa, 2003)) with hierarchical classifier (*Decision Tree*) (Safavian and Landgrebe, 1991). The obtained accuracy based on micro F1 score is 0.629, 0.625, and 0.632 for the *Hindi, Meitei and Multi* dataset, and this shared task ranking is based on obtaining instance F1 score. Our experiment placed rank at 3rd (Hindi), 2nd (Meitei), and 4th (Multi) with instance F1 score as 0.263, 0.267, and 0.258 respectively for the different datasets.

The rest of the paper is assembled in different sections. Section 2 provides a survey made upon social media content to identify aggression and bias and methodologies implemented. Later Section 3, describe the details of the experiment performed over the shared task, begins with dataset description, technique and model used, and the result with error analysis obtained over this experiment. Last Section 4 draws the conclusion and further scope suggested towards the better outcome of the topic.

2 Literature survey

Aggression, gender and communal bias identification are the new research topics in the field of NLP. Few specific and related work in this topic make use of feature extraction techniques like BOW (Kwok and Wang, 2013), dictionary (Tulkens et al., 2016), word and character level ngram (Pérez and Luque, 2019) and lexicons based (Alorainy et al., 2019; Cryan et al., 2020) ANN-based advance feature embedding techniques such as GloVe (Kumari and Singh, 2020; Zhang et al., 2018; Khan-

Proceedings of the 18th International Conference on Natural Language Processing:
Shared Task on Multilingual Gender Biased and Communal Language Identification, pages 35–40
National Institute of Technology - Silchar, December 16-19, 2021.

delwal and Kumar, 2020), Fast-Text (Kumari and Singh, 2020; Khandelwal and Kumar, 2020; Jha and Mamidi, 2017), Word2Vec (Mossie and Wang, 2020) and BERT (Liu et al., 2020; Minot et al., 2021; Cryan et al., 2020) are also seen reported. Multi-lingual model on aggressive identification using frequency based feature extraction (Khandelwal and Kumar, 2020; Datta et al., 2020; Martinc et al., 2018; Modha et al., 2018) has shown improvement over the earlier methods. Above mentioned techniques observed in gender bias classification (Martinc et al., 2018; Leavy, 2019; Jha and Mamidi, 2017; Cryan et al., 2020). Communal bias text identification is another challenging and new area under NLP. There is comparatively less work related to communal bias text identification, related work includes (Khanday et al., 2021; Chang, 2021; Smith-Vidaurre et al., 2020; Lourie et al., 2021). As mentioned earlier, machine learning algorithm plays a promising role in different classification problems. The data structure and multiclass property of the dataset pulls the attention of hierarchical tree based classification. Decision tree classifier is widely employed with good performance over multiclass problem (Farid et al., 2014; Shao et al., 2013; Polat and Güneş, 2009). Relatively, its implementation over area of text classification like aggression, hatespeech and gender-bias is seen in (Yuvaraj et al., 2021; Modha et al., 2018; Kamiran et al., 2010) and these techniques outperformed in many other text classification task (Khanday et al., 2021; Kamiran et al., 2010; Farid et al., 2014).

3 System architecture

This section discusses the detail of the used dataset provided by the shared task organizer and its implementation.

3.1 Dataset

The dataset for the shared task is a multilingual dataset which comprises of 3 different languages *Meitei, Bangla, Hindi* (Kumar et al., 2021b). Separate dataset was provided for *Meitei, Bangla and Hindi* task. In total, it contains 12000 and 3000 samples for training and testing. It is an annotated dataset with three label *aggression, gender bias, and communal bias* of which aggression is a three-way multiclass problem and *gender bias and communal bias* are binary class problem. Table 1 explains the instance's contribution to the training and validation dataset. However, instances density concerning each class is shown in Table 2, where different 12 combinations are found and demonstrated in the dataset column of the table. Collectively it is a multiclass-multioutput problem, where it comprises of 3 different classes which describe the level of *Aggression, Gender bias, and Communal bias*. Aggression category is a multiclass problem with three different level *OAG: Overtly aggressive, CAG: Covertly aggressive, NAG: Non-aggressive*, whereas other two classes are binary class classification problem with *GEN: gendered, NGEN: non-gendered and COM: communal, NCOM: non-communal.*

Dataset	Training set	Validation set	Testing set
Meitei	2209	1000	1020
Hindi	4615	1000	1002
Bangla	2391	1000	967
Multi-lingual	9214	2997	2989

Table 1: Dataset description with instances figure

3.2 Experiment

The experiment for the shared task is carried out with three major phases, namely, pre-processing, feature extraction, and classification (Kumar et al., 2021a). The pre-processing stage aims to remove words or characters, which represent noise to the dataset. Prior to pre-process step, we explore the dataset and end with a few observations.

- All the instances are mostly short text, and it highly signifies social media content like comments on youtube or Facebook.
- The instances in the dataset for the concern languages are in code-mixed with English.
- Apart from it, the instances in all the dataset represent casual expression and use the shortened expressions.

The first pre-processing step includes converting all the instances to lowercase, resulting in an overall increase in word frequency. This step aims to normalize the valuable samples for the sentiment classification, such as digits having a minor role in sentiment identification. Hence removal of the number is carried out as part of pre-processing step. As mentioned above, all the datasets are code-mixed, and therefore for stopword removal, we consider the English stopword list for the Hindi and the dataset for stopword removal. However, for the Meitei dataset, we add 58 words with minimal

	Hindi		Meitei		Multi	
Dataset	*Train*	*Valid*	*Train*	*Valid*	*Train*	*Valid*
CAG, GEN, COM	5	1	18	5	32	7
CAG, GEN, NCOM	118	20	51	21	291	104
CAG, NGEN, COM	149	28	94	39	289	85
CAG, NGEN, NCOM	528	120	861	406	154	601
NAG, GEN, COM	5	4	0	0	7	6
NAG, GEN, NCOM	116	40	3	0	182	67
NAG, NGEN, COM	35	14	3	1	98	39
NAG, NGEN, NCOM	1133	247	882	369	2672	895
OAG, GEN, COM	63	13	18	9	136	35
OAG, GEN, NCOM	643	147	58	20	135	429
OAG, NGEN, COM	760	136	41	14	932	203
OAG, NGEN, NCOM	1060	230	180	116	1677	536

Table 2: Different instances contribution in both train and validatation for *[hindi, meitei, multi]* languages.

Stopwords	**Lists**
Meitei	adubu,aduga,akhoina,ashhh,asida,asiga, asina,asumna,atoppa,bjp,ebanigi,eduna, ei,eibudi,eibusu,eigee,eigidi,eigita, eihaki,eihakpu,eihakse,eihakti,einadi, elshi,esadi,gonna,gumna,haaaa,haiba, haina,hekta,hoi,hyduna,hyrga,jaaye, karigi,karisu,keino,keisu,khara,ma,makhoi, masibu,nang,nangbu,nangdi,nangga,nanggi, nangi,nangna,nangse,nangsu,ngasidi,ngkna, pakpi,thembi,yaishnagi,yenglk.

Table 3: Meitei dataset stopwords list

sentiment intensity. There is no specific stopword list for Meitei language, however being a native speaker, we identify a few words of a total 58, which contribute minimally in deciding the class of text and shown in table 3. The added terms are purely based on the dataset with high occurrences with a low degree of sentiment, for example, *keino [what], nang [you], nangi [yours], nangga [with you]* etc. The multi-lingual dataset comprises of *Hindi, Meitei and Bangla* languages; therefore, we extend the stopwords list used in the individual Meitei dataset as mentioned above. The social media text often contains *link and references*, and punctuation. In this phase, removing such Html/link and punctuation is carried out. Terms with character lengths less than three usually are less meaningful and contribute high density to the dataset. Social media text, in general, is found to use abbrev terms for the words like *u for you, ng for nang* etc. Usually, these terms bypass the stopword removal step. Part of pre-processing initiates the removal of such terms with a character length less than 3.

Lastly, pre-processing handles the concept of expanding contractions for Metei language and implemented over Meitei and Multi-lingual datasets. Misspell and abbrev terms with character lengths above three are observed with a high degree in the datasets. Collectively 296 words undergo the expansion-contraction phase, where it is normalized to its based form or single acceptable word, example include *ebema, ebenma to ebemma, fhare to phare, hairk to hairak* etc. is normalized [1]. Listed contraction words are highly used in the written form of meitei text.

Feature extraction aims to represent the raw data in a manageable form. This experiment uses frequency-based feature extraction techniques for all the datasets. *TFIDF* is a widely used feature extraction technique in the field of information retrieval. A numerical statistic based on word importance's over the instances or the dataset. A language-independent weighting factor is built on term occurrences for the instances in the dataset. Equation 1 elaborate the TFIDF computation formula, with *t, d, df, n* as the term, document, document frequency and size of dataset.

$$tf.idf(t,d) = \left[\frac{\text{Total count of } t \text{ in } d}{\text{Total words in } d}\right] \times \left[\frac{log(1+n)}{1+df(t)} + 1\right] \quad (1)$$

The classification problem for this experiment is a multitask classification that exhibits a multiclass-multioutput form, where each instance possesses a set of non-binary properties. The estimator needs to operate on several joint classification tasks. This experiment considers the decision tree classifier as the classification algorithm. A non-parametric algorithm is applicable both in classification and regression. A tree-structured classifier based on CART algorithm (classification and regression tree algorithm) with different nodes *root: entire dataset, internal: dataset features with decision rules and*

[1] https://github.com/debinamaibam/Manipuri-contraction-word-list-repository.git

leaf: decision outcome. CART model is a binary tree where two child nodes are formed with every split. The decision tree splitting process is based on the rule set upon decision node result different sub-nodes and tree formation. Lastly, it develops different decision tree nodes with the best attribute and no further possible classification naming the final node as a leaf node. Implementation of the classifier is based on python scikit-learn library, with parameters as random state as 0, Gini criterion for split, and minimum split sample to consider as 2.

3.3 Result Analysis

The experiment begins with the training of 4615, 2209, 9214 instances of *Hindi, Meitei and Multilingual* dataset. A well-designed pre-processing is carried out to filter out words, characters, or links less productive in classification. The processed text is passed for the feature generation stage, where we adopt *TFIDF* as the extraction technique to generate features for each sample based upon occurrences. The feature vector generated for the dataset mentioned above are *[4615 * 38273], [2209 * 40531] and [9214 * 92942]* sizes represented in [X * Y] with X representing the number of instances in the dataset and Y denote the total number of features generated by TFIDF vectorizer. The feature is developed upon the word with unigram range for ngrams and l2 normalization. This normalization technique is used for performance enhancement measures and aims to minimize the mean cost means the sum of the squares of each sample is always up to 1. These features are passed upon decision tree classifier for the classification task. Best attribute selection for the root and sub-nodes is one of the challenging units. Attribute selection measures are established using 2. Equation 2 elaborate the computation of Gini index, where p_i signify probability of instances being classified to particular class. The purity and impurity are measured during tree creation in CART.

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2 \tag{2}$$

$$MicroF1 - score = 2 * \frac{micro - precision * micro - recall}{micro - precision + micro - recall} \tag{3}$$

The experiment is executed to estimate the tree split quality as Gini, with minimum samples needed for split as two and minimum leaf node as 1. Result validation is based upon the micro-F1 score 3. Micro F1-score measures aggregated contributions of all the classes, where 1 denotes the best score and 0 as the worst. Overall and Individual micro-F1 scores for each of the class is returned. Table 4 displays the achievement accuracy of the model over different datasets.

3.4 Error analysis

All three datasets possess 12 class combinations with a high imbalance nature of class density, as shown in Table 2. Imbalance of dataset sequel in the classifier performance degradation. For example, the model is trained with 1024, 888, 297 samples as CAG, OAG, and NAG for aggressive class and 174 and 2035 samples as COM and NCOM for communal bias in the meitei dataset, which shows a clear imbalance nature. There exist techniques like *resampling* to work out such an issue. However, for this dataset, implementing an oversample or undersample might affect the other way, as each sample is linked to 3 labels with 12 different combinations. Therefore, we bypass the resampling technique to maintain data originality and proceed risk-free. Another possible factor to compromise with the selected classifier is, of the three classes, one class behaves multilabel and the other two as a binary class, resulting in the classification task as the multiclass-multioutput problem.

4 Conclusion

Related to the ICON-2021 shared task, we participated in subtask 1 on multilingual gender-biased and communal language identification for the Hindi, Meitei, and multilingual datasets. Our system is built upon the TFIDF feature technique with Decision Tree as a classifier and obtained an F1-score of 0.629, 0625, and 0.632. In the future, we aim to build the multilingual model by embedding relative lexicon and enhancing frequency-based features extensively.

References

Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65.

