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

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

Abstract

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

1 Introduction

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

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

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

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

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

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

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

2 Related Work

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

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

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

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

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

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

3 Task and Dataset

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

4 Methodology

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

4.1 Formal Representation

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

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

4.2 Model

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

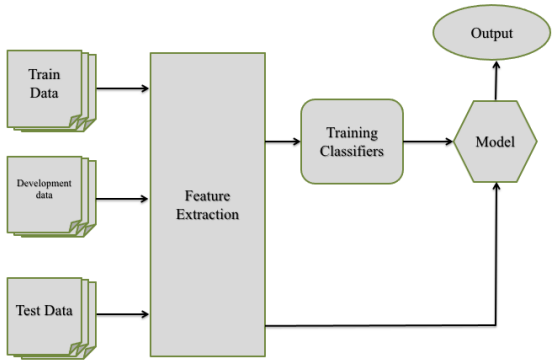

Figure 1: The general framework of presented model.

The model consists of the following stages:

4.2.1 Tokenization

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

4.2.2 Feature Extraction

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

4.2.3 Training the Classifier

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

5 Experiments and Results

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

5.1 Performance Evaluation

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

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

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

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

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

6 Conclusion

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

References

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

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

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

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

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

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

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

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

Ritesh Kumar, Bornini Lahiri, Akanksha Bansal, Enakshi Nandi, Laishram Niranjana Devi, Shyam Ratan, Siddharth Singh, Akash Bhagat, and Yogesh Dawer. 2021a. Comma@icon: Multilingual gender biased and communal language identification task at icon-2021. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON): COMMA@ICON 2021 Shared Task*, Silchar, India. NLP Association of India (NLPAI).

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

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating aggression identification in social media. In *Proceedings of the*

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

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

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

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

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

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

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

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

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

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

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

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