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

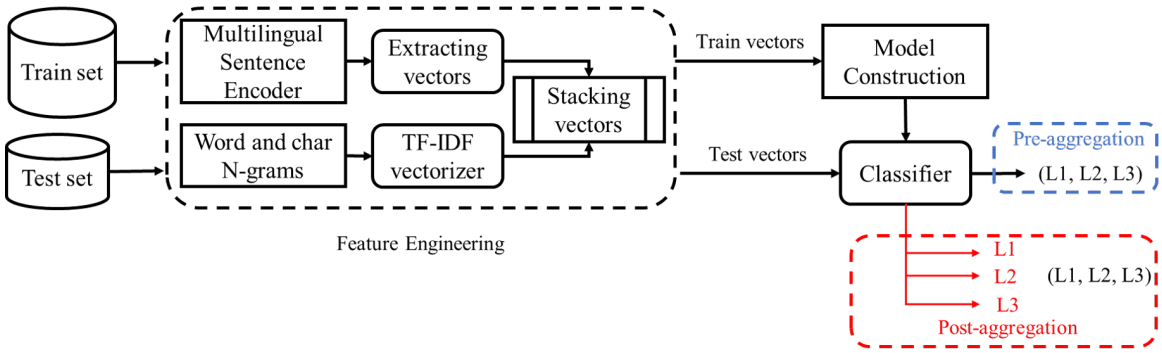

Figure 1: Overview of the proposed methodology

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.