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.