nlpBDpatriots at BLP-2023 Task 1: A Two-Step Classification for Violence Inciting Text Detection in Bangla

Md Nishat Raihan*, **Dhiman Goswami***, **Sadiya Sayara Chowdhury Puspo***, **Marcos Zampieri**
George Mason University
{dgoswam|mraihan2|spuspo|mzampier}@gmu.edu

Abstract

In this paper, we discuss the nlpBDpatriots entry to the shared task on Violence Inciting Text Detection (VITD) organized as part of the first workshop on Bangla Language Processing (BLP) co-located with EMNLP. The aim of this task is to identify and classify the violent threats, that provoke further unlawful violent acts. Our best-performing approach for the task is two-step classification using back translation and multilinguality which ranked 6^{th} out of 27 teams with a macro F1 score of 0.74.

1 Introduction

In an era dominated by social media platforms such as Facebook, Instagram, and TikTok, billions of individuals have found themselves connected like never before, enabling them to swiftly share their thoughts and viewpoints. The growth of social networks provides people all over the world with unprecedented levels of connectedness and enriched communication. However, social media posts often abound with comments containing varying degrees of violence, whether expressed overtly or covertly (Kumar et al., 2018, 2020). To combat this worrisome trend, social media platforms established community guidelines and standards that users are expected to adhere to.[1,2].Violations of these rules may result in the removal of offensive content or even the suspension of user accounts. Given the vast amount of user-generated content on these platforms, manually scrutinizing and filtering potential violence is a very challenging task. This moderation approach is limited by moderators' capacity to keep pace, comprehend evolving slang and language nuances, and navigate the complexity of multilingual content (Das et al., 2022). To address this issue, several social media platforms turn to AI and NLP models capable of detecting inappropriate content across a range of categories such as aggression and violence, hate speech, and general offensive language (Zia et al., 2022; Weerasooriya et al., 2023).

The shared task on Violence Inciting Text Detection (VITD) (Saha et al., 2023a) aims to categorize and discern various forms of communal violence, aiming to shed light on mitigating this complex phenomenon for the Bangla speakers. For this task, we carry out various experiments presented in this paper. We employ various models and data augmentation techniques for violent text identification in Bangla.

2 Related Work

Violence Identification in Bangla Several works have been done on building datasets similar to this task and training models on those data. Such datasets include the works of (Remon et al., 2022; Das et al., 2022), which mostly gather data by social media mining. However, most of the datasets are comparatively small in size. One of the larger datasets is prepared by Romim et al. (2022), which consists of 30,000 user comments from YouTube and Facebook, annotated using crowdsourcing.

While most works focus primarily on the datasets, they also present some experimental analysis. Das et al. (2022) evaluates transformer-based models like m-BERT, XLM-RoBERTa, IndicBERT, and MuRIL. XLM-RoBERTa excels with ample training and MuRIL performs well in joint training, while m-BERT and IndicBERT show proficiency in zero-shot scenarios. However, the most notable work here is done by Jahan et al. (2022) who introduces BanglaHateBERT, a re-trained BERT model for abusive language detection in Bangla. It is trained on a large-scale Bangla offensive, abusive, and hateful corpus. The authors collect and annotate a balanced Bangla hate speech dataset and use

*These three authors contributed equally to this work.

[1] https://transparency.fb.com/policies/community-standards/hate-speech

[2] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 179–184

it to pretrain BanglaBERT. The proposed model, BanglaHateBERT, outperforms other BERT models and CNN-based models in detecting hate speech on benchmark datasets.

Related Shared Tasks Zampieri et al. (2019, 2020) organized OffensEval, a series of shared tasks identifying and categorizing offensive language in tweets organized at SemEval 2019 and 2020. At OffensEval, participants trained a variety of models ranging from machine learning to deep learning approaches. While BERT and other transformed dominated the leaderboard in 2020, systems' performance in 2019 was more varied with traditional ML classifiers and ensemble-based approaches achieving competition performance along with deep learning approaches. Another shared task, MEX-A3T track at IberLEF 2019 (Aragon et al., 2019), focused on author profiling and aggressiveness detection in Mexican Spanish tweets. Additionally, Modha et al. (2021) presents an overview of the HASOC track at FIRE 2021 for hate speech and offensive content detection in English, Hindi, and Marathi, where the highest accuracy is achieved on the Marathi dataset.