Wafa Alorainy, Pete Burnap, Han Liu, and Matthew L Williams. 2019. "the enemy among us" detecting cyber hate speech with threats-based othering lan-

Dataset	instances F1	Overall micro F1	Agression micro-F1	GenderBias micro-F1	CommunalBias micro-F1
Hindi	0.263	0.629	0.479	0.726	0.682
Meitei	0.267	0.625	0.344	0.682	0.849
Multi	0.258	0.632	0.413	0.694	0.791

Table 4: Model performance over different datasets

guage embeddings. *ACM Transactions on the Web (TWEB)*, 13(3):1–26.

Christian V Baccarella, Timm F Wagner, Jan H Kietzmann, and Ian P McCarthy. 2018. Social media? it's serious! understanding the dark side of social media. *European Management Journal*, 36(4):431–438.

Ricardo Baeza-Yates. 2018. Bias on the web. *Communications of the ACM*, 61(6):54–61.

Li-jing Arthur Chang. 2021. Detecting asian values in asian news via machine learning text classification. In *Advances in Data Science and Information Engineering*, pages 123–128. Springer.

Jenna Cryan, Shiliang Tang, Xinyi Zhang, Miriam Metzger, Haitao Zheng, and Ben Y Zhao. 2020. Detecting gender stereotypes: Lexicon vs. supervised learning methods. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–11.

Anisha Datta, Shukrity Si, Urbi Chakraborty, and Sudip Kumar Naskar. 2020. Spyder: Aggression detection on multilingual tweets. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 87–92.

Dewan Md Farid, Li Zhang, Chowdhury Mofizur Rahman, M Alamgir Hossain, and Rebecca Strachan. 2014. Hybrid decision tree and naïve bayes classifiers for multi-class classification tasks. *Expert systems with applications*, 41(4):1937–1946.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*, pages 7–16.

Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*, pages 869–874. IEEE.

Akib Mohi Ud Din Khanday, Qamar Rayees Khan, and Syed Tanzeel Rabani. 2021. Identifying propaganda from online social networks during covid-19 using machine learning techniques. *International Journal of Information Technology*, 13(1):115–122.

Anant Khandelwal and Niraj Kumar. 2020. A unified system for aggression identification in english code-mixed and uni-lingual texts. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pages 55–64.

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.

Kirti Kumari and Jyoti Prakash Singh. 2020. Ai_ml_nit_patna@ trac-2: Deep learning approach for multi-lingual aggression identification. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 113–119.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence*.

Susan Leavy. 2019. Uncovering gender bias in newspaper coverage of irish politicians using machine learning. *Digital Scholarship in the Humanities*, 34(1):48–63.

Han Liu, Peter Burnap, Wafa Alorainy, and Matthew Williams. 2020. Scmhl5 at trac-2 shared task on aggression identification: Bert based ensemble learning approach.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32, 000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13470–13479.

Matej Martinc, Blaz Skrlj, and Senja Pollak. 2018. Multilingual gender classification with multi-view deep learning. In *Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*.

Joshua R Minot, Nicholas Cheney, Marc Maier, Danne C Elbers, Christopher M Danforth, and Peter Sheridan Dodds. 2021. Interpretable bias mitigation for textual data: Reducing gender bias in patient notes while maintaining classification performance. *arXiv preprint arXiv:2103.05841*.

Sandip Modha, Prasenjit Majumder, and Thomas Mandl. 2018. Filtering aggression from the multilingual social media feed. In *Proceedings of the first*

workshop on trolling, aggression and cyberbullying (TRAC-2018), pages 199–207.

Zewdie Mossie and Jenq-Haur Wang. 2020. Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, 57(3):102087.

Juan Manuel Pérez and Franco M Luque. 2019. Atalaya at semeval 2019 task 5: Robust embeddings for tweet classification. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 64–69.

Kemal Polat and Salih Güneş. 2009. A novel hybrid intelligent method based on c4. 5 decision tree classifier and one-against-all approach for multi-class classification problems. *Expert Systems with Applications*, 36(2):1587–1592.

S Rasoul Safavian and David Landgrebe. 1991. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674.

Yuan-Hai Shao, Wei-Jie Chen, Wen-Biao Huang, Zhi-Min Yang, and Nai-Yang Deng. 2013. The best separating decision tree twin support vector machine for multi-class classification. *Procedia Computer Science*, 17:1032–1038.

Grace Smith-Vidaurre, Marcelo Araya-Salas, and Timothy F Wright. 2020. Individual signatures outweigh social group identity in contact calls of a communally nesting parrot. *Behavioral Ecology*, 31(2):448–458.

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A dictionary-based approach to racism detection in dutch social media. *arXiv preprint arXiv:1608.08738*.

Natarajan Yuvaraj, Victor Chang, Balasubramanian Gobinathan, Arulprakash Pinagapani, Srihari Kannan, Gaurav Dhiman, and Arsath Raja Rajan. 2021. Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification. *Computers & Electrical Engineering*, 92:107186.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*, pages 745–760. Springer.

LUC at ComMA-2021 Shared Task: Multilingual Gender Biased and Communal Language Identification without using linguistic features

Rodrigo Cuéllar-Hidalgo
CENIDET/COLMEX
rcuellar@colmex.mx

Julio de Jesús Guerrero-Zambrano
IIM
julio@jzambrano.xyz

Dominic Forest
Université de Montréal
EBSI
dominic.forest@umontreal.ca

Gerardo Reyes-Salgado
CENIDET
gerardo.rs@cenidet.tecnm.mx

Juan-Manuel Torres-Moreno
Université d'Avignon
LIA
juan-manuel.torres@univ-avignon.fr

Abstract

This work aims to evaluate the ability that both probabilistic and state-of-the-art vector space modeling (VSM) methods provide to well known machine learning algorithms to identify social network documents to be classified as aggressive, gender biased or communally charged. To this end, an exploratory stage was performed first in order to find relevant settings to test, i.e. by using training and development samples, we trained multiple algorithms using multiple vector space modeling and probabilistic methods and discarded the less informative configurations. These systems were submitted to the competition of the ComMA@ICON'21 Workshop on Multilingual Gender Biased and Communal Language Identification.

1 Introduction

The introduction of the Internet and its democratization in the public sphere has fostered the emergence of many sociological phenomena. This opens the possibility of forming friendly relations and information sharing from online networking platforms (Arroyo-Fernández et al., 2018). As the organizers say: *"Aggression and its manifestations in different forms have taken unprecedented proportions with the tremendous growth of Internet and social media."* [1]. The challenge ComMA@ICON 2021 is an interesting task in order to automatically discover and understand the pragmatic and structural aspects of such forms of language usage (Waseem et al., 2017; Dadvar et al., 2013).

Our international LUC[2] team[2] have worked in such challenge using several approaches mainly without linguistic features.

This paper is organized as follows: the section 2 presents some relevant state of the art work related to automatic aggression identification, section 3 describes the methodology and the ComMA@ICON'21 dataset used, section 4 shows the results and finally the section 5 analyzes and discusses the contributions of this paper.

2 Related work

There already exist software designed to identify aggression and cyberbullying in social medias, e.g. CyberPatrol. The main drawback of these systems is that they are based on keyword filtering, which is a limitation because no statistical features of texts are taken into account. Further, these keyword filtering methods require manual maintenance. To overcome the limitations of keyword filtering systems, (Yin et al., 2009) is one of the former attempts to detect cyberbullying by using statistical features: word frequency, analysis of feelings (use of pronouns in the second person, insults, etc.) and context. (Dinakar et al., 2021) built a system that can detect bullying elements in commentaries of YouTube videos. These were classified according to different representative categories (sexuality, intelligence, race and physical attributes). The classification revealed weaknesses and an increase in false positive cases. Researchers emphasized the importance of using common sense to understand users' goals, emotions, and relationships, thereby disambiguating and contextualizing language. In (Berry and Kogan, 2010) the authors were also interested in a word search method based on a bag of words (BoW) system incorporating sentiment and contextual analysis. They build a decision tree that predicted intimidating messages with an accuracy rate of 0.93. The researchers also have developed

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

[2] Team composed by LIA (Laboratoire Informatique d'Avignon/Université d'Avignon, France), EBSI/Université de Montréal (Québec, Canada), CENIDET (Centro Nacional de Investigación y Desarrollo Tecnológico, Cuernavaca, México) and COLMEX (the Colegio de México, CDMX, México).

Proceedings of the 18th International Conference on Natural Language Processing: Shared Task on Multilingual Gender Biased and Communal Language Identification, pages 41–45
National Institute of Technology - Silchar, December 16-19, 2021.

the Chatcoder software to detect malicious activities online (Kontostathis et al., 2012).

Another study tested a system that allows users of a website to control the messages posted on their web pages: it customized vocabulary filtering criteria using a machine learning method that automatically labeled the contents. This approach had limitations because it was unable to measure relationships between terms beyond a certain semantic level (Dadvar et al., 2012). (Nahar et al., 2014) provided a concrete method for detecting online harassment by measuring the score of sent and received messages (and thus their degree of involvement in a conversation) using the Hyper-link Induced Topic Research algorithm (HITS). The authors also proposed a graphical model that identifies the aggressors and their most active victims. Other studies have attempted to go further by seeking to take into account more specific characteristics. (Dadvar et al., 2012) tried to establish a system based on language features characterizing the author's genre of comments on MySpace. Their results revealed an improvement in the detection of bullying when this information is taken into account. As we can see, recent work defines the means to respond to the cyberbullying phenomenon that is becoming more and more widespread as the use of the web does.

The problem of identifying offensive and abusive language is a more difficult task than expected due to the prevalence of 5 factors, according to (Nobata et al., 2016), described below:

1. The intentional obfuscation of words that lead to false positives.
2. The difficulty in identifying racial slurs since depending on the target group, this can be offensive or flattering.
3. The grammatical fluidity that leads to false negatives.
4. The limit of sentences, that is, abusive language can be articulated in more than one sentence.
5. The sarcasm, which is even difficult for a human to interpret, implies a lot of knowledge about the context of the message (geographical, historical, social, etc).

Other aspects detected by Nobata et al. (2016) corresponds to the heterogeneity in the approach to the problem itself that causes too much noise and confusion, such as the fact of only addressing specific aspects, specific contexts, different definitions for certain terms and / or domain, and finally different sets of assessment.

At present, the task of classifying text in "agglutinating" or "morphologically rich" languages, leaves aside classical preprocessing and is replaced by the use of deep neural networks and word embeddings, since they take into account the semantic distance of the words in context, which is very useful in categorization tasks, which is not the case with the classic bag of words methods. A clear example of this is the implementation of fastText for the Turkish language by Kuyumcu et al. (2019).

3 Methodology and data-sets

In this section we present the methodology and the data-sets used in our study.

3.1 Data-sets

For this task, the ComMA organizers have provided a multilingual data-set with a total of 12,000 instances for training and development and an overall 3,000 instances for testing. The corpus is in three Indian languages Meitei, Bangla (an Indian variety) and Hindi and English. Several instances are expressed in two our more languages. Refer to the challenge website for further information[3].

From the organizer's website, the corpus is labelled as follows:

1. **Aggression level**. This category gives a classification in 'Overtly Aggressive'(OAG), 'Covertly Aggressive'(CAG) or 'Non-aggressive'(NAG) text.
2. **Gender**. This category classify the text as 'gendered'(GEN) or 'non-gendered'(NGEN).
3. **Communal**. The task is to develop a binary classifier for classifying the content as 'communal' (COM) or 'non-communal'(NCOM).

We confirmed that the data-set furnished (both training and validation) are too small in order to employ Deep Learning methods. We then chose to use mainly classical probabilistic and VSM (Salton et al., 1983) oriented algorithms.

[3] https://competitions.codalab.org/competitions/35482#learn_the_details-datasets

Corpus	Instances	Tokens	Chars
Training	9,000	186,017	1 585,979
Developing	3,000	55,996	473,403
Testing	3,000	82,367	815,104

Table 1: Basic statistics from ComMA corpus.

3.2 Pre-processing

Because there are a mixture of several languages (Meitei, Bangla, Hindi and English) and often the data-set instances presents two or more languages mixed, we decided of avoid any linguistic pre-processing. Indeed, neither stemmer, filtering or lemmatizer was used in our study. The only pre-processing that was carried out was the removal of NaN and extraction of basic characteristics of each message, which are listed below:

- Number of words.
- Number of sentences.
- Number of scores.
- Number of numbers.
- Number of unrecognizable characters (emojis).

Using a simple tokenizer written in Python.

3.3 Algorithms

To tackle the problem presented in this challenge, we develop LUC, a multi-classifier that uses the following algorithms:

- Nearest-Neighbor algorithm (KNN) (Altman, 1992);
- Naive Bayes method (Lewis, 1998);
- Support Vector Machine algorithm (SVM) (Cortes and Vapnik, 1995);
- Random Forest algorithm (Breiman, 2001);
- Generalized Boosted Regression Models (GBM) [4];
- Adaboost (Freund and Schapire, 1997);
- Neural Networks (NN) based algorithms (Hopfield, 1982).

