HUB@DravidianLangTech-EACL2021: Identify and Classify Offensive Text in Multilingual Code Mixing in Social Media

Bo Huang
School of Information Science
and Engineering Yunnan University,
Yunnan, P.R. China
hublucashb@gmail.com

Yang Bai
School of Information Science
and Engineering Yunnan University,
Yunnan, P.R. China
baiyang.top@gmail.com

Abstract

This paper introduces the system description of the HUB team participating in DravidianLangTech - EACL2021: Offensive Language Identification in Dravidian Languages. The theme of this shared task is the detection of offensive content in social media. Among the known tasks related to offensive speech detection, this is the first task to detect offensive comments posted in social media comments in the Dravidian language. The task organizer team provided us with the code-mixing task data set mainly composed of three different languages: Malayalam, Kannada, and Tamil. The tasks on the code mixed data in these three different languages can be seen as three different comment/post-level classification tasks. The task on the Malayalam data set is a five-category classification task, and the Kannada and Tamil language data sets are two six-category classification tasks. Based on our analysis of the task description and task data set, we chose to use the multilingual BERT model to complete this task. In this paper, we will discuss our fine-tuning methods, models, experiments, and results.

1 Introduction and Background

Social media platforms are playing an increasingly important role in people's modern social life. Even applications in academic exchanges and technological dissemination are becoming more and more popular (Sugimoto et al., 2017). In recent years, various offensive comments directed at individuals, groups, races, and countries that have appeared in social media have attracted attention in academic and industrial fields (Zampieri et al., 2019; Davidson et al., 2017). Even more worrying is that offensive comments/posts spread very quickly on social media (Mathew et al., 2019).

As far as the current situation is concerned, the COVID-19 virus is spreading and raging on a global scale. The work of Lyu et al. showed us the age distribution of users who used controversial terms on social media during the COVID-19 virus epidemic, with the total share of the 18-24 and 25-34 age groups being 49% (Depoux et al., 2020). Combining the work of Depoux and others, we can realize that what is more terrifying than the speed and harm of the virus is the public panic caused by rumors and hostile comments on social media (Lyu et al., 2020).

What makes us feel encouraged is that similar issues that are currently appearing on social media have been highly valued in the academic and industrial fields (Ahmad and Murad, 2020). However, the current automation technology is mostly applied to some languages with a large number of users (such as English, Spanish user groups, etc.) (Zahiri and Ahmadvand, 2020; Rangel et al., 2020; Pamungkas et al., 2018). The Dravidian languages were first documented in Tamili (Tamil-Brahmi) script engraved on cave walls in Tamil Nadu's Madurai and Tirunelveli districts in the 6nd century BCE. According to the 14th century Sanskrit text Lilatilakam, which is a grammar of Manipravalam, the spoken languages of modern-day Kerala and Tamil Nadu were identical, and they were referred to as "Dramia" (Tamil). Malayalam split from Tamil after 16th century. The earliest known inscriptions in Sanskrit are from the 1st century BCE, such as the Ayodhya Inscription of Dhana and Ghosundi-Hathibada. Sanskrit borrowed many words and grammatical structure from Tamil, Pali and Prakrit. Tamil languages is one of the longest-surviving classical languages in the world which is older than any surviving language in India.

For Tamil, there few works on sentiment analysis (Thavareesan and Mahesan, 2019, 2020a,b). Combined with our analysis of the negative impact of negative posts/comments in social media, we

Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pages 203–209
April 20, 2021 ©2021 Association for Computational Linguistics

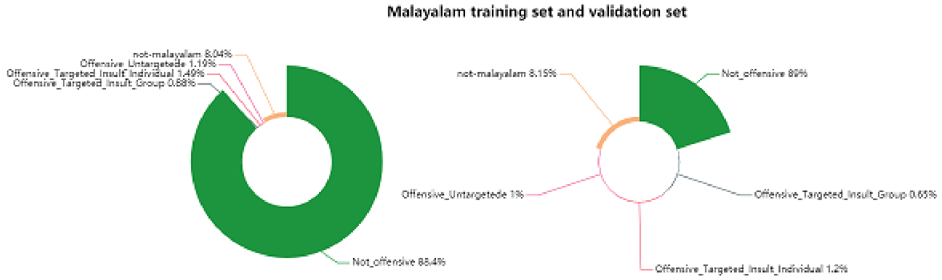

Figure 1: Labels distribution of Malayalam training set and validation set. In the training set, Not offensive: 88.4%, Not Malayalam: 8.04%, Offensive Targeted Insult Individual: 1.49%, Offensive Untargetede: 1.19%, Offensive Targeted Insult Group: 0.88%. In the validation set, Not offensive: 89%, Not Malayalam: 8.15%, Offensive Targeted Insult Individual: 1.2%, Offensive Untargetede: 1%, Offensive Targeted Insult Group: 0.65%.