3 Dataset

The VITD shared task (Saha et al., 2023b) provides the participants with a Bangla dataset including 2700 instances for training and 1330 instances for development. The blind test set contains 2016 instances. The dataset (Saha et al., 2023a) has been annotated using three labels: Non-Violence, Direct-Violence, and Passive-Violence. This three-class annotated dataset differs from similar datasets where a binary annotation is used (Romim et al., 2022; Wadud et al., 2021). The data distribution per label is shown in Table 1.

Label	Train	Dev	Test
Non-Violence	51%	54%	54%
Passive-Violence	34%	31%	36%
Direct-Violence	15%	15%	10%

Table 1: Label-wise data distribution across training, development, and test datasets.

4 Methodologies

4.1 Models

Statistical ML Classifiers In our experiments, we use statistical machine learning models like Logistic Regression and Support Vector Machine using TF-IDF vectors.

Transformers We test multiple transformer models pre-trained on Bangla. Our initial experiments include Bangla-BERT (Kowsher et al., 2022) which is only pre-trained on Bangla corpus. We fine-tune the model on the train set and evaluate it on the dev set with empirical hyperparameter tuning. We then use multilingual transformer models like multilingual-BERT (Devlin et al., 2019) and xlm-roBERTa (Conneau et al., 2020), which are pre-trained on 104 and 100 different languages respectively, including Bangla. We also do the same hyperparameter tuning with both models. Lastly, we use MuRIL (Khanuja et al., 2021), another transformer pre-trained in 17 Indian languages including Bangla.

Task Fine-tuned Models We use BanglaHateBERT (Jahan et al., 2022) as a task fine-tuned model which is developed on existing pre-trained BanglaBERT (Kowsher et al., 2022) model and retrained with 1.5 million offensive posts.

Prompting We prompt gpt-3.5-turbo model (OpenAI, 2023) from OpenAI for this classification task. We use the API to prompt the model, while providing a few examples for each label and ask the model to label the dev and test set.

4.2 Data Augmentation

Given the relatively small size of the VITD dataset, we implement a few data augmentation strategies to expand its size. First, we use Google's Translator API (Google, 2021) to translate the train and dev set to 3 other languages that are very similar to Bangla (Hindi, Urdu, and Tamil). Bangla, Hindi, Urdu belong to Indo-Aryan language branch and Tamil from Dravidian language brach, though, all of these languages have cultural interaction in south-east asian region. The native speakers of these languages live in closer geographic proximity. Moreover, these languages have similar morphosyntactic features. So, translating Bangla text to those languages do not hamper structural and grammatical integrity of the sentences. Therefore, we combine these new synthetic datasets with the original train dataset and finetune the multilingual transformer models on them.

The second approach to augment the dataset is back translation. We again use the Translator API to translate the original train and dev set to a few