[4]https://github.com/gbm-developers/gbm

In the first system (S1), the individual outputs of the algorithms Naive Bayes, SVM, Random Forest, GBM, Adaboost and a multi-layer perceptron were combined in a single output using a mixing strategy that assembles all of the models that were created using the previous algorithms. In order to combine the predictions of the previously mentioned estimators, it was necessary to stack the outputs of each individual estimator and use a final estimator to compute the final prediction.

The stacking classifier responsible to compute the final estimation takes every individual estimator as input and creates a final estimator by training cross-validated predictions of the base estimators. For each estimator, the final classifier computes the prediction probability and final prediction, to return a final estimation based on a logistic regression of the inputs.

In order to achieve better results, each one of the tasks were trained and executed independently (by language and category) and the results were combined at the end. Accuracy was measured in order to keep track of the metrics of the results.

In the second system (S2), we also explored the relevance of the K nearest neighbors (KNN) algorithm alone to perform these supervised classification tasks. This algorithm is well known in the field of machine learning. It is both simple and efficient. It consists in assigning to each document of the test set the category with the highest frequency among its K nearest neighbors. The cosine measurement was used to evaluate the distance between the vectors representing each document. In addition, we varied two main parameters during the learning phase. On the one hand, we have varied the value of K, that is to say the number of neighbors to be considered. We systematically compared the results using 1, 2, 3, 4, 5, 10, 15, 20, 25 and 50 neighbors. We also varied the number of features to describe the documents. As mentioned previously, no preprocessing was applied to reduce the size of the initial lexicon. Based on the frequency of strings in the entire corpus and by evaluating the correlation between the most frequent strings and the categories to predict (using a simple Chi2 measure), we used from 500 to 30,000 strings of characters (in increments of 500) to describe the training corpus. During the learning phase, we obtained the best results using 30,000 features and K = 1. It is therefore this combination that we applied to the test corpus (system 5).

	KNN (S2)	Combined (S1)
Accuracy	0.98	0.7746

Table 2: Developing set evaluation of our best classifiers.

F1	KNN (S2)	Combined (S1)
Instance F1	0.234	0.213
Overall micro-F1	0.615	0.616
A. micro-F1	0.446	0.389
G. bias micro-F1	0.675	0.693
C. bias micro-F1	0.726	0.766

Table 3: Testing set evaluation for our classifiers. See: https://competitions.codalab.org/competitions/public_submissions/35482.

4 Results

We present our results in two parts. In the first one we show the accuracy of methods on the training set. In the second one, the official F1 measures from the CodaLab website of the organizers are showed in their "Results" page[5].

4.1 Training and Developing data-sets

For the training set, our main results are showed in the Table 2.

In the training phase, the accuracy was our principal criterion to optimize on the Developing dataset.

4.2 Testing set

The main results on the testing set are shown in the Table 3. It is possible to see that our classifier LUC obtains the 3rd rank on the multilingual task, with a overall-F1 measure of **0.234** (see Addendum).

5 Discussion and Conclusions

This task was very challenging. The corpus (train, developing and test) are very noisy at several levels (syntactic, lexical, semantic, etc.), and the instances presents two or more languages mixed. There are instances that make no sense. By example, the instances from train and developing sets:

```
C575.31 rrrrrrrrrrrrrrrrrrrrrrrrr
rrrrrrrrrrrrrrrrrrrrrrrrrrrrrurrr
rrrrrrrrrrrrryrrrr (NAG,NGEN,NCOM)

C575.690.1  XXXXXXXXXXXXXXXXXXXXX
(NAG,NGEN,NCOM)

C579.732.1 G (NAG,NGEN,COM)
```

where the string of "r" of the instance C575.31 represents a string of a dog's emojis, is a good example of semantic noise. What was the logic (if any) of the annotators in assigning the classes (NAG,NGEN,NCOM) to this instance?

However, our strategy has obtained the third rank on overall-F1 measure and the second rank on the Aggression micro-F1 measure. The methodology that we adopted in this study allowed us to observe in broad strokes the complexity of the challenge.

We think that our model may be enriched by other model sentences representation, as followed in (Arroyo-Fernández et al., 2019), in order to outperform our results. Finally, sentences classified with labels 'aggression', 'genre' or 'communal levels' may be used in other NLP tasks, by example on the summaries generation guided by a specific context and coming from social network text data (Torres-Moreno, 2014).

Acknowledgments

This work was partially financed by the COLMEX (Mexico), CENIDET (Mexico), Instituto de Ingenieros de Morelos (Mexico), LIA (Université d'Avignon, France) and by the EBSI (Université de Montréal, Québec).

Addemdum

We dispute the results mentioned on the challenge's website. Indeed, the organizers of the challenge allowed certain teams to benefit from an additional day for the submission of their result. This violates the original rules of the challenge and has favored some teams. Strangely, the teams that benefited from this advantage finished in the top 3 positions. We therefore consider, by virtue of the initial regulations that we respected, that we finished in 3rd place in the task on the multilingual corpus. We are also very amazed at how the results have evolved over time. The results were first published online on the challenge website (we were then 3rd), then these results were modified (we were then ranked 5th). Finally, the non-public excel sheet was sent to the teams with new results, without any explanation (we ended up in 6th place).

[5]See: https://competitions.codalab.org/competitions/35482#results

References

N. S. Altman. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

Ignacio Arroyo-Fernández, Dominic Forest, Juan-Manuel Torres-Moreno, Mauricio Carrasco-Ruiz, Thomas Legeleux, and Karen Joannette. 2018. Cyberbullying detection task: the EBSI-LIA-UNAM system (ELU) at coling'18 TRAC-1. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, TRAC@COLING 2018, Santa Fe, New Mexico, USA, August 25, 2018*, pages 140–149. Association for Computational Linguistics.

Ignacio Arroyo-Fernández, Carlos-Francisco Méndez-Cruz, Gerardo Sierra, Juan-Manuel Torres-Moreno, and Grigori Sidorov. 2019. Unsupervised sentence representations as word information series: Revisiting tf–idf. *Computer Speech Language*, 56:107–129.

Michael W. Berry and Jacob Kogan, editors. 2010. *Text Mining: Applications and Theory*. Wiley, Chichester, UK.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

M. Dadvar, Franciska M.G. de Jong, Roeland J.F. Ordelman, and Rudolf Berend Trieschnigg. 2012. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*, pages 23–25. Ghent University. Null ; Conference date: 24-02-2012 Through 24-02-2012.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *Advances in Information Retrieval*, pages 693–696, Berlin, Heidelberg. Springer Berlin Heidelberg.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2021. Modeling the detection of textual cyberbullying. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(3):11–17.

Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

J J Hopfield. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558.

April Kontostathis, Andy Garron, Kelly Reynolds, Will West, and Lynne Edwards. 2012. Identifying predators using chatcoder 2.0. In *CLEF*.

Birol Kuyumcu, Cuneyt Aksakalli, and Selman Delil. 2019. An automated new approach in fast text classification (fasttext): A case study for turkish text classification without pre-processing. *PervasiveHealth: Pervasive Computing Technologies for Healthcare*, pages 1–4.

David D. Lewis. 1998. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, 1398, pages 4–15, Chemnitz, DE. Springer Verlag, Heidelberg, DE.

Vinita Nahar, Xue Li, Hao Lan Zhang, and Chaoyi Pang. 2014. Detecting cyberbullying in social networks using multi-agent system. *Web Intell. Agent Syst.*, 12(4):375–388.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153. International World Wide Web Conferences Steering Committee.

Gerard Salton, Edward A Fox, and Harry Wu. 1983. Extended boolean information retrieval. *Communications of the ACM*, 26(11):1022–1036.

Juan-Manuel Torres-Moreno. 2014. *Automatic Text Summarization*. ISTE Ltd, John Wiley & Sons, Inc., London.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian Davison, April Edwards, and Lynne Edwards. 2009. Detection of harassment on web 2.0.

ARGUABLY at ComMA@ICON: Detection of Multilingual Aggressive, Gender Biased, and Communally Charged Tweets using Ensemble and Fine-Tuned IndicBERT

Guneet Singh Kohli, Prabsimran Kaur, Dr. Jatin Bedi
Computer Science and Engineering Department,
Thapar Institute of Engineering and Technology, Patiala, Punjab, India

Abstract

The proliferation in Social Networking has increased offensive language, aggression, and hate-speech detection, which has drawn the focus of the NLP community. However, people's difference in perception makes it difficult to distinguish between acceptable content and aggressive/hateful content, thus making it harder to create an automated system. In this paper, we propose multi-class classification techniques to identify aggressive and offensive language used online. Two main approaches have been developed for the classification of data into aggressive, gender biased, and communally charged. The first approach is an ensemble-based model comprising of XGBoost, LightGBM, and Naive Bayes applied on vectorized English data. The data used was obtained using an Indic Transliteration on the original data comprising of Meitei, Bangla, Hindi and English language. The second approach is a BERT-based architecture used to detect misogyny and aggression. The proposed model employs IndicBERT Embeddings to define contextual understanding. The results of the models are validated on the ComMA v 0.2 dataset.

1 Introduction

A burgeon in Social Networking has been seen in the past few years. The number of platforms and users has increased by 77% from 2014 to 2021. Social Media, due to its easy accessibility and freedom of use, has transformed our communities and how we communicate. One of the widespread impacts can be seen through trolling, cyberbullying, or sharing aggressive, hateful, misogynistic content vocalized through platforms like Facebook, Twitter, and YouTube. The intensity and hostility lying in aggressive words, abusive language, or hate speech is a matter of grave concern. These are used to harm the victim's status, mental health, or prestige (Beran and Li, 2005; Culpeper, 2011). This articulation of hatefulness often travels from the online to the offline domain, resulting in organized riot-like situations and unfortunate casualties, which causes disharmony in society. Hence, it has become crucial for scholars and researchers to take the initiative and find methods to identify the source and articulation of aggression.

Aggression is a feeling of anger or antipathy that results in hostile or violent behavior and readiness to attack or confront. According to (Kumar et al., 2018c), one can express aggression in a direct, explicit manner (Overtly Aggressive) or in an indirect, sarcastic way (Covertly Aggressive). Hate speech can be used to attack a person or a group of people based on their color, gender, race, sexual orientation, ethnicity, nationality, religion (Nockleyby, 2000). Misogyny or Sexism is a subset of hate speech '(Waseem and Hovy, 2016) and targets the victim based on gender or sexuality (Davidson et al., 2017; Bhattacharya et al., 2020).

While it is essential to identify hate speech in social networks, it is rather time-consuming to perform manually, considering the massive amount of data at hand. Thus, there is a need to build an automated system for the identification of such aggression. However, distinguishing between acceptable content and hateful content is challenging due to the subjectivity of definitions and varying perceptions of the same content by different people, thus making it tedious to build an automated AI system. Regardless, numerous studies exist that have explored different aspects of hateful and aggressive language and their computational modeling and automatic detection, such as toxic comments. To this end, several workshops such as 'Abusive Language Online' (ALW) (Roberts et al., 2019), 'Trolling, Aggression and Cyberbullying' (TRAC) (Kumar et al., 2018b), and Semantic Evaluation (SemEval) shared task on Identifying Offensive Language in Social Media (OffensEval) (Zampieri et al., 2020)

Proceedings of the 18th International Conference on Natural Language Processing:
Shared Task on Multilingual Gender Biased and Communal Language Identification, pages 46–52
National Institute of Technology - Silchar, December 16-19, 2021.

have been organized.

This paper presents our system for Shared Task on "Multilingual Gender Biased and Communal Language Identification @ ICON 2021" (Kumar et al., 2021a). Two approaches have been implemented developed for the classification of data into **aggressive**, **gender biased**, or **communally charged**.

1. An ensemble-based model comprising of XGBoost, LightGBM, and Naive Bayes was applied on vectorized English data. This data was obtained using an Indic Transliteration on the original data comprising of Meitei, Bangla, Hindi and English language.

2. A BERT-based architecture to detect misogyny and aggression. The proposed model employs IndicBERT Embeddings to define contextual understanding.

2 Related Work

Recently there has been an increase in the studies exploring different aspects of hate speech, sexism detection, aggressive language, and their computational modeling and automatic detection, such as trolling (Cambria et al., 2010; Kumar et al., 2014; de la Vega and Ng, 2018; Mihaylov et al., 2015), racism (Greevy and Smeaton, 2004; Greevy, 2004; Waseem, 2016), online aggression (Kumar et al., 2018a), cyberbullying (Xu et al., 2012; Dadvar et al., 2013), hate speech (Kwok and Wang, 2013; Djuric et al., 2015; Burnap and Williams, 2015; Davidson et al., 2017; Malmasi and Zampieri, 2017, 2018; Waseem and Hovy, 2016), and abusive language (Waseem et al., 2017; Nobata et al., 2016; Mubarak et al., 2017). The prevalent misogynistic and sexist comments, posts, or tweets on social media platforms have also come into light. (Jha and Mamidi, 2017) analyzed sexist tweets and categorized them as hostile, benevolent, or other. (Sharifirad and Matwin, 2019) provided an in-depth analysis of sexist tweets and further categorized them based on the type of harassment. (Frenda et al., 2019) performed linguistic analysis to detect misogyny and sexism in tweets.

Prior studies have explored aggressive and hateful language on platforms like Twitter (Xu et al., 2012; Burnap and Williams, 2015; Davidson et al., 2017). Using Twitter data, (Kwok and Wang, 2013) proposed a supervised approach to categorize the text into racist and non-racist labels to detect anti-black hate speech on social media platforms. (Burnap and Williams, 2015) used an ensemble-based classifier to capture the grammatical dependencies between words in Twitter data to anticipate the increasing cyberhate behavior using statistical approaches. (Nobata et al., 2016) curated a corpus of user comments for abusive language detection and applied machine learning-based techniques to identify subtle hate speech. (Gambäck and Sikdar, 2017) used convolutional layers on word vectors to detect hate speech. (Parikh et al., 2019) provided the largest dataset on sexism categorization and applied a BERT based neural architecture with distributional and word level embeddings to perform the classification task. BERT based approaches also have become prevalent recently (Nikolov and Radivchev, 2019; Mozafari et al., 2019; Risch et al., 2019).

There have also been an increasing number of shared Tasks on Agression Indentification. (Kumar et al., 2018a) aimed to identify aggressive tweets in social media posts in Hindi and English datasets. (Samghabadi et al., 2018) used lexical and semantic features and logistic regression for the Hindi and English Facebook datasets. (Orasan, 2018) used machine learning methods such as SVM and random forest on word embeddings for aggressive language identification. (Raiyani et al., 2018) used fully connected layers on highly pre-processed data. (Aroyehun and Gelbukh, 2018)Aroyehun and Gelbukh (2018) used data augmentation and deep learning for aggression identification.

3 Task Description

The shared task focuses on the multi-label classification to identify the different aspects of aggression and offensive language usage on social media platforms. We have been provided with a multilingual, ComMA v 0.2 (Kumar et al., 2021b) dataset consisting of 12,000 samples for training and an overall 3,000 samples for testing in four Indian languages **Meitei, Bangla, Hindi, and English**. We were required to classify each sample into one of the following labels: aggressive, gender biased, and communally charged.

3.1 Sub-Task A

The first task focuses on aggression identification. It requires us to develop a classifier that can classify the text into **'Overtly Aggressive'(OAG), 'Covertly Aggressive'(CAG),** and

Example	Language	Original	Transliterated	Label
1	Bangla	Media Tao bikrri hoye giyeche	Media Tao bikrri hoye giyeche	*en*
2	Bangla	গরুর মুতখাছ্তো।	Garura muta khācchē.	*ba*
3	Hindi	नंगे घूम हमे क्या	nange ghoom hame kya	*hi*
4	Hindi	Bjp bhagayo Des bachayo	Bjp bhagayo Des bachayo	*en*
5	English	Very nice new	Very nice new	*en*

Figure 1: Examples of the data in the provided dataset and the transliteration performed

'Non-aggressive'(NAG).

3.2 Sub-Task B

The second task deals with aggression identification. It requires us to develop a binary classifier that can classify the text as **'gendered'(GEN)** or **'non-gendered'(NGEN)**.

3.3 Sub-Task C

The third task focuses on aggression identification. It requires us to develop a binary classifier that can classify the text as **'communal' (COM)** and **'non-communal'(NCOM)**.

4 Methodology

4.1 Data Preparation

To get better accuracy, we require a dataset in English language. Therefore, the multilingual input dataset have been passed through the spacy-langdetect toolkit[1]. This toolkit consists of a pipeline for custom language detection. The sentence is categorized into the language it belongs to, i.e., Hindi, Bangla, or English, depending upon the probability assigned to that sentence. The sentences belonging to the Hindi language were given the label "hi," those belonging to Bangla were given the label "ba," and sentences in English were given the label "en." All the sentences belonging to the "hi" and "ba" labels were transliterated, the process of transferring a word from the alphabet of one language to another, to provide us with a uniform multilingual dataset in English.

We must note that the labeling done is based on the language it is written in (as shown in example 3 Figure 1) rather than the language itself (as shown in example 1 Figure 1), which indicates that if the words used are those of English, irrespective of the language, it will be given the label "en". Such sentences do not require transliteration. This data thus prepared has been used in both the proposed architectures as discussed below.

[1]https://spacy.io/universe/project/spacy-langdetect/

4.2 Boosted Voting Ensembler

Machine learning algorithms generally require a numerical input; however, the data is in text form. Thus, the data must be converted to its numerical representation. Count Vectorization technique was

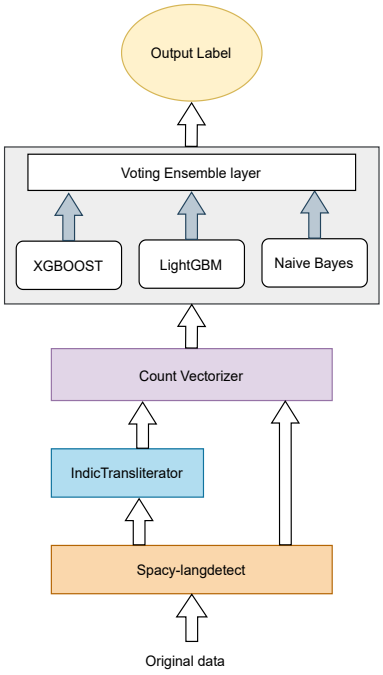


Figure 2: Architecture of Boosted Voting Ensembler

used to transform the data into a vector based on the frequency (count) of each word that occurs in the entire text. It creates a matrix in which a column of the matrix represents each unique word, and each text sample from the document is a row in the matrix. The value of each cell is nothing but the count of the word in that particular text sample. This matrix is then passed through the state-of-the-art models, XGBoost, LightGBM, and the traditional Naive Baye that form the ensemble voting classifier. Each individual model gives a label to the sentence and the number of labels with the highest vote is chosen as the final label.

Language	Instance	Overall micro	Aggression micro	Gender Bias	Communal Bias
Bangla	0.252	0.659	0.442	0.669	0.866
Hindi	0.161	0.582	0.402	0.702	0.642
Multilingual	0.165	0.59	0.361	0.632	0.777

Table 1: Test Results obtained from Boosted Voting Ensembler approach

Language	Instance	Overall micro	Aggression micro	Gender Bias	Communal Bias
Bangla	0.204	0.668	0.341	0.732	0.876
Hindi	0.098	0.625	0.439	0.796	0.639
Multilingual	0.153	0.566	0.357	0.558	0.783

Table 2: Test Results obtained from IndicBERT approach

4.3 IndicBERT Fine-Tuned

For initializing weights of the ALBERT layer, we use "ai4bharat/indic-bert"[2] pre-trained weights for English, Hindi, and Bengali. Before feeding the data into IndicBERT transformer architecture, it must be encoded. Encoding involves the tokenization and padding of sentences to the maximum specified length, which was 150 in our case. In case the length of the sentence exceeds 150, then the sentence is truncated. The encoded sentences

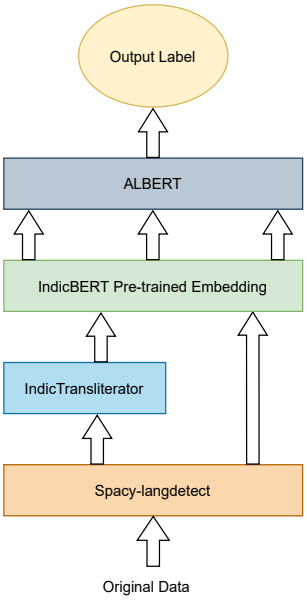


Figure 3: Architecture of Fine-Tuned IndicBERT

are then processed to yield contextually rich pre-trained embeddings. The embeddings are then passed through the IndicBERT transformer, a multilingual ALBERT model trained on large-scale corpora, covering 12 major Indian languages, which gives us the final label.

[2]https://indicnlp.ai4bharat.org/indic-bert/

5 Experimentation and Results

5.1 Boosted Voting Ensembler

The pre-processed data obtained was passed through the voting classifier comprising of xgboost, LightGBM, and conventional Multinomial Naïve Bayes which calculated the outputs from individual models and performed voting to yield the final label. The proposed approach was tested on the three variations of the dataset namely: Multilingual Hindi, Meitei, English, Bangla, purely Hindi, and purely Bangla text. Three sets of label classifications i.e., Aggression, Gender Bias, and Communal Bias were involved corresponding to each sentence which had to be predicted using the proposed pipeline. In reference to Table 1, it can be observed that the Aggression analysis attributed to relatively lower F1 scores of 0.361 in Multilingual, 0.442 in Bangla, 0.402 in Hindi which corresponds to the fact that the various categories of Aggressions tend to have overlapping contextual meanings which are difficult to segregate while performing the classification task. The Gender Bias and Communal Bias being Binary classification tasks observed significantly higher F1 scores in comparison to the aggression task and also showed the strength of our proposed approach to handle these specific category use cases. From the table it can be seen that in Gender Bias the F1 scores achieved for multilingual is 0.632, for Bangla its 0.669, and for Hindi 0.702 whereas in the case of Communal Bias these scores move even higher except in the case Hindi i.e., the F1 scores achieved for multilingual is 0.777, for Bangla its 0.866 and for Hindi 0.642. Overall, the model performance is satisfactory in the binary classification task of Gender and Communal Bias prediction however the results observe a significant fall when dealing

with aggression analysis which highlights the shortcomings of the system in handling the overlapping context among the three aggression labels. The application of Ensemble in the given problem helps us in leveraging the individual powers of XGBoost, LightGBM, and Naïve Bayes and yields results that are more robust and can handle the unknown inputs better. In the future, the inclusion of better embeddings like glove and BERT which capture the underlying semantic and lexical relations could improve the performance of the methodology manifolds.

5.2 IndicBERT

In this section, we discuss the performance of Indic Bert methodology on the processed data. The approach was again tested upon the multilingual, Hindi, and Bangla data, and the observed results are highlighted in Table 2. The Indic Bert is able to achieve an F1 score of 0.558 for multilingual, 0.796 for Hindi, and 0.732 for Bangla in the case of Gender Bias. For the communal bias, the same high-performing trend can be observed with Indic Bert generating scores of 0.876 in Bangla, 0.639 in Hindi, and 0.783 for multilingual. The Aggression analysis again came out as the low-performing task with Indic Bert giving scores of 0.341 for Bangla,0.439 for Hindi, and 0.357 for the Multilingual data. The system performed well in many tasks when compared with the ensemble technique especially in handling the binary classification tasks. However, this pipeline again lacks in performing well on the aggression tasks thus highlighting the shortcomings in handling contextual overlaps in many sentences.

5.3 Comparisons

On close observations of results of both the pipelines the Indic Bert seems to have performed well in individual tasks. For Aggression Analysis Indic Bert outperforms the Ensemble approach in multilingual data and Hindi data. In Gender Bias Indic Bert takes the lead for Hindi and Bangla data and for Communal Bias it beats the Ensemble technique in Bangla and Multilingual data. Though Indic Bert seems to be outperforming the Ensemble approach in more individual tasks the instance F1 score indicates the performance of the model in predicting the three categories together is higher for the ensemble model than its deep learning counterpart. The instance F1 scores for all the languages is higher for the ensemble approach which shows its adaptability over all the categories together. Indic Bert takes lead in Bangla and Hindi in the case of overall micro F1 score but is not able to outperform the ensemble approach in multilingual data. The robustness provided by the ML technique makes it a better performing system.

6 Conclusion

The paper describes our experimentation over ComMa v 0.2 dataset consisting of Multilingual, Bangla, Hindi, and English data to perform analysis on aggression, communal bias, and gender bias. We have proposed two strategies Boosted Voting Ensemble and IndicBERT fine-tuned in this paper. The Boosting Voting Ensemble outperforms IndicBERT in terms of instance F1 scores that showcase the robustness of our proposed approach as well its capabilities in handling all three labels efficiently. However, it should also be noted that IndicBERT majorly outperforms the Ensemble approach in the individual task, highlighting its power in understanding contextual meanings related to Aggression, Communal Bias, and Gender Bias. The F1 scores for aggression are relatively on the lower side because of the contextual overlaps between the output labels, which was not the case in Gender and Communal Bias. In the future, the inclusion of better embeddings like glove and BERT which capture the underlying semantic and lexical relations could improve the performance of the methodology manifolds. The application of Ensembling techniques in a deep learning setting could be another set of experimentations to be considered.