believe that it is very meaningful and necessary to study the task of Offensive Language Identification in Dravidian Languages-EACL 2021[1] (Chakravarthi et al., 2021). The dataset in the task is the corpus obtained by Chakravarthi and others from social media comments/posts. Use this corpus to create the Dravidian Languages code-mixed dataset for Tamil-English, Malayalam-English, and Kannada-English (Chakravarthi et al., 2020a,b; Hande et al., 2020). Code-mixing is mixing of two or more languages in the conversation (Jose et al., 2020; Priyadharshini et al., 2020; Mandl et al., 2020). What we need to accomplish is to design an automated system and enter a Youtube comment into the system. Then let the system automatically detect which of the categories Not-offensive, Offensive-untargeted, Offensive-targeted-individual, Offensive-targeted-group, Offensive-targeted-other, or Not-in-indented-language the comment content should be classified into. There is usually a lot of noise in the text of social media comments. Also, the data set for this task is text with code mixed types. Combining these characteristics, we chose to use the pre-trained language model multilingual BERT that achieved excellent results in tasks in the natural language processing field to complete this task (Pires et al., 2019). Regarding the problem of code-mixing language, we try to quote Tf-Idf to alleviate the adverse effect of code-mixing on the result. In the next part 3 and part 4, we will introduce our methods and experiments in the task in detail.

2 Related Work

Recently, various issues arising on online social media platforms on the Internet have been receiving attention from many parties. Mossie et al. showed us that they use deep learning methods to process some posts on social media. Their purpose is to predict the target groups that may be subjected to hate attacks. This work provides a very valuable reference for relevant governments and organizations to formulate measures to protect vulnerable groups (Mossie and Wang, 2020). Williams et al. use data science methods to analyze the connection between data from crime, census, and Twitter, revealing to us that hate crime in the digital age is a complete process, not a discrete event (Williams et al., 2020). Velásquez et al. used mathematical analysis and modeling methods to detect and evaluate malicious content related to COVID-19 on online social media. Obtained the critical point of the virus spreading at multiple levels (Velásquez et al., 2020).

Vidgen et al. used machine learning methods to perform a quantitative analysis of whether Islamophobic hate speech in social media may be strong or weak (Vidgen and Yasseri, 2020). The appearance of a large number of fake news on social media has induced hostile comments to a certain extent. To detect fake news in social networks, Zhou et al. proposed a theory-driven fake news detection model. This method investigates fake news from multiple levels, involving lexical semantics, social psychology, and supervised learning method models. The contribution of this work is not only to increase the recognition rate of fake news but also to enhance the interpretability of fake news. What is more worthy of our attention is that the method of Zhou et

[1] https://dravidianlangtech.github.io/2021/

al. can also detect bad news early when the content information is limited (Zhou et al., 2020).

3 Data And Methods

3.1 Data Description and Analysis

The task description shows us that the data used in the task comes from some comments and posts on YouTube. The three data sets include training and validation sets composed of code mixed languages. The three code mixed languages are mainly Tamil, Malayalam, and Kannada. The label distribution probabilities in the training set and validation set of the three different code mixed languages provided by the task organizer are very similar. There are five different types in the Malayalam dataset labels, the Tamil data sets, and Kannada data sets are six different types of labels. Compared with the Malayalam language, an "Offensive Targeted Insult Other" label is added. Also, in the three different Dravidian language data sets, the data volume distribution of different tags is very unbalanced. This unbalanced data label distribution is most prominent on the Malayalam dataset. The proportion of "Not offensive" labels in the Malayalam language data set is 88.4%, and the sum of the other four labels only accounts for 21.6%.

In addition to the above-mentioned feature of label ratio distribution, another feature of the data set of this task is code-mixing. Code mixing means that the text in a piece of text may contain two or more languages. Also, since the data set comes from comments/posts on social media, there are many special symbols and emojis in the text. These data characteristics are all difficult points we need to face.

