NLP_CUET at BLP-2023 Task 1: Fine-grained Categorization of Violence Inciting Text using Transformer-based Approach

Jawad Hossain, Hasan Mesbaul Ali Taher, Avishek Das and **Mohammed Moshiul Hoque**
Department of Computer Science and Engineering
@Chittagong University of Engineering & Technology, Chattogram-4349, Bangladesh
{u1704039, u1804038}@student.cuet.ac.bd, {avishek, moshiul_240}@cuet.ac.bd

Abstract

The amount of online textual content has increased significantly in recent years through social media posts, online chatting, web portals, and other digital platforms due to the significant increase in internet users and their unprompted access via digital devices. Unfortunately, the misappropriation of textual communication via the Internet has led to violence-inciting texts. Despite the availability of various forms of violence-inciting materials, text-based content is often used to carry out violent acts. Thus, developing a system to detect violence-inciting text has become vital. However, creating such a system in a low-resourced language like Bangla becomes challenging. Therefore, a shared task has been arranged to detect violence-inciting text in Bangla. This paper presents a hybrid approach (GAN+Bangla-ELECTRA) to classify violence-inciting text in Bangla into three classes: *direct*, *passive*, and *non-violence*. We investigated a variety of deep learning (CNN, BiLSTM, BiLSTM+Attention), machine learning (LR, DT, MNB, SVM, RF, SGD), transformers (BERT, ELECTRA), and GAN-based models to detect violence inciting text in Bangla. Evaluation results demonstrate that the GAN+Bangla-ELECTRA model gained the highest macro f_1-score (74.59), which obtained us a rank of 3rd position at the BLP-2023 Task 1.

1 Introduction

Violence-inciting text refers to textual content that promotes or glorifies acts of violence or harm towards individuals, groups, or entities, including hate speech and extremist ideologies. Detecting such text is crucial for preventing harmful activity and maintaining safety on social media. Social media's widespread use by diverse religious and cultural factions has led to weaponization, inciting hatred and causing communal violence, resulting in significant loss of life and destruction. This issue persists not only in a specific geographical region but also globally, escalating the longstanding issue. This paper aims to classify various forms of communal violence to illuminate this complex phenomenon and contribute to its mitigation.

Violence has evolved with society's advancements, with physical and psychological abuse now predominantly occurring online and on social networks, even though it was once face-to-face (Golem et al., 2018). Previous studies reveal that social media platforms incite political and religious violence, thereby threatening communal harmony and societal stability (Patton et al., 2014). Social networks have become a virtual civilization where people share views, feelings, photos, videos, and blogs. However, there is no defined mechanism for restricting violent content on these platforms (Yadav and Manwatkar, 2015). In recent years, tech giants like Facebook, YouTube, and Twitter have been striving to achieve this goal (Ghanghor et al., 2021). However, it is impossible to manually monitor these violent inciting contents that surf social media (Sharif and Hoque, 2022). Therefore, developing such a system for detecting violence-inciting text is crucial to reducing illegal behavior and maintaining a clean information ecosystem.

This work aims to build a system that can detect violence inciting text from Bangla text concerning three different categories. This work's key contributions are illustrated in the following:

- Developed a hybrid model using GAN and Bangla-ELECTRA to detect and classify violence-inciting Bangla texts into three groups: direct violence (DV), passive violence (PV), and non-violence (NV).
- Investigated the model's effectiveness in detecting and classifying violence-inciting texts by comparing several ML, DL, and transformer-based models and analyzed in-depth errors, offering valuable insights into violence-inciting text detection.

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 241–246

2 Related Work

While providing platforms for individual freedom of expression, social media and other blogging platforms can facilitate antisocial conduct, including hate speech, cyberbullying, and online harassment (Karim et al., 2021). Several works has been conducted to detect aggressive comments (Sharif and Hoque, 2022), abusive comment, hate speech (Das et al., 2021), trolling (Zampieri et al., 2019). However, few studies have been conducted to detect violence-inciting text. Though several works have been done in high-resource languages, leaving low-resource languages like Bangla out of the focus. To identify abusive language, Eshan and Hasan, 2017 utilized a dataset comprising 2.5k instances of abusive Bangla text and evaluated the performance of several ML models (RF, NB, and SVM) and achieved a maximum accuracy of 85% using SVM with linear kernel and tri-gram features. Kumar et al., 2018 categorized 15k English and Hindi comments on aggression into overtly aggressive, covertly aggressive, and non-aggressive categories, expanding the corpus to include Bangla aggressive comments (Kumar et al., 2020). Aroyehun and Gelbukh, 2018 studied the effectiveness of DNN models in detecting aggression using enhanced data and pseudo-labeled samples. Ishmam and Sharmin, 2019 classified 5k Bangla abusive Facebook comments into six categories using a GRU-based model, achieving 70.10% accuracy. The introduction of BERT-based models significantly enhanced performance, surpassing all previous models on these datasets (Risch and Krestel, 2020, Sharif et al., 2021). Sharif et al., 2021 presented a Bangla aggressive text dataset, and later, they extended the previous dataset to create a new novel dataset named *BAD*. They used a transformer-based ensemble technique to identify and categorize aggressive texts in Bangla, achieving the highest weighted scores of 93.43% (coarse-grained) and 93.11% (fine-grained). As per our exploration, none of the past studies addressed classifying the violence-inciting texts in Bangla. This work uses a hybrid approach incorporating GAN and Bangla-ELECTRA models to address the downstream task.