References

Segun Taofeek Aroyehun and Alexander Gelbukh. 2018. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 90–97.

Tanya Beran and Qing Li. 2005. Cyber-harassment: A study of a new method for an old behavior. *Journal of educational computing research*, 32(3):265.

Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr Ojha. 2020. Developing a multilingual annotated corpus of misogyny and aggression. *arXiv preprint arXiv:2003.07428*.

Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine

classification and statistical modeling for policy and decision making. *Policy & internet*, 7(2):223–242.

Erik Cambria, Praphul Chandra, Avinash Sharma, and Amir Hussain. 2010. Do not feel the trolls. *ISWC, Shanghai*.

Jonathan Culpeper. 2011. *Impoliteness: Using language to cause offence*, volume 28. Cambridge University Press.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696. Springer.

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.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30.

Simona Frenda, Bilal Ghanem, Manuel Montes-y Gómez, and Paolo Rosso. 2019. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5):4743–4752.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the first workshop on abusive language online*, pages 85–90.

Edel Greevy. 2004. *Automatic text categorisation of racist webpages*. Ph.D. thesis, Dublin City University.

Edel Greevy and Alan F Smeaton. 2004. Classifying racist texts using a support vector machine. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 468–469.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the second workshop on NLP and computational social science*, pages 7–16.

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, Shervin Malmasi, and Marcos Zampieri. 2018a. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11.

Ritesh Kumar, Atul Kr Ojha, Marcos Zampieri, and Shervin Malmasi. 2018b. 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)*.

Ritesh Kumar, Aishwarya N Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018c. Aggression-annotated corpus of hindi-english code-mixed data. *arXiv preprint arXiv:1803.09402*.

Srijan Kumar, Francesca Spezzano, and VS Subrahmanian. 2014. Accurately detecting trolls in slashdot zoo via decluttering. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, pages 188–195. IEEE.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence*.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting hate speech in social media. *arXiv preprint arXiv:1712.06427*.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202.

Todor Mihaylov, Georgi Georgiev, and Preslav Nakov. 2015. Finding opinion manipulation trolls in news community forums. In *Proceedings of the nineteenth conference on computational natural language learning*, pages 310–314.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A bert-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, pages 928–940. Springer.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on arabic social media. In *Proceedings of the first workshop on abusive language online*, pages 52–56.

Alex Nikolov and Victor Radivchev. 2019. Nikolov-radivchev at semeval-2019 task 6: Offensive tweet classification with bert and ensembles. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 691–695.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

J Nockleyby. 2000. 'hate speech in encyclopedia of the american constitution. *Electronic Journal of Academic and Special librarianship*.

Constantin Orasan. 2018. Aggressive language identification using word embeddings and sentiment features. Association for Computational Linguistics.

Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2019. Multi-label categorization of accounts of sexism using a neural framework. *arXiv preprint arXiv:1910.04602*.

Kashyap Raiyani, Teresa Gonçalves, Paulo Quaresma, and Vitor Beires Nogueira. 2018. Fully connected neural network with advance preprocessor to identify aggression over facebook and twitter. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pages 28–41.

Julian Risch, Anke Stoll, Marc Ziegele, and Ralf Krestel. 2019. hpidedis at germeval 2019: Offensive language identification using a german bert model. In *KONVENS*.

Sarah T Roberts, Joel Tetreault, Vinodkumar Prabhakaran, and Zeerak Waseem. 2019. Proceedings of the third workshop on abusive language online. In *Proceedings of the Third Workshop on Abusive Language Online*.

Niloofar Safi Samghabadi, Deepthi Mave, Sudipta Kar, and Thamar Solorio. 2018. Ritual-uh at trac 2018 shared task: aggression identification. *arXiv preprint arXiv:1807.11712*.

Sima Sharifirad and Stan Matwin. 2019. When a tweet is actually sexist. a more comprehensive classification of different online harassment categories and the challenges in nlp. *arXiv preprint arXiv:1902.10584*.

Luis Gerardo Mojica de la Vega and Vincent Ng. 2018. Modeling trolling in social media conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 656–666.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.

sdutta at ComMA@ICON: A CNN-LSTM Model For Hate Detection

Sandip Dutta[†], **Utso Majumder**[†], **Sudip Kumar Naskar**[‡]
[†]Department of ETCE, [‡]Department of CSE
Jadavpur University
Kolkata, India
sandip28dutta@gmail.com, utso1201@gmail.com, sudip.naskar@gmail.com

Abstract

In today's world, online activity and social media are facing an upsurge of cases of aggression, gender-biased comments and communal hate. In this shared task, we used a CNN-LSTM hybrid method to detect aggression, misogynistic and communally charged content in social media texts. First, we employ text cleaning and convert the text into word embeddings. Next we proceed to our CNN-LSTM based model to predict the nature of the text. Our model achieves 0.288, 0.279, 0.294 and 0.335 Overall Micro F1 Scores in multilingual, Meitei, Bengali and Hindi datasets, respectively, on the 3 prediction labels.

1 Introduction

Identifying aggressive and abusive atrocities on the internet is an important field of study in today's world. Researchers are striving to develop remedial measures to combat such online content.

In order to efficiently carry out these tasks, the research community have proposed several Machine Learning models, to enhance the efficiency of handling large sets of data and accurately assessing them. The extent of accuracy, however, is a point of concern, since ML models are entirely dependent on large, comprehensive training datasets. Models are prone to poor performance due to lack of properly curated datasets. Conventional models and ensembles are more reliable in these cases, as their data is easily interpreted.

The work is designed to identify objectionable and abusive content on online platforms, as either aggressive, gender based or communally charged. The objective of the model is to demarcate the overlapping aspects of the three types of contents being investigated, and also if this intersectionality could be useful to the task. The task includes multilingual datasets to widen the spectrum of potentially abusive content and to challenge the models.

2 Related Works

Important research contributions have been made in the domain of aggression detection in text (Razavi et al., 2010; Kumar et al., 2018, 2020) and offensive language (Nobata et al., 2016). Gender bias and communally charged content detection have been investigated in research work such as Anzovino et al. (2018), Kiritchenko and Mohammad (2018) and Davidson et al. (2017) respectively. Aforementioned works are different in terms of the target subject they investigate. The NLP research fraternity has analysed the pragmatic and structural features of such forms of hate speech (Djuric et al., 2015; Dadvar et al., 2013) and developing systems that could automatically detect and handle these (Waseem et al., 2017; Zampieri et al., 2019).

Although the most prevalent language for predicting model datasets is English, there are some other languages on which works have been reported, for example, in Hindi (Mandla et al., 2021).

However, on a general note, any predictive model built on historical data may inadvertently inherit human biases based on gender or ethnicity (Sweeney, 2013; Datta et al., 2015; Sun et al., 2019).

3 Model Description

The prediction pipeline is described in Figure (1). The task required us to detect aggression, misogyny and communal hatred in text data in multiple languages. Additional challenge was introduced by code mixing and code switching.

We use a CNN-LSTM based neural network for our prediction task. The steps undertaken are presented here.

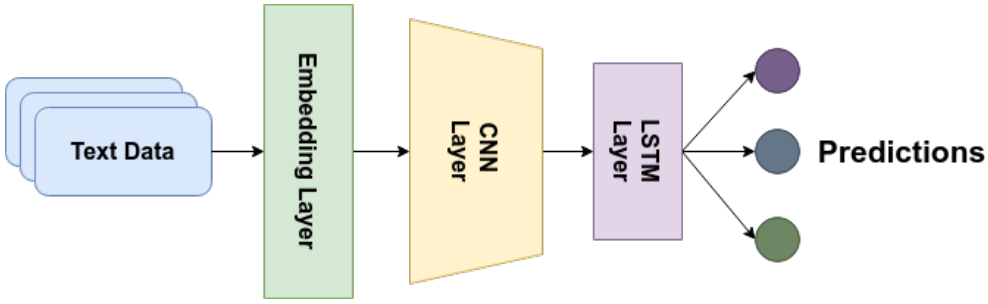
3.1 Text Data Cleaning

The data was cleaned using the following steps:

- **Hashtag, User Handle and URL Removal**: Hashtags and user handles provide redundant

Proceedings of the 18th International Conference on Natural Language Processing:
Shared Task on Multilingual Gender Biased and Communal Language Identification, pages 53–57
National Institute of Technology - Silchar, December 16-19, 2021.

Figure 1: Model Diagram



information and these were removed using regular expressions.

- **Punctuation Removal**: Punctuation introduces noise in the text and inflates the vocabulary size. It was also cleaned using regular expressions.

3.2 Word Embedding Vectorization

The word embedding layer converts the sentences into dense word vectors (Mikolov et al., 2013). These provide valuable information to subsequent layers regarding the words.

3.3 CNN - LSTM Model

The combination of CNN and RNN based models (Wang et al., 2016) provides certain advantages. The CNN layer captures global information while LSTM takes care of sequential information.

The CNN layer specializes in identifying informative features from text. The LSTM layer is designed to capture subtle patterns and regularities in sequences. They allow modeling non-markovian dependencies looking at the context window around a focus word, while zooming-in on informative sequential patterns in that window (Goldberg, 2017).

4 Experiments and Results

4.1 Dataset

A multilingual dataset with a total of 12000 samples for training and development and an overall 3000 samples for testing in four Indian languages Meitei, Bangla (Indian variety), Hindi and English, were provided for the task. Each language data was divided into train, validation and test sets. Each data point contains text that is code-mixed with English or their respective varieties of English (i.e. English used in the context of these languages) (Kumar et al., 2021b).

For the task (Kumar et al., 2021a), the contents are categorized broadly into three levels, namely aggression, gender bias and communal bias. The dataset, for each level, is marked at different specific labels or classifications:

- **Level A - Aggression** : This level gives a 3-way classification in between 'Overtly Aggressive'(OAG), 'Covertly Aggressive'(CAG) and 'Non-aggressive'(NAG) text data.

- **Level B - Gender Bias** : At this level the classifier will need to classify the text as 'gendered'(GEN) or 'non-gendered'(NGEN).

- **Level C - Communal Bias** : At the level C, the task is to develop a binary classifier for classifying the text as 'communal' (COM) and 'non-communal'(NCOM).

The task could be approached as three separate classification tasks or a multi-label classification task or a structured classification task. The final submission file contains the labels for each of the three levels as one single predicted tuple.

4.2 Experimental Setup

Figure (1) shows our entire classification model. We create our entire model using Tensorflow (Abadi et al., 2015) and Keras (Chollet et al., 2015). The train, validation and test data was used as is given in (Kumar et al., 2021b).

The random number seed was set to 2833. We selected the maximum sequence length to be of 256 tokens. A vocabulary size of 85000 words was chosen per language for the classification task.

The word embedding dimension was taken to be 50. The Convolution layer gave a 64 dimensional output which was then fed to LSTM layer with `units` hyperparameter set to 100. This output was further fed into the final prediction layer.

Table 1: Predictions by Our Model

Text	Aggression		Misogyny		Communal	
	Actual	Predicted	Actual	Predicted	Actual	Predicted
Chi Chi.A Abar MP. Banglar Lajja	CAG	NAG	GEN	GEN	NCOM	COM
Are kyo apni izzat nilam kar rhi ho	OAG	NAG	GEN	GEN	NCOM	COM
Sunila ekai khangdabi nmaidud khupak thaninge	OAG	CAG	GEN	GEN	NCOM	COM
Aur ye bumbedkar waale bhi bahut madarchod hai	OAG	OAG	GEN	GEN	NCOM	COM

Table 2: Model Scores on Task

Language	Instance F1	Overall Micro F1	Agg. Micro F1	Gen. Micro	Comm. Micro
Multilingual	0.02	0.288	0.376	0.281	0.208
Meitei	0.007	0.279	0.388	0.311	0.138
Bangla	0.006	0.294	0.438	0.339	0.107
Hindi	0.047	0.335	0.44	0.204	0.361

Table 3: Model Scores Comparison on Task

Team Name	Instance F1 Scores			
	Multilingual	Meitei	Bengali	Hindi
Team_BUDDI	0.371	-	-	0.398
Hypers	0.322	0.129	0.223	0.336
Beware Haters	0.294	0.322	0.292	0.289
sdutta	0.02	0.007	0.006	0.047
MUCIC	0.000	0.000	0.000	0.000

We chose Cross Entropy as the loss function for all the 3 prediciton tasks. All other hyperparameters were kept to their default values as is defined in (Chollet et al., 2015).

We trained the model for 12 epochs on a Intel Xeon CPU with Early Stopping enabled. The code[1] was run in the Google Colab environment.

The scores obtained are shown in Table (2).