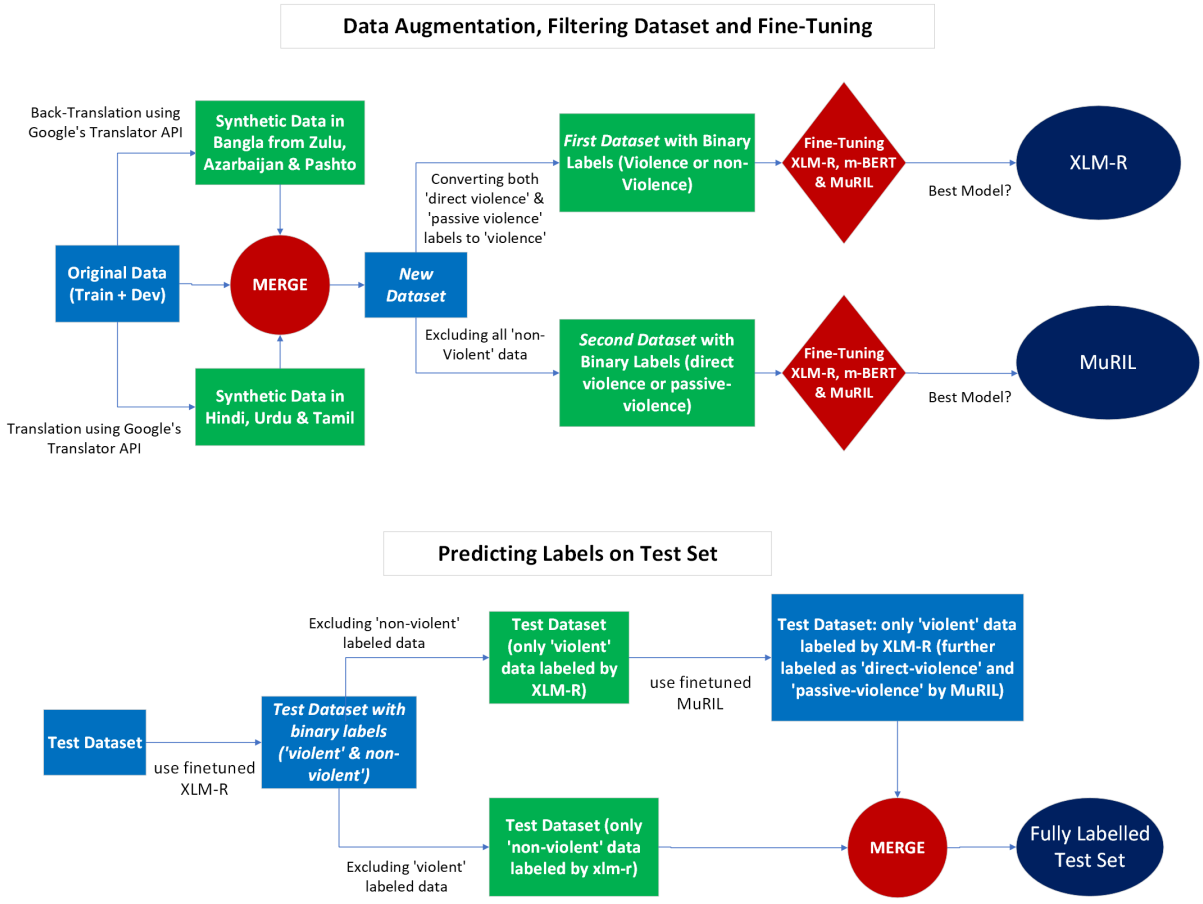

Figure 1: Two-step Classification with Data Augmentation

different low-resource languages like Zulu, Pashto, and Azarbijani as the intermediary language for back translation, in order to add more context. Zulu is from Niger-Congo, Pashto is Indo-Iranian and Azabijani is from Turkic language family. As these languages does not have any cultural interaction with Bangla, back translating from these languages will make three additional version of same sentences with versatility. Then we combine these data with the original dataset. We observe that xlm-roBERTa produces a better macro F1 than the first approach, but still the same as it was on the original data, 0.73.

4.3 Two-step Classification with Data Augmentation

Finally, we combine the two dataset augmentation techniques discussed previously. After combining the synthetic data with the original train set, we have a *New Dataset* that is 7 times the size of the original train set. We generate two different datasets using this *New Dataset*. For the *First Dataset*, we convert all the labels in the *New Dataset* to either Violent (1) or non-Violent (0). And for the *Second Dataset*, we only keep the violent data (both Direct and Passive) from the *New Dataset*.

We finetune mBERT, MuRIL and xlm-roBERTa on both binary labeled *First Datatset* and *Second Dataset* and save their model weights. xlm-roBERTa outperforms the other two when fine-tuned the *First Dataset* and *MuRIL* outperforms the other two when fine-tuned on the *Second Dataset*. For the test set, we first use the finetuned xlm-roBERTa to label the whole dataset as either violent or non-violent data. We then separate all the data from the test set that are labeled as 'violent' by the finetuned xlm-roBERTa model and use the fine-tuned MuRIL model to predict the 'active violence' and 'passive violence' labels. Finally, we merge this with all the 'non-violent' labeled datasets from the first step. Thus, we get all the predicted labels for the test set using 2-step classification by two fine-tuned models. The whole procedure is demonstrated in Figure 1.