3.2 Methods

Combining our analysis and understanding of task description and task data set, we choose to develop our system based on multilingual BERT. Because as far as we know, the multilingual corpus used by multilingual BERT in the pre-training stage includes Malayalam, Tamil, and Kannada[2]. The structure of multilingual BERT is the same as that of BERT. The difference is that a corpus with a richer variety of languages than BERT is used in the pre-training phase. These corpora involve more than 100 text corpora of different language types. Therefore, the multilingual BERT has a strong advantage in cross-language. In the previous content, we have analyzed the characteristics of code mixed text, so we try to use the Tf-Idf algorithm to weight the output of multilingual BERT. We hope that using this method can reduce the impact of code-mixing on the results.

In our system, in the first step, we input text data into the multilingual BERT model, and also process the same text data using the Tf-Idf algorithm. In the second step, we take the output of the last layer of(last_layer_output) the multilingual BERT, and then use the Tf-Idf algorithm to get the text encoding and the last_layer_output to do the weighting operation. We can get a shape that is the same as the output of the last_layer_output. We call it the weighted_output. In the third step, we input the last_layer_output and weighted_output into the same CNN block and use two different linear classifiers(Classifier_0, Classifier_1) to classify the results. The last step is to perform arithmetic average operation(Mean) on the results of the two different linear classifiers into the system Finally output the result. We provide the code implementation of our system[3].

4 Experiment and Results

4.1 Data Preprocessing

Regarding this part of the data preprocessing, we mainly use the Tf-Idf algorithm to get weighted_output. To ensure that the shape of the text encoding processed by the Tf-Idf algorithm is the same as the shape of (last_layer_output), we delete the part of the text encoding that exceeds the maximum sentence length, and for the text encoding less than the maximum sentence length, we perform zero-padding operations.

4.2 Experiment setting

The BERT-base-multilingual-cased pre-training language model we use in our system is from the version released by Hugging Face[4]. The CNN block is composed of Conv2d convolution. The convolution kernel uses three different sizes of 3, 4, and 5. Use the maximum pooling operation and the activation function to select ReLU. Finally, the three convolution results of different sizes are

[2] https://github.com/google-research/bert/blob/master/multilingual.md

[3] https://github.com/Hub-Lucas/hub-at-offensive-2021

[4] https://huggingface.co/bert-base-multilingual-cased/tree/main

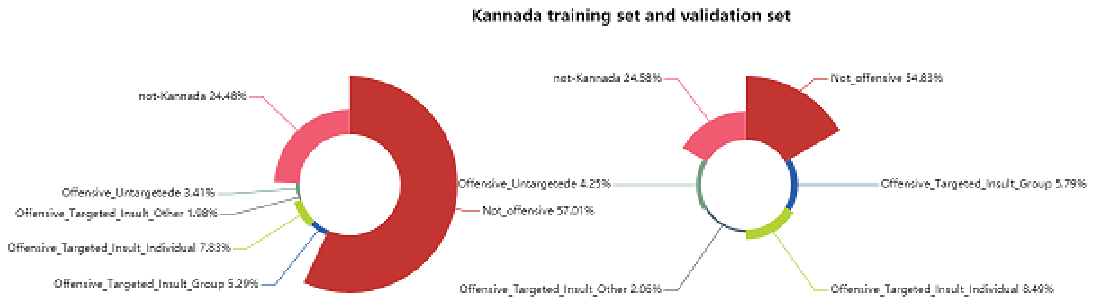

Figure 2: Labels distribution of Kannada training set and validation set. In the training set, Not offensive: 57.01%, Not Kannada: 24.48%, Offensive Targeted Insult Individual: 7.83%, Offensive Targeted Insult Group: 5.29%, Offensive Untargetede: 3.41%, Offensive Targeted Insult Other: 1.98%. In the validation set, Not offensive: 54.83%, Not Kannada: 24.58%, Offensive Targeted Insult Individual: 8.49%, Offensive Targeted Insult Group: 5.79%, Offensive Untargetede: 4.25%, Offensive Targeted Insult Other: 2.06%.