3 Task and Dataset Descriptions

Task organizers[1] created a gold standard corpus to detect violence-inciting language in social media. To address this phenomenon, Saha et al., 2023 developed a Violence Inciting Text Detection (VITD) corpus[2] in the Bangla language. The task aims to implement a system that can detect offensive texts. The corpus consists of the text of three different classes: *non-violence*, *passive violence*, and *direct violence*. According to Saha et al., 2023, the definition of each class is illustrated in the following:

- **Direct Violence (DV):** Texts expressing explicit threats fall under direct violence.
- **Passive Violence (PV)**: Texts containing abusive or derogatory use of language.
- **Non-Violence (NV)**: The non-violence category consists of any discussions conducted by texts that do not involve any form of violence.

The dataset (VITD) accumulated 6046 texts from YouTube comments in Bangla. VITD is related to nine violent incidents during the previous 10 years. The task aims to quickly distinguish between violent threats to stop further incitement to violent acts. Contribution to the identification and prevention of stimulation to violent acts online is the primary goal of this task.

Table 1 illustrates the detailed statistics of the dataset. The dataset consists of training, validation, and test sets containing 2700, 1330, and 2016 texts. The dataset is imbalanced as there are more non-violence samples than direct and passive violence combined. The non-violence class includes the highest data (1389 texts) with 7128 unique words.

Table 1: Summary of the dataset statistics.

Classes	Train	Valid	Test	Total words
DV	389	196	201	13202
PV	922	417	719	39423
NV	1389	717	1096	54333
Total	2700	1330	2016	106958

We further analyzed the dataset in terms of sentence length. Figure 1 shows the length-frequency distribution of the dataset. The analysis of the length-frequency distribution revealed that there were fewer than 50 text samples whose text length was more than 128 words. Thus, this work used a maximum input sentence size of 128 words. The minimum sentence length is one word, whereas the average length is 18 words.

[1] https://blp-workshop.github.io/sharedtasks

[2] https://github.com/blp-workshop/blp_task1

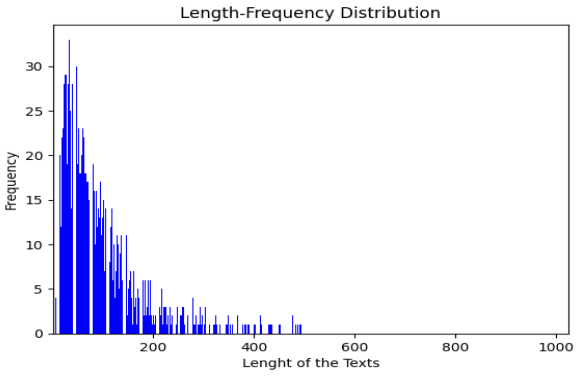

Figure 1: Length-frequency distribution of the dataset

4 Methodology

This work exploited several ML, DL, and transformer-based models to address the task. After investigating all models' performance, this work proposes a hybrid method using GAN and Bangla-ELECTRA to detect and classify violence-inciting Bangla texts. We used the 'scikit-learn' and TensorFlow library to build ML and DL models. Figure 2 shows an abstract view of the proposed system.

First, the unwanted characters (URLs, punctuation, and whitespace) are removed from the texts. We apply different feature extraction techniques (i.e., TF-IDF, Word2Vec) to extract the textual features. This work employed six traditional ML models, such as logistic regression (LR), decision tree (DT), support vector machine (SVM), multinomial naive Bayes (MNB), random forest (RF), and stochastic gradient descent (SGD). We also used three DL methods, such as CNN, BiLSTM, and BiLSTM, with attention.

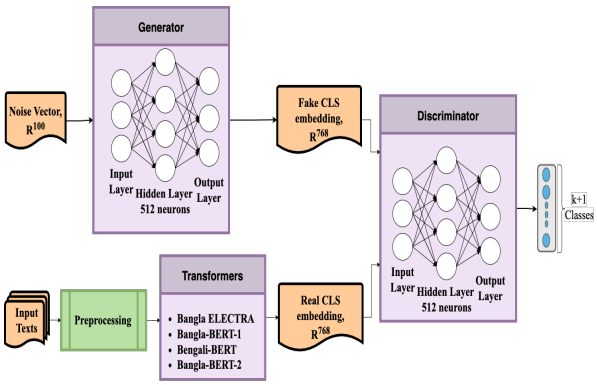

Figure 2: Proposed hybrid model using GAN and Bangla-ELECTRA to detect and classify violence inciting Bangla texts.

This work employed four transformers fetched from HuggingFace[3] library. We built the transformer models with PyTorch library to tackle the task, such as Bangla-ELECTRA (Bhattacharjee et al., 2021), Bangla-BERT-1 (Sarker, 2020), Bangla-BERT (Joshi, 2022), and Bangla-BERT-2 (Kowsher et al., 2022).

4.1 GAN+Bangla-ELECTRA

In the GAN, we used 2 sub-networks: generator and discriminator. The generator takes input noise and outputs fake data, which tries to resemble the original data distribution. The discriminator is trained over a $(k+1)$-class objective: the *true* examples are classified in one of the target $(1, ..., k)$ classes, while the generated samples are classified into $k+1$ class. The adversarial training procedure is applied (Goodfellow et al., 2020). The generator is penalized each time the discriminator discovers its output as fake. The discriminator is penