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:

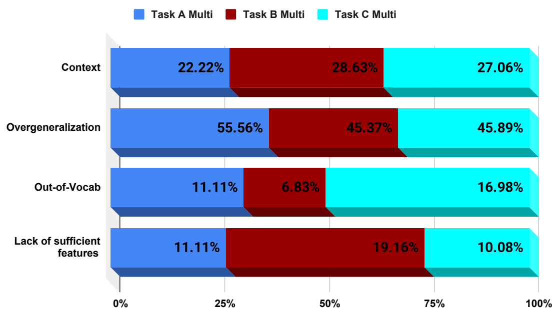

Figure 1: Error types proportion in each sub-task

- **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.