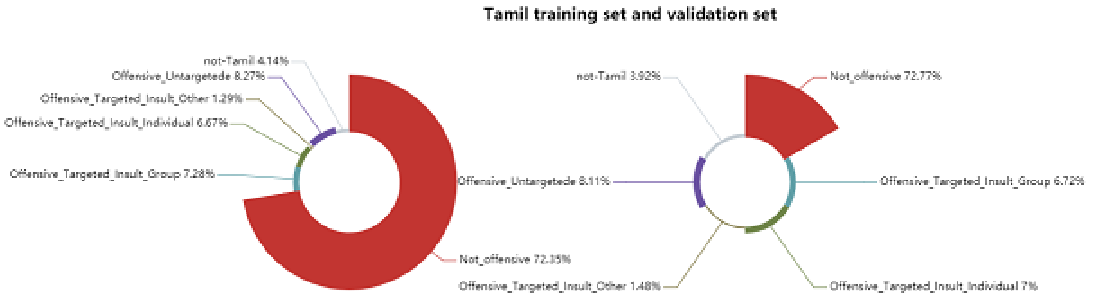

Figure 3: Labels distribution of Tamil training set and validation set.
In the training set, Not offensive: 72.23%, Offensive Untargetede: 8.27%, Offensive Targeted Insult Group: 7.28%, Offensive Targeted Insult Individual: 6.67%, Not-Tamil: 4.14%, Offensive Targeted Insult Other: 1.29%.
In the validation set, Not offensive: 72.77%, Offensive Untargetede: 8.11%, Offensive Targeted Insult Group: 6.72%, Offensive Targeted Insult Individual: 7%, Not Tamil: 3.92%, Offensive Targeted Insult Other: 1.48%.

Language	F1	Precision	Recall
Malayalam	0.91	0.91	0.92
Tamil	0.78	0.78	0.79
Kannada	0.70	0.71	0.73

Table 1: The results of our model and method on the validation set. The validation set data is provided by the task organizer team.

Language	F1	Precision	Recall
Our Malayalam	0.91	0.89	0.93
Our Tamil	0.74	0.73	0.78
Our Kannada	0.64	0.65	0.69

Table 2: The results of our model and method on the test set. The score of the test set comes from the ranking list announced by the task organizer team.

stitched together (256+256+256), the output dimension of the CNN block is 768 dimensions. The loss function of the two different classifiers is the CrossEntropyLoss function provided by PyTorch. In the experiment, we adjust the parameters according to the scores of different language data sets on the validation set.

- **Malayalam data set**: The epoch, batch size, maximum sequence length, and learning rate for the data set are 5, 32, 70, and 3e-5, respectively.
- **Kannada data set**: The epoch, batch size, maximum sequence length, and learning rate for the data set are 4, 32, 80, and 4e-5, respectively.
- **Tamil data set**: The epoch, batch size, maximum sequence length, and learning rate for the data set are 5, 32, 70, and 2e-5, respectively.

4.3 Analysis of Results

The task evaluation index specified by the task organizer team in the task description is the weighted average F1 score. The leaderboard finally announced by the task organizer not only has the ranking of each team, but also the F1 score, Precision, and Recall of the results submitted by each team.

Language	F1	Precision	Recall
Top1 Malayalam	0.97	0.97	0.97
Top1 Tamil	0.78	0.78	0.78
Top1 Kannada	0.73	0.78	0.75

Table 3: The results of the Top1 team in the test set. The score of the test set comes from the ranking list announced by the task organizer team.

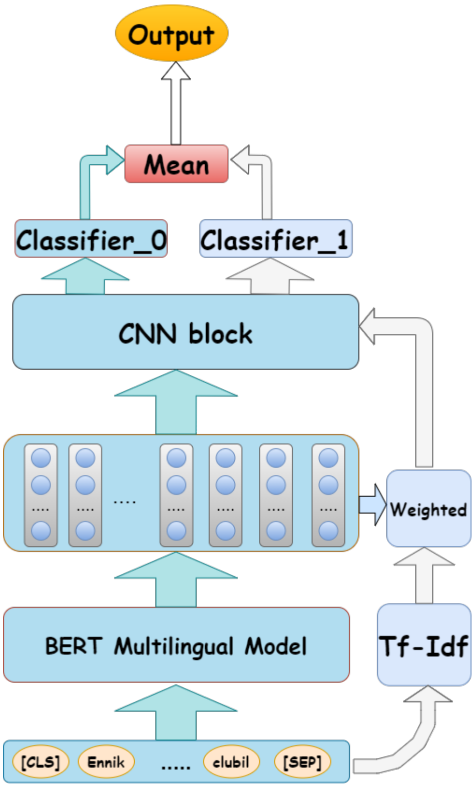

Figure 4: The model structure we used in this task.

Comparing our result score on the Malayalam test set and the result score on the Malayalam data validation set are very close. This shows that our training on the Malayalam language data set is normal. On the Kannada data and Tamil data, there is a large gap between our test set score and the validation set score. This result is likely to be overfitting during the training process. Because the Malayalam data set also has data imbalance, and the result score of the validation set of Malayalam is very close to the result score of the test set. This shows that our result scores on Kannada data and Tamil data are likely to be caused by the model overfitting. By comparing the scores of three different language test data sets, we can see that the prediction results of the two test sets of Tamil and Kannada are compared with the scores submitted by the top1 team, and there is a large gap.

