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

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

Abstract

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

1 Introduction

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

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

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

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

have been organized.

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

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

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

2 Related Work

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

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

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

3 Task Description

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

3.1 Sub-Task A

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

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

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

'Non-aggressive'(NAG).

3.2 Sub-Task B

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

3.3 Sub-Task C

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

4 Methodology

4.1 Data Preparation

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

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

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

4.2 Boosted Voting Ensembler

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

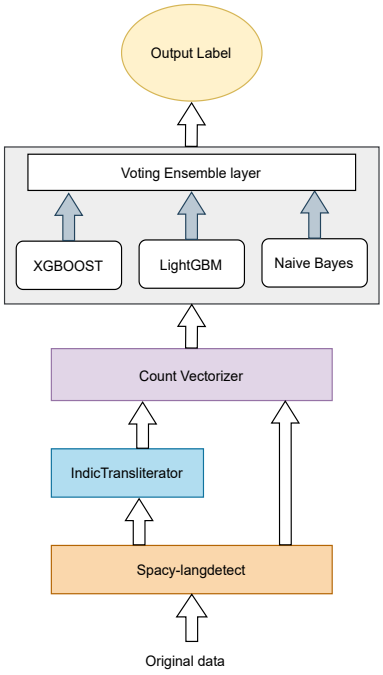

Figure 2: Architecture of Boosted Voting Ensembler

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

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

Table 1: Test Results obtained from Boosted Voting Ensembler approach

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

Table 2: Test Results obtained from IndicBERT approach

4.3 IndicBERT Fine-Tuned

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

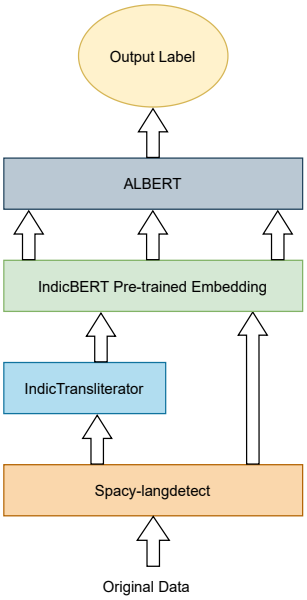

Figure 3: Architecture of Fine-Tuned IndicBERT

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

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

5 Experimentation and Results

5.1 Boosted Voting Ensembler

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

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

5.2 IndicBERT

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

5.3 Comparisons

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

6 Conclusion

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

References

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

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

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

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

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

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

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

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

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

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

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

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

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

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

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

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

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

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

Ritesh Kumar, Atul Kr Ojha, Marcos Zampieri, and Shervin Malmasi. 2018b. Proceedings of the first workshop on trolling, aggression and cyberbullying (trac-2018). In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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