5 Results and Analysis

5.1 Results

At the start of the shared task, three baseline macro F1 scores have been provided by the organizers. For BanglaBERT, XLM-R and mBERT, the provided baselines are 0.79, 0.72, and 0.68 respectively. The results of our experiments are shown in Table 2.

Models	Dev	Test
Logistic Regression	0.55	0.56
Support Vector Machine	0.61	0.63
BanglaBERT	0.66	0.67
mBERT	0.71	0.67
MuRIL	0.81	0.72
XLM-R	0.79	0.73
BanglaHateBERT	0.59	0.60
GPT 3.5 Turbo	0.46	0.43
XLM-R (Self-transfer Learning)	0.79	0.72
XLM-R (Multilinguality)	0.78	0.72
XLM-R (Back Translation)	0.77	0.73
XLM-R, MuRIL (Two-step)	**0.84**	**0.74**

Table 2: Dev and test macro F-1 score for all evaluated models and procedures.

Among the statistical machine learning models, we use logistic regression and support vector machine. For logistic regression, we achieve a macro F1 score of 0.56 and for the support vector machine the F1 is 0.63. For transformer-based models, we use BanglaBERT, mBERT, MuRIL and XLM-R where we get the best F1 score of 0.73 by XLM-R. Task fine-tuned model BanglaHateBERT scores 0.60 macro F1.

A few shot learning procedure is used by using GPT3.5 Turbo. We give a few instances of each label as prompt and got 0.43 F1 which is significantly lower than our other attempted approaches. This is because GPT3.5 is still not enough efficient for any downstream classification problem in Bangla like this shared task.

We also perform some customization in our approach instead of directly using the existing models. We use transfer learning. Instead of using the basic idea of transfer learning by fine-tuning a model with a larger dataset of the same label, we translate the train set to English with Google Translator API and used XLM-R on that data. Then we use that finetune model and perform the same procedure over the actual Bangla train set. We refer this procedure as *self-transfer learning* and the F1 score from this procedure is 0.72.

Introducing multilinguality to many downstream tasks proves to be effective. So we also opt for this procedure by translating the train data using Google Translator API to Hindi, Urdu, and Tamil as they are grammatically less diverse and vocabulary is close in contact among the native speakers of these languages. That is how we make the size of our train set three times higher than the original one and got a 0.72 F1 score.

On the other hand, we use Zulu, Azerbaijan, and Pashto - 3 very diverse languages from Bangla for back translation. So, we also get the size of our train set three times higher than the original Bangla one with significantly different translations for each instance. And we get a 0.73 F1 score for that.

Moreover, we use a two-step classification with the data achieved by multilinguality and back translation. Along with these data, we also merge our original Bangla train set. Then, we perform two separate streams of classification. At first, instead of direct and passive violence, we convert them as violence and finetune by XLM-R, mBERT, and MuRIL to classify violence and non-violence where XLM-R performs the best. Then we use the same procedure with the same models to classify direct and passive violence from the merged labels of violence where MuRIL performs the best. Following this procedure, we achieve our best macro F1 score of 0.74 for this shared task.

5.2 Analysis

In terms of text length, the model attains a perfect macro F1 score of 1.000 for texts of 10 words or fewer but struggles with longer texts, evidenced by a macro F1 of only 0.329 for texts of 500-1000 words (Figure 2, Table 3). Though, it maintains respectable F1 scores for text lengths commonly encountered in the dataset, future work should focus on enhancing F1 score for texts with direct violence content.