5 Conclusion

On the three different language data sets provided by the task organizer, we combined the Tf-Idf algorithm and the output of the multilingual BERT model and introduced the CNN block as a shared layer. Experimental results prove that our conjecture is feasible, but our method needs to be improved. Especially to eliminate the over-fitting phenomenon in the training phase. When we rechecked our work, we also discovered our omissions. On the Kannada and Tami data sets, the verification set and the test set are different in the setting of the maximum sentence length. At the same time, our model has many areas that can be improved. For example, the use of classifiers and the replacement of CNN blocks are areas that we can try to optimize in future work. Stop words and some special symbols can be deleted in data preprocessing. Try to use data enhancement and other methods on the problem of data imbalance. We also hope that our system and methods can give other teams that pay attention to such tasks some inspiration and reference. In future work, we will not only improve our methods and systems but also continue to pay attention to related code-mixing fields progress.

References

Araz Ramazan Ahmad and Hersh Rasool Murad. 2020. The impact of social media on panic during the covid-19 pandemic in iraqi kurdistan: online questionnaire study. *Journal of Medical Internet Research*, 22(5):e19556.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020a. A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020b. Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210, Marseille, France. European Language Resources association.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan V, Elizabeth Sherly, and John Philip McCrae. 2021. Findings of the shared task on Offensive Language Identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

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.

Anneliese Depoux, Sam Martin, Emilie Karafillakis, Raman Preet, Annelies Wilder-Smith, and Heidi Larson. 2020. The pandemic of social media panic travels faster than the covid-19 outbreak.

Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2020. KanCMD: Kannada CodeMixed dataset for sentiment analysis and offensive language detection. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 54–63, Barcelona, Spain (Online). Association for Computational Linguistics.

Navya Jose, Bharathi Raja Chakravarthi, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2020. A Survey of Current Datasets for Code-Switching Research. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 136–141.

Hanjia Lyu, Long Chen, Yu Wang, and Jiebo Luo. 2020. Sense and sensibility: Characterizing social media users regarding the use of controversial terms for covid-19. *IEEE Transactions on Big Data*.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 29–32, New York, NY, USA. Association for Computing Machinery.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pages 173–182.

Zewdie Mossie and Jenq-Haur Wang. 2020. Vulnerable community identification using hate speech detection on social media. *Information Processing & Management*, 57(3):102087.

Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, Viviana Patti, et al. 2018. 14-exlab@ unito for ami at ibereval2018: Exploiting lexical knowledge for detecting misogyny in english and spanish tweets. In *3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2018*, volume 2150, pages 234–241. CEUR-WS.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Mani Vegupatti, and John P. McCrae. 2020. Named Entity Recognition for Code-Mixed Indian Corpus using Meta Embedding. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 68–72.

Francisco Rangel, Anastasia Giachanou, Bilal Ghanem, and Paolo Rosso. 2020. Overview of the 8th author profiling task at pan 2020: Profiling fake news spreaders on twitter. In *CLEF*.

Cassidy R Sugimoto, Sam Work, Vincent Larivière, and Stefanie Haustein. 2017. Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and technology*, 68(9):2037–2062.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019. Sentiment Analysis in Tamil Texts: A Study on Machine Learning Techniques and Feature Representation. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 320–325.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020a. Sentiment Lexicon Expansion using Word2vec and fastText for Sentiment Prediction in Tamil texts. In *2020 Moratuwa Engineering Research Conference (MERCon)*, pages 272–276.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020b. Word embedding-based Part of Speech tagging in Tamil texts. In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, pages 478–482.

Nicolás Velásquez, R Leahy, N Johnson Restrepo, Yonatan Lupu, R Sear, N Gabriel, Omkant Jha, and NF Johnson. 2020. Hate multiverse spreads malicious covid-19 content online beyond individual platform control. *arXiv preprint arXiv:2004.00673*.

Bertie Vidgen and Taha Yasseri. 2020. Detecting weak and strong islamophobic hate speech on social media. *Journal of Information Technology & Politics*, 17(1):66–78.

Matthew L Williams, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. 2020. Hate in the machine: anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology*, 60(1):93–117.

Sayyed M Zahiri and Ali Ahmadvand. 2020. Crab: Class representation attentive bert for hate speech identification in social media. *arXiv preprint arXiv:2010.13028*.

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. *arXiv preprint arXiv:1902.09666*.

Xinyi Zhou, Atishay Jain, Vir V Phoha, and Reza Zafarani. 2020. Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice*, 1(2):1–25.