4.3 Error Analysis

Our model underperforms severely and seems to overfit on certain categories. Some predicitons are shown in Table (1). As is summarized in Table (3), our model provides suboptimal performance in the task compared to other models.

The aggression predictions seem somewhat better than other classes. However, for all the tasks, the performance is not satisfactory.

The main reason for this problem is the huge imbalance in the dataset. The number of data points in one class hugely surpasses other classes. This tends to make the model predict the majority class only. Even enabling early stopping to prevent overfitting gave a poor result due to the high imbalance in this model.

We identified some issues to be cautious of while training on this dataset which are listed below.

- The data is highly imbalanced which can cause severe overfitting. The model will predict only the majority class, which will result in good scores on the train data, but in practice, it will not be beneficial. One can change the loss function to weigh each sample differently during loss calculations. Moreover, a totally different loss function can be used to handle this imbalance.

- There is a lot of code mixing and code switching in this dataset. Code mixing and code switching can inflate the vocabulary size, as there will be multiple representations of the same word. A lot of the texts also contain unicode characters. This further aggravates the problem and can limit the performance of

[1] https://github.com/Dutta-SD/CoMMA_ICON

models in learning good representations of the data. Unicode normalisation can alleviate this problem partially.

These problems severely limit the performance of the model in this dataset. One needs to be aware of these pitfalls before training models.

5 Conclusion

Our model performs moderately on the aggression labels. However, in gender-bias and communally charged labels, it significantly under-performs. Out of the four datasets, the model performs the best on Hindi dataset, but accuracy declines in Meitei and Multilingual datasets.

In the future, we aim to re train the model using sample weighting to obtain better results. We also aim to train using larger models to obtain better results.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer.

Francois Chollet et al. 2015. Keras.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696. Springer.

Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. pages 29–30.

Yoav Goldberg. 2017. Neural network methods for natural language processing. *Synthesis lectures on human language technologies*, 10(1):1–309.

Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems.

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, 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, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating aggression identification in social media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 1–5.

Thomas Mandla, Sandip Modha, Gautam Kishore Shahi, Amit Kumar Jaiswal, Durgesh Nandini, Daksh Patel, Prasenjit Majumder, and Johannes Schäfer. 2021. Overview of the hasoc track at fire 2020: Hate speech and offensive content identification in indo-european languages.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Amir H Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Canadian Conference on Artificial Intelligence*, pages 16–27. Springer.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review.

Latanya Sweeney. 2013. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54.

Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. 2016. Cnn-rnn: A unified framework for multi-label image classification.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media.

MUCIC at ComMA@ICON: Multilingual Gender Biased and Communal Language Identification using n-grams and Multilingual Sentence Encoders

F. Balouchzahi[1,a], **O. Vitman**[1,b], **H.L. Shashirekha**[2,c], **G. Sidorov**[1,d], **A. Gelbukh**[1,e]
[1]Instituto Politécnico Nacional, Centro de Investigación en Computación, CDMX, Mexico
[2]Department of Computer Science, Mangalore University, Mangalore, India
[a]frs_b@yahoo.com, [c]hlsrekha@gmail.com,
{[b]ovitman2021, [d]sidorov, [e]gelbukh}@cic.ipn.mx

Abstract

Social media analytics are widely being explored by researchers for various applications. Prominent among them are identifying and blocking abusive contents especially targeting individuals and communities, for various reasons. The increasing abusive contents and the increasing number of users on social media demands automated tools to detect and filter the abusive contents as it is highly impossible to handle this manually. To address the challenges of detecting abusive contents, this paper describes the approaches proposed by our team MUCIC for Multilingual Gender Biased and Communal Language Identification shared task (ComMA@ICON) at International Conference on Natural Language Processing (ICON) 2021. This shared task dataset consists of code-mixed multi-script texts in Meitei, Bangla, Hindi as well as in Multilingual (a combination of Meitei, Bangla, Hindi, and English). The shared task is modeled as a multi-label Text Classification (TC) task combining word and char n-grams with vectors obtained from Multilingual Sentence Encoder (MSE) to train the Machine Learning (ML) classifiers using Pre-aggregation and Post-aggregation of labels. These approaches obtained the highest performance in the shared task for Meitei, Bangla, and Multilingual texts with instance-F1 scores of 0.350, 0.412, and 0.380 respectively using Pre-aggregation of labels.

1 Introduction

In the past few years, the spread of internet is gradually increasing the user-generated content over various platforms. Consequently, aggressive and hateful content like trolling, cyberbullying, flaming, abusive language, etc. is also growing alarmingly Butt et al. (2021). These abusive contents targeting individuals and communities for various reasons is creating negative impact on individuals as well as on the society Fazlourrahman et al. (2021c). Detection of such abusive contents on social media is a crucial task. Filtering these contents manually is almost an impossible task due to the increasing number of social media users as well as increasing abusive contents. This demands an automated abusive content detection system that aims to reduce the abusive contents and discourage users from demonstrations of any form of aggression. Recently, several shared tasks such as Sexism Identification in Social Networks Rodríguez-Sánchez et al. (2021), Arabic Misogyny Identification Mulki and Ghanem (2021), etc. have explored the detection of abusive contents in different languages.

To tackle the challenges of detecting the abusive contents on social media, in this paper, we team MUCIC, present two ML approaches proposed for ComMA@ICON shared task at ICON 2021 Kumar et al. (2021a). The shared task is defined as a three-level (Level A, B and C) multi-label TC task for code-mixed multi-script texts in three languages: Meitei, Bangla, Hindi as well as in Multilingual (a combination of Meitei, Bangla, Hindi, and English). While Level A is a multi-class classification task with three categories, Level B and C are binary classifications. The shared task could be approached as three separate classification tasks or a multi-label classification task or a structured classification task. However, the final submission file must contain the labels for each of the three levels as one single predicted tuple.

The shared task is modeled as a multi-label TC task combining word and char n-grams with vectors obtained from MSE to train three ML classifiers using Pre-aggregation and Post-aggregation of labels. ML classifiers, namely: Logistic Regression (LR), Support Vector Machine (SVM) and Random Forest (RF) are ensembled as a soft voting classifier. The results released by shared task organizers show that the proposed approaches obtained

Proceedings of the 18th International Conference on Natural Language Processing:
Shared Task on Multilingual Gender Biased and Communal Language Identification, pages 58–63
National Institute of Technology - Silchar, December 16-19, 2021.

highest performance for Meitei, Bangla, and Multilingual texts using Pre-aggregation of labels.
The rest of the paper is organized as follows: Section 2 throws some light on some of the recent works in detecting abusive contents in general, followed by the proposed methodology to detect gender biased and communal language identification in Section 3. Experiments and results are brought out in Section 4 and the paper finally concludes in Section 5.

2 Related Work

Hateful content detection is a very challenging task. In the last few years, there have been several studies proposing several methods for the classification of offensive and hateful speech Waseem et al. (2017); Hardaker (2013); Dadvar et al. (2013); Davidson et al. (2017). Few researchers have also shown that taking context dependencies into account can improve hateful speech detection system considerably Dadvar et al. (2013); Zhang et al. (2018); Dinakar et al. (2012).

Several studies have looked into different types of abusive languages like hate speech, cyberbullying, and trolling. Waseem et al. (2017) suggested classification of abusive language as a two-fold topology that considers whether (i) abusive content is either directed towards a specific individual or a general one and (ii) abusive language is explicit (unambiguous in its potential) or implicit (does not immediately apply or denote abuse).
Dadvar et al. (2013) approached cyberbullying detection as a TC task using content-based, cyberbullying-specific and user-based features to train a SVM to classify comments as bullying or non-bullying. This study proves that incorporating user's context such as comments history and characteristics can considerably improve the performance of cyberbullying detection tools.

Zampieri et al. (2019) compiled the Offensive Language Identification (OLI) dataset in English with tweets annotated using a fine-grained three-layer annotation scheme to distinguish whether the language is offensive or not along with its type and target. Among the experiments conducted using SVM, Bidirectional Long Short-Term-Memory (BiLSTM) and Convolutional Neural Network (CNN), CNN models outperformed other models for OLI, its type and target with macro-F1 scores of 0.80, 0.69, and 0.47 respectively.

The Hindi-English code-mixed dataset developed by Kumar et al. (2018) is crawled from the public pages of Facebook and Twitter consisting of the posts about the issues that are expected to be discussed more among the Indians. With approximately 18k tweets and 21k Facebook comments, the dataset was annotated with different levels and types of aggression, such as physical threat, sexual aggression, gender aggression, etc. using the Crowdflower platform.

Nobata et al. (2016) developed a ML based method to detect hate speech in online user comments. Data was sampled from comments posted on Yahoo! Finance and annotated by New Yahoo's in-house trained raters. Experiments were performed by training Vowpal Wabbit's regression model using n-grams, linguistic, syntactic and distributional semantics features as well as different types of embeddings combined with the standard Natural Language Langauge (NLP) features. The models with a combination of all the features achieved best F1-scores of 0.795 and 0.817 for Finance and News data respectively.

In spite of several techniques to detect abusive language in code-mixed script, very few works focus on Indian languages. This provides lot of scope to carry out experimentation on Indian particularly low-resource languages and also multilingual text and script.

3 Methodology

Inspired by Fazlourrahman et al. (2021a,b,d); Fazlourrahman and Shashirekha (2021) in utilizing various types and combinations of n-grams for code-mixed multi-scripts TC tasks, this work transforms word and char n-grams in the range (1, 3) to Term Frequency–Inverse Document Frequency (TF-IDF) vectors and stacks them with vectors extracted from MSE[1]. The stacked vectors are then used to train the ML classifiers. Range of word and char n-grams and the vector size of all the features for all the languages of the shared task are given in Table 1.

Two approaches used for labels aggregation to train ML classifiers are described below:

- **Pre-aggregation approach:** a single classifier is trained with a tuple of three labels for each sentence as one label. So, the prediction on each test sample consists of one label which in fact is a combination of three labels.

[1]https://tfhub.dev/google/universal-sentence-encoder-multilingual/3

- **Post-aggregation approach:** three individual classifiers are trained with one label each in the tuple of labels and the three predictions on each test set are aggregated (as required by the organizers for the purpose of submitting the predictions for evaluation) as a tuple.

The difference between the two approaches lies in aggregating the labels as shown in Figure 1. While the blue dotted part indicates the model's prediction using Pre-aggregation approach, red dotted part indicates that of Post-aggregation approach. Both the approaches use the same feature engineering step.

Model construction part consists of soft voting ensemble of RF, SVM, and LR classifiers. The classifiers are selected based on their success in Fazlourrahman et al. (2021a,d) for code-mixed multi-script TC tasks.

The classifiers are empowered with hyper-parameter tuning using GridSearchCV module from Sklearn library[2]. A set of random values are assigned for each parameter corresponding to a particular classifier and then GridSearchCV is used to determine the best value for each parameter. However, the limitation of hyper-parameter tuning is that it requires a lot of time to find the best value for each parameter. Owing to the time constraints, hyper-parameter tuning is done only for mulilingual dataset and those parameter values are in turn used for all the datasets. However, hyper-parameter tuning for each dataset separately is expected to enhance the performance of the classifiers. The final values of parameters for each classifier are presented in Table 2.

4 Experiments and Results

4.1 Dataset

The dataset used in this work is provided by the organizers of ComMA@ICON at ICON 2021 shared task Kumar et al. (2021b). It consists of a multi-label TC task in four languages, namely: Meitei, Bangla, Hindi as well as in Multilingual (a combination of Meitei, Bangla, Hindi, and English). The datasets are made up of a combination of native script of intended language and transliterated form as well as English language making the task more challenging. Further, the dataset is designed for the multi-label TC task at three levels as given below:

- **Level A:** a multi-class classifier defined as Aggression Identification to categorize texts into one of the three classes, namely: Overtly Aggressive (OAG), Covertly Aggressive (CAG) and Non-aggressive (NAG)

- **Level B:** a binary classifier defined as Gender Bias Identification task to classify text as either gendered (GEN) or non-gendered (NGEN)

- **Level C:** a binary classifier defined as Communal Bias Identification task to classify text as either communal (COM) or non-communal (NCOM).

Participants were provided with the labeled training and development sets and unlabeled test sets. The statistics of the training sets are given in Table 3. For evaluating the models, 1,002, 962, 1,020, and 2,989 unlabeled texts in Meitei, Bangla, Hindi as well as in Multilingual respectively were provided as test sets. The details of the dataset are given in task website[3].

4.2 Results

The predictions on the test sets are evaluated using two major metrics, namely: instance-F1 and micro-F1. Based on instance-F1 score, all labels in the predicted tuple should be the same as gold labels and the weighted average score of each label will be considered for micro-F1.