Text Length	Macro F1	Count	Percentage
(0, 10]	1.000	1	0.050
(10, 20]	0.836	34	1.687
(20, 50]	0.820	528	26.190
(50, 100]	0.736	632	31.349
(100, 200]	0.673	571	28.323
(200, 300]	0.606	156	7.738
(300, 500]	0.627	80	3.968
(500, 1000]	0.329	14	0.694

Table 3: Performance analysis based on text length.

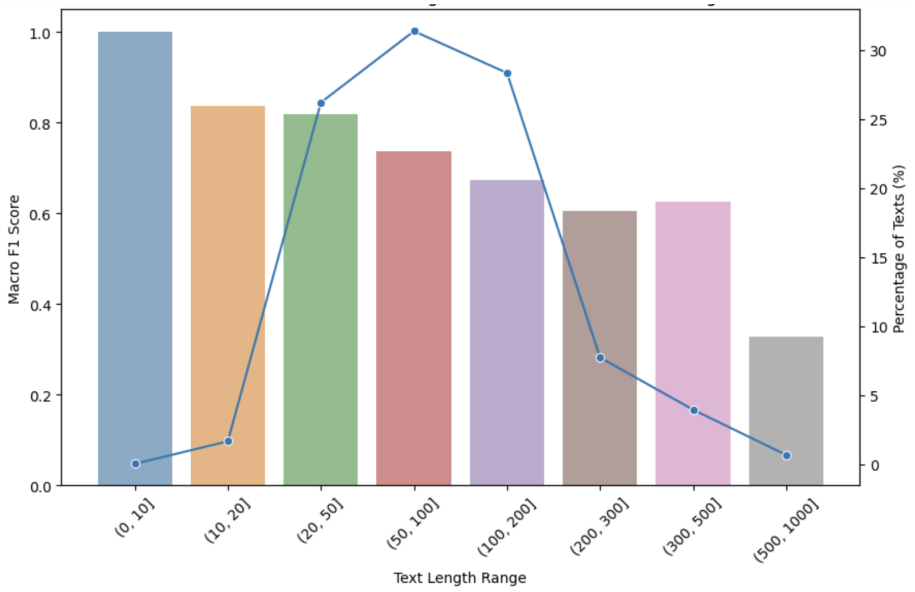

Figure 2: Performance analysis based on text length.

Our model is tasked with categorizing text into one of three labels: non-offensive, direct violence, and passive violence. The confusion matrix, displayed in Figure 3, depicts the performance of the model across these categories. It's pivotal to recognize that in our task, an ideal model would demonstrate high precision and recall across all three labels.

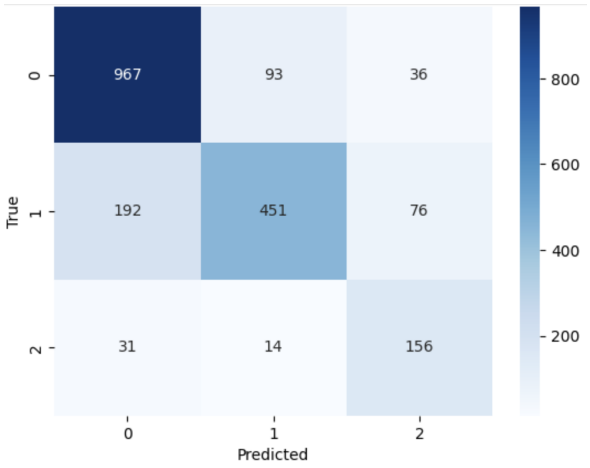

Figure 3: Confusion Matrix

The model categorizes text into non-violence (label 0), passive violence (label 1), and direct violence (label 2) with an overall macro F1 score of 0.74. It particularly excels in identifying non-violence texts. It also demonstrates aptitude in recognizing passive violence texts. However, it faces challenges in the realm of direct violence.

6 Conclusion

In this paper we described the nlpBDpatriots approach to the VITD shared task. We evaluated various models on the data provided by the shared task organizers, namely statistical machine learning models, transformer-based models, few shot prompting, and some customization with transformer-based models with multilinguality, back translation, and two-step classification. We show that the two-step classification procedure with multilinguality and back translation is the most successful approach achieving a macro F1 score of 0.74.