The results obtained with both Pre-aggregation and Post-aggregation approaches are presented in Table 4. It can be observed that the models obtained zero instance-F1 for all the four languages using Post-aggregation approach. On contrary to this, using Pre-aggregation approach, the models obtained very high results and the best performance among all the participants. Comparison of the performances in terms of instance-F1 of our models with that of the other participants is presented in Table 5. The results reveal that Pre-aggregation approach achieved highest instance-F1 in the shared task for Meitei, Bangla, and Multilingual texts with instance-F1 scores of 0.350, 0.412, and 0.380 respectively. On the other hand, Post-aggregation approach was more successful in obtaining highest overall micro-F1 scores of 0.723 and 0.690 for Bangla and Meitei texts respectively.

[2]https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[3]https://sites.google.com/view/comma-at-icon2021/overview

Dataset	Feature	Range	Size
Multilingual	Char n-grams	(1, 3)	54,135
	Word n-grams	(1, 3)	271,545
	Multilingual Sentence Encoder	-	512
Meitei	Char n-grams	(1, 3)	12,088
	Word n-grams	(1, 3)	74,404
	Multilingual Sentence Encoder	-	512
Bangla	Char n-grams	(1, 3)	20,810
	Word n-grams	(1, 3)	42,423
	Multilingual Sentence Encoder	-	512
Hindi	Char n-grams	(1, 3)	38,614
	Word n-grams	(1, 3)	160,469
	Multilingual Sentence Encoder	-	512

Table 1: Range and size of features

Classifier	Parameters
RF	max_features='sqrt', n_estimators=1000
SVM	C=100, degree=1, gamma=0.1, kernel='rbf', probability=True
LR	C=10, penalty='l2', solver='liblinear'

Table 2: Parameters and their values for the classifiers

Language	Level A			Level B		Level C	
	OAG	NAG	CAG	GEN	NGEN	COM	NCOM
Hindi	2,526	1,289	800	3,665	950	3,598	1,017
Bangla	1,274	782	335	1,489	902	2,087	304
Meitei	1,024	888	297	2,061	148	2,035	174
Multilingual	4,096	2,959	2,159	7,215	1,999	7,720	1,494

Table 3: Statistics of the training set

Language	Approach	instance-F1	Overall micro-F1	Aggression micro-F1	Gender Bias micro-F1	Communal Bias micro-F1
Hindi	**Post-agg**	0	0.697	0.606	0.801	0.683
	Pre-agg	0.341	0.706	0.620	0.808	0.690
Bangla	**Post-agg**	0	0.723	0.509	0.772	0.890
	Pre-agg	0.412	0.718	0.517	0.746	0.890
Meitei	**Post-agg**	0	0.690	0.484	0.716	0.871
	Pre-agg	0.350	0.681	0.462	0.713	0.868
Multilingual	**Post-agg**	0	0.701	0.534	0.764	0.806
	Pre-agg	0.380	0.705	0.540	0.759	0.816

Table 4: Performance of the proposed approaches (Pre-agg: Pre-aggregation, Post-agg: Post-aggregation)

Language	Metric	Pre-agg	Post-agg	Team_BUDDI	Hypers	Beware Haters	MUM	BFCAI
Hindi	**instance-F1**	0.341	0	**0.398**	0.336	0.289	0.343	0.304
	Overall micro-F1	0.706	0.697	**0.709**	0.683	0.668	0.691	0.678
Bangla	**instance-F1**	**0.412**	0	-	0.223	0.292	0.390	0.391
	Overall micro-F1	0.718	**0.723**	-	0.579	0.704	0.708	0.695
Meitei	**instance-F1**	**0.350**	0	-	0.129	0.322	0.326	0.317
	Overall micro-F1	0.681	**0.690**	-	0.472	0.672	0.661	0.664
Multilingual	**instance-F1**	**0.380**	0	0.371	0.322	0.294	0.359	0.342
	Overall micro-F1	0.705	0.701	**0.713**	0.685	0.658	0.691	0.671

Table 5: Comparison of the performances of the proposed methodology (Pre-agg and Post-agg) with the top performing teams in the shared task

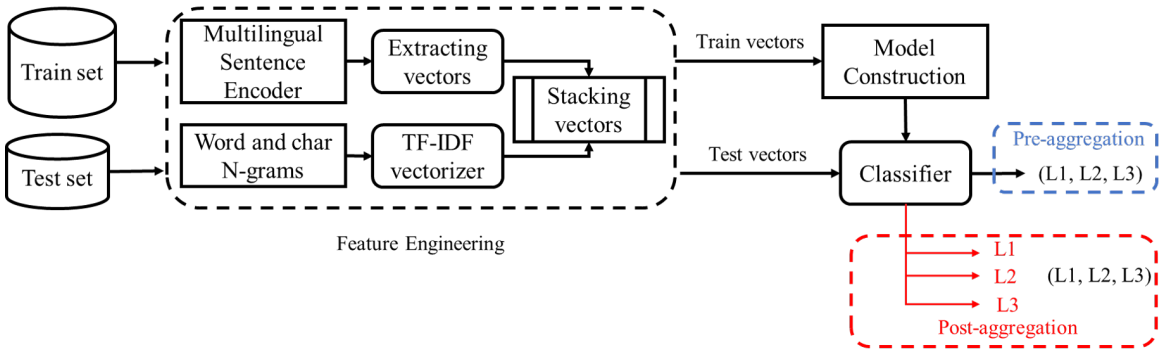


Figure 1: Overview of the proposed methodology

The advantages of proposed approaches over baselines Kumar et al. (2021b) that use combination of word and char n-grams are (i) hyper-parameter tuning using GridSearchCV, and (ii) ensembling ML classifiers as voting classifier to make a robust classifier for TC.

MSE is used as English language is a major component in any code-mixed texts. However, choosing MSE was not a good choice as it failed to encode the complete dataset efficiently mainly because it does not support Indian languages.

5 Conclusion and Future Work

This paper describes the models submitted by the team MUCIC for ComMA@ICON shared task at ICON 2021 for gender biased and communal language identification. The shared task is a three level multi-label TC task for code-mixed multi-scripts texts in Meitei, Bangla, Hindi as well as in Multilingual. Our previous work on code-mixed multi-scripts TC tasks is extended for this shared task with stacked word and char n-grams combined with MSE vectors as features using Pre-aggregation and Post-aggregation of labels. A soft ensemble of three ML classifiers empowered by hyper-parameter tuning using GridSearchCV are trained with the stacked features for the three level multi-label TC task. The results of the shared task provided by the organizers show the highest results using Pre-aggregation approach for Meitei, Bangla, and Multilingual texts with instance-F1 scores of 0.350, 0.412, and 0.380 respectively. This illustrates the efficiency of the proposed approaches.

Acknowledgments

The work was done with the partial support from the Mexican Government through the grant A1-S-47854 of the CONACYT, Mexico, grants 20211784, 20211884, and 20211178 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

Sabur Butt, Noman Ashraf, Grigori Sidorov, and Alexander F. Gelbukh. 2021. Sexism Identification using BERT and Data Augmentation - EXIST2021. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain, September, 2021*, volume 2943 of *CEUR Workshop Proceedings*, pages 381–389. CEUR-WS.org.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving Cyberbullying Detection with User Context. In *European Conference on Information Retrieval*, pages 693–696. Springer.

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.

Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Commonsense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):1–30.

B Fazlourrahman, B K Aparna, and H L Shashirekha. 2021a. MUCS@DravidianLangTech-EACL2021: COOLI-Code-Mixing Offensive Language Identification. In *Proceedings of the First Workshop on*

Speech and Language Technologies for Dravidian Languages, pages 323–329, Kyiv. Association for Computational Linguistics.

B Fazlourrahman, B K Aparna, and H L Shashirekha. 2021b. MUCS@LT-EDI-EACL2021: CoHope-Hope Speech Detection for Equality, Diversity, and Inclusion in Code-Mixed Texts. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 180–187, Kyiv. Association for Computational Linguistics.

B Fazlourrahman, S Grigori, and H L Shashirekha. 2021c. Arabic Misogyny Identification. In *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, Hyderabad, India, December 13-17, 2021*, CEUR Workshop Proceedings. CEUR-WS.org.

B Fazlourrahman and H L Shashirekha. 2021. LASACo: A Study of Learning Approaches for Sentiments Analysis in Code-Mixing Texts. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 109–118, Kyiv. Association for Computational Linguistics.

B Fazlourrahman, H L Shashirekha, and S Grigori. 2021d. A Comparative Study of Syllable and Char Level N-grams for Dravidian Multi-Script and Code-Mixed Offensive Language Identification. In *International Workshop on Soft Computing and Advances in Intelligent Systems, SC-AIS-2021*, Mexico. Journal of Intelligent and Fuzzy Systems.

Claire Hardaker. 2013. "Uh.... not to be nitpicky, but... the past tense of drag is dragged, not drug.": An Overview of Trolling Strategies. *Journal of Language Aggression and Conflict*, 1(1):58–86.

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, Aishwarya N Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated Corpus of Hindi-English Code-mixed Data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Hala Mulki and Bilal Ghanem. 2021. ArMI at FIRE2021: Overview of the First Shared Task on Arabic Misogyny Identification. In *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation*. CEUR.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021. Overview of exist 2021: Sexism Identification in Social Networks. *Procesamiento del Lenguaje Natural*, 67:195–207.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*, pages 745–760. Springer.

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[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 Pre-processing, 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.

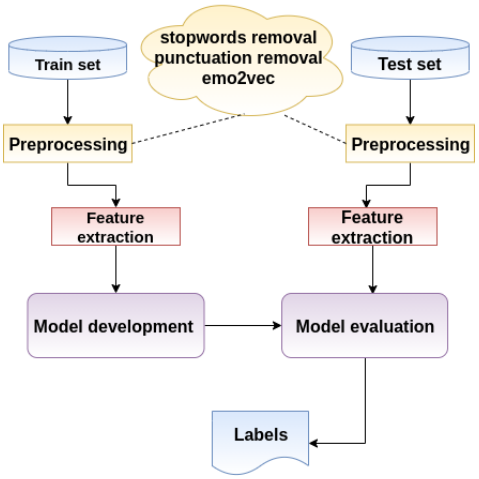


Figure 1: Structure of the ML classifier using Elastic-net

- **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.

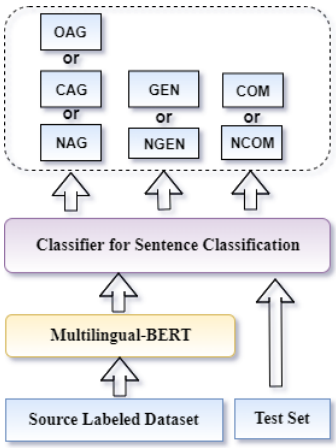


Figure 2: Structure of DL classifier using fine-tuned mBERT

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

Training set					
Task	**Label**	**Meitei**	**Bangla**	**Hindi**	**Multi lingual**
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.

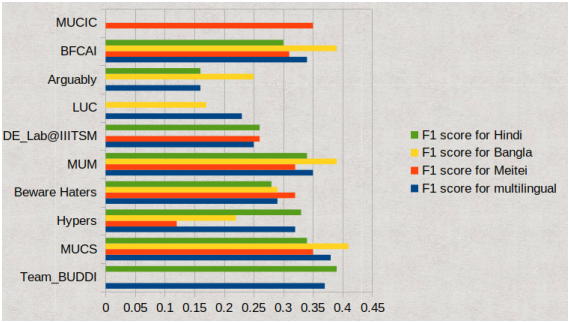


Figure 3: Comparison of instance-F1 scores of the proposed model with top performing models of the shared task

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.

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

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

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.

BFCAI at ComMA@ICON 2021: Support Vector Machines for Multilingual Gender Biased and Communal Language Identification

Fathy Elkazzaz, Fatma Sakr, Rasha Orban, Hamada Nayel
Department of Computer Science
Faculty of Computers and Artificial Intelligence
Benha University, Egypt
{fathy.elkazzaz,fatma.sakr}@fci.bu.edu.eg
{rasha.abdelkreem,hamada.ali}@fci.bu.edu.eg

Abstract

This paper presents the system that has been submitted to the multilingual gender biased and communal language identification shared task by BFCAI team. The proposed model used Support Vector Machines (SVMs) as a classification algorithm. The features have been extracted using TF/IDF model with unigram and bigram. The proposed model is very simple and there are no external resources are needed to build the model.

1 Introduction

The manner in which humans communicate and their communities has been completely changed due to the widespread in Social Network Platforms. The integration of communication and data sharing through platforms like YouTube, Facebook, and Twitter caused the emergence and vocalization of hate between users. The intensity and hostility lying in the written expressions is a matter of grave concern. Articulations of hatefulness often cause breaking down or weakening communities. As the impact of such articulation travels from online to offline domain, resultant reactions frequently lead to incidents like organized riot-like situations and unfortunate casualties to ultimately broaden the scope of marginalization of individuals as well as communities (Bhattacharya et al., 2020). There exists widespread, simmering distrust, hatred and insult towards specific groups of people. Users of social network platforms in most of nations are predisposed both to believe disinformation and to share misinformation about discriminated groups in face-to-face as well as in social network platforms (Kumar et al., 2020).