Our two-step approach towards solving the problem presented for this shared task shows promising results. However, the relatively small size of the dataset made it difficult for the other pre-trained models to learn informative features that would help them perform classification. Also, the dataset contains three imbalanced labels making it easy for the models to overfit. Our approach with data augmentation and two-step classification generates good results, but it is still below one of the three baseline results announced by the organizers prior to the start of the competition.

Acknowledgment

We would like to thank the VITD shared task organizing for proposing this interesting shared task. We further thank the anonymous reviewers for their valuable feedback.

References

Mario Aragon, Miguel Angel Carmona, Manuel Montes, Hugo Jair Escalante, Luis Villaseñor-Pineda, and Daniela Moctezuma. 2019. Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets. In *Proceedings of IberLEF*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Mithun Das, Somnath Banerjee, Punyajoy Saha, and Animesh Mukherjee. 2022. Hate speech and offensive language detection in bengali. In *Proceedings of AACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Google. 2021. Google cloud translation api documentation. Accessed: 2023-08-28.

Md Saroar Jahan, Mainul Haque, Nabil Arhab, and Mourad Oussalah. 2022. BanglaHateBERT: BERT for abusive language detection in Bengali. In *Proceedings ResTUP*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

M Kowsher, Abdullah As Sami, Nusrat Jahan Prottasha, Mohammad Shamsul Arefin, Pranab Kumar Dhar, and Takeshi Koshiba. 2022. Bangla-bert: transformer-based efficient model for transfer learning and language understanding. *IEEE Access*, 10:91855–91870.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of TRAC*.

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

Sandip Modha, Thomas Mandl, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Tharindu Ranasinghe, and Marcos Zampieri. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech. In *Proceedings of FIRE*.

OpenAI. 2023. Gpt-3.5 turbo fine-tuning and api updates. Accessed: 2023-08-28.

Nasif Istiak Remon, Nafisa Hasan Tuli, and Ranit Debnath Akash. 2022. Bengali hate speech detection in public facebook pages. In *Proceedings of ICISET*.

Nauros Romim, Mosahed Ahmed, Md Saiful Islam, Arnab Sen Sharma, Hriteshwar Talukder, and Mohammad Ruhul Amin. 2022. Bd-shs: A benchmark dataset for learning to detect online bangla hate speech in different social contexts. In *Proceedings of LREC*.

Sourav Saha, Jahedul Alam Junaed, Maryam Saleki, Mohamed Rahouti, Nabeel Mohammed, and Mohammad Ruhul Amin. 2023a. Blp-2023 task 1: Violence inciting text detection (vitd). In *Proceedings of the 1st International Workshop on Bangla Language Processing (BLP-2023)*.

Sourav Saha, Jahedul Alam Junaed, Maryam Saleki, Arnab Sen Sharma, Mohammad Rashidujjaman Rifat, Mohamed Rahout, Syed Ishtiaque Ahmed, Nabeel Mohammad, and Mohammad Ruhul Amin. 2023b. Vio-lens: A novel dataset of annotated social network posts leading to different forms of communal violence and its evaluation. In *Proceedings of BLP*.

Md Anwar Hussen Wadud, Md Abdul Hamid, Muhammad Mostafa Monowar, and Atif Alamri. 2021. L-boost: Identifying offensive texts from social media post in bengali. *Ieee Access*, 9:164681–164699.

Tharindu Cyril Weerasooriya, Sujan Dutta, Tharindu Ranasinghe, Marcos Zampieri, Christopher M Homan, and Ashiqur R KhudaBukhsh. 2023. Vicarious offense and noise audit of offensive speech classifiers: Unifying human and machine disagreement on what is offensive. In *Proceedings of EMNLP*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of SemEval*.

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). In *Proceedings of SemEval*.

Haris Bin Zia, Ignacio Castro, Arkaitz Zubiaga, and Gareth Tyson. 2022. Improving zero-shot cross-lingual hate speech detection with pseudo-label fine-tuning of transformer language models. In *Proceedings of ICWSM*.