In recent times, there have been a large number of studies exploring various aspects of hateful and aggressive language and their computational modelling and automatic detection such as toxic comments, trolling, racism, online aggression, cyberbullying, hate speech, abusive language and offensive language (Bhattacharya et al., 2020; Nayel and L, 2019; Nayel, 2020; Nayel et al., 2019; Nayel, 2019).

Prior studies have explored the use of aggressive and hateful language on different platforms such as Twitter and Facebook posts. One of the recent studies was to make use of YouTube comments for computational modelling of aggression and misogyny. Some of the earlier studies on computational modelling of misogyny have focused almost exclusively on tweets (Mubarak et al., 2017; Nayel et al., 2021; Chowdhury et al., 2020). Also, all of these studies have focused on either English or European languages like Greek (Pitenis et al., 2020), Italian (Fersini et al., 2020) and Spanish (Costa-jussà et al., 2020; Chiruzzo et al., 2020). The use of a wide range of aggressive and hateful content on social media becomes interesting as well as challenging to study in context to India which is a secular nation with religious as well as linguistic and cultural heterogeneity (Chowdhury et al., 2020).

The broader aim of research in this area is to understand how communal and sexually threatening misogynistic content is linguistically and structurally constructed. In addition, how this kind of contents evaluated by the other participants in the discourse (Bhattacharya et al., 2020). Data originating from social media is multi-lingual data, which makes the automatic analysis of social media is incredibly challenging. In addition, people of the multi-language countries such as India always use code-mixed contents. To convey challenges of automatic multilingual abusive harmful content detection, we present the model has been submitted to ComMA@ICON shared task at ICON 2021. In this paper, a machine learning based model will be developed to detect the misogyny, gender biased

Proceedings of the 18th International Conference on Natural Language Processing:
Shared Task on Multilingual Gender Biased and Communal Language Identification, pages 70–74
National Institute of Technology - Silchar, December 16-19, 2021.

and communal languages on social media. The system will use a supervised text classification model that would be trained using a dataset annotated at two levels with labels pertaining to sexual and communal aggression.

In this paper, a demonstration of the submitted system by BFCAI team to ComMA@ICON2021 shared task is given. The rest of the paper is organized as follows; related work is outlined in section 2, section 3 presents the task and dataset, methodology and algorithms have been used to develop our system are described in section 4. In section 5, experimental settings and discussion of the results are given. Conclusion and future work are given in the last section.

2 Related Work

A lot of research works have been done in this area. Bolukbasi et al. (2016) provide a way to investigate gender bias observed in well-known word embeddings, e.g., word2vec (Mikolov et al., 2013). They use a set of binary gender pair to gather a gender subspace. For in-explicitly gendered words, the difficulty of the word embeddings that project onto this subspace can be removed to de-bias the embeddings within the gender direction. They furthermore endorse a softer model that balances reconstruction of the precise embeddings at the same time as minimizing part of the embeddings that project onto the gender subspace.

The Word Embedding Association Test (WEAT) was performed by Caliskan et al. (2017). It is entirely predicated on the hypothesis. Also, it phrases embeddings that are closer collectively in the high dimensional area and are semantically nearer. They find strong evidence of social biases in pre-trained phrase embeddings.

Gonen and Goldberg (2019) perform experiments on the use of the de-biasing strategies proposed by Bolukbasi et al. (2016) and Zhao et al. (2018). They explain that bias elimination approaches primarily based on gender routes are inefficient in getting rid of all factors of bias. In an excessive dimensional space, the spatial distribution of the gender-impartial phrase embeddings stay nearly identical after de-biasing. This permits a gender-impartial classifier to nevertheless select the cues that encode different semantic factors of bias. Zhao et al. (2020) create a multilingual European languages dataset for bias evaluation.

Table 1: A glance of shared task and the associated sub-tasks

Subtask	Labels	Description
A	OAG	Content is overtly aggressive
	CAG	Content is covertly aggressive
	NAG	Content is non-aggressive
B	GEN	Content is gendered
	NGEN	Content is non-gendered
C	COM	Content is communal
	NCOM	Content is non-communal

They recommended numerous approaches for quantifying bias from both intrinsic and extrinsic perspectives. Experimental outcomes display that choosing a specific alignment target space and using BERT improve performance. They pick out the embeddings aligned to a gender-wealthy language to lessen the unfairness.

3 Task and Dataset

The aim of Multilingual Gender Biased and Communal Language Identification (ComMA@ICON) shared task is to identify aggressive, gender biased or communally charged contents in text (Kumar et al., 2021a). The shared task encompass Hindi, Meitei, Bangla (Indian variety) and English. Lately, hate speech related research gained a great interest in the area of computational linguistics. The shared task is divided into three sub-tasks (**A**, **B** and **C**) to identify aggressive, gender biased and communal biased contents respectively. Table 1 gives a glance of the shared task and the associated sub-tasks. The corpus is a multilingual dataset consists of 12,000 samples for training and development and an overall 3,000 samples for testing in the proposed languages. Tags contained in this dataset represent the aggression, gender bias and communal bias concepts. The full details of the dataset are given by Kumar et al. (2021b).

4 Methodology

In this section, we present details of the proposed model and the algorithms used.

4.1 Formal Representation

Given a set of comments $C = \{c_1, c_2, ..., c_n\}$, where each comment contains a set of words $c_i = \{w_1, w_2, ..., w_k\}$ and the categories $A =$

$\{OAG, CAG, NAG\}$, $B = \{GEN, NGEN\}$ and $C = \{COM, NCOM\}$ for the sub-tasks A, B and C respectively. The given task is formulated as a multi-label classification problem, where an unlabelled comment is assigned with multiple class labels one from each class **A**, **B** and **C**. The proposed model will assign the triple (a, b, c) such that, $a \in A$, $b \in B$ and $c \in C$ for each given an unlabelled comment.

4.2 Model

The general structure of the presented model is given in Fig. 1. Machine learning algorithms have been used to create the proposed model. As an input for the classification algorithms we extracted Term Frequency/Inverse Document Frequency (TF/IDF) for each instance in the training, development and the blind test set.

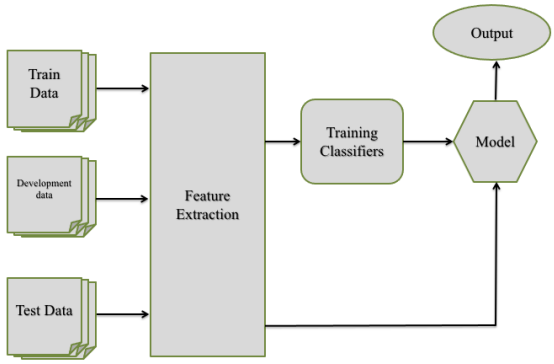


Figure 1: The general framework of presented model.

The model consists of the following stages:

4.2.1 Tokenization

The first step to build any classifier is to represent the input data. In this step, each comment c_k has been tokenized into a set of terms to get n-gram (*unigram* and *bigram*) bag of words.

4.2.2 Feature Extraction

In this phase, a TF/IDF vector has been computed for all the comments in the training and development sets. This vector will be used as an input for developing the classifier. TF/IDF has been calculated as described in (Nayel and Shashirekha, 2017).

4.2.3 Training the Classifier

The features that have been extracted are used as input for training the classifiers. Support Vector Machines (SVMs), Linear classifier, Multilayer Perceptron (MLP) and Multinomial Naive Bayes (MNB) algorithms have been used separately as classification algorithms as well as the ensemble approach to train the model (Nayel and L, 2019). Different classifiers namely, Linear classifier, SVM, MNB, SGD, MLP and Ensemble are created for the given task.

5 Experiments and Results

The following experimental settings have been used while training our classifiers: Stochastic Gradient Descent (SGD) optimization algorithm has been used for optimizing the parameters of linear classifier. "Hinge" function has been used as a loss function for linear classifier and SVM uses the linear kernel while training. The activation function has been used while training MLP was logistic function and there exist 20 neurons in the hidden layer. The hard voting technique was used for ensemble approach. We used `python` programming languages and the library `sklearn`, which contains an integrated set of functions for machine learning framework. The experiments have been conducted on MacBook Air device with 8 GB memory, 1.8 GHz Intel core i5.

5.1 Performance Evaluation

Instance-F1 and micro-F1 are two standard evaluation metrics used for multi-label classification problems. They have been used for evaluation and ranking the submissions of the participants. Instance-F1 is the F-measure averaging on each instance in the test set, while micro-F1 gives a weighted average score of each class and is generally considered a good metric in cases of class-imbalance.

Table 2 shows the instance-F1 and micro-F1 of our submission for all sub-tasks over all language comments. We could submit only SVM output due to time restriction. The performance of our system among all submissions is very interesting, although it is very simple and dependent from any external resources.

Raw data has been used for training the classifiers, we did not apply any preprocessing. This may affect the performance of our model. In addition, we did not use any lexical analysis for the data. In addition, the usage of classical representation for the texts detained the model performance. Using state-of-the-art representation such as word embeddings would improve the model performance.

Table 2: Instance F1 and Micro-F1 for SVM and all languages of the test set

Language	Instance F1	Micro-F1			
		Overall	Aggression	Gender Bias	Communal Bias
Multi	0.340	0.669	0.454	0.765	0.790
Meiti	0.317	0.664	0.438	0.692	0.862
Bangala	0.359	0.665	0.471	0.644	0.882
Hindi	0.304	0.678	0.568	0.799	0.668

6 Conclusion

In this paper, a machine learning approaches have been used for creating a model for detecting the multilingual gender biased and communal contents. Presented model achieved good results compared to its simplicity. Extension of our work includes using deep learning models to develop the classifier and test it on much bigger dataset.

References

Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr. Ojha. 2020. Developing a multilingual annotated corpus of misogyny and aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 158–168, Marseille, France. European Language Resources Association (ELRA).

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Luis Chiruzzo, Santiago Castro, and Aiala Rosá. 2020. HAHA 2019 dataset: A corpus for humor analysis in Spanish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5106–5112, Marseille, France. European Language Resources Association.

Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J. Jansen, and Joni Salminen. 2020. A multi-platform Arabic news comment dataset for offensive language detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6203–6212, Marseille, France. European Language Resources Association.

Marta R. Costa-jussà, Esther González, Asuncion Moreno, and Eudald Cumalat. 2020. Abusive language in Spanish children and young teenager's conversations: data preparation and short text classification with contextual word embeddings. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1533–1537, Marseille, France. European Language Resources Association.

Elisabetta Fersini, Debora Nozza, and Giulia Boifava. 2020. Profiling Italian misogynist: An empirical study. In *Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language*, pages 9–13, Marseille, France. European Language Resources Association (ELRA).

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating aggression identification in social media. In *Proceedings of the*

Second Workshop on Trolling, Aggression and Cyberbullying, pages 1–5, Marseille, France. European Language Resources Association (ELRA).

Tomás Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56, Vancouver, BC, Canada. Association for Computational Linguistics.

Hamada Nayel. 2020. NAYEL at SemEval-2020 task 12: TF/IDF-based approach for automatic offensive language detection in Arabic tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2086–2089, Barcelona (online). International Committee for Computational Linguistics.

Hamada Nayel, Eslam Amer, Aya Allam, and Hanya Abdallah. 2021. Machine learning-based model for sentiment and sarcasm detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 386–389, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Hamada A. Nayel. 2019. NAYEL@APDA: Machine Learning Approach for Author Profiling and Deception Detection in Arabic Texts. In *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019*, volume 2517 of *CEUR Workshop Proceedings*, pages 92–99. CEUR-WS.org.

Hamada A. Nayel and Shashirekha H. L. 2019. DEEP at HASOC2019: A machine learning framework for hate speech and offensive language detection. In *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019*, volume 2517 of *CEUR Workshop Proceedings*, pages 336–343. CEUR-WS.org.

Hamada A. Nayel, Walaa Medhat, and Metwally Rashad. 2019. BENHA@IDAT: Improving Irony Detection in Arabic Tweets using Ensemble Approach. In *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019*, volume 2517 of *CEUR Workshop Proceedings*, pages 401–408. CEUR-WS.org.

Hamada A. Nayel and H. L. Shashirekha. 2017. Mangalore-University@INLI-FIRE-2017: Indian Native Language Identification using Support Vector Machines and Ensemble approach. In *Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation, Bangalore, India, December 8-10, 2017.*, volume 2036 of *CEUR Workshop Proceedings*, pages 106–109. CEUR-WS.org.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in Greek. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5113–5119, Marseille, France. European Language Resources Association.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

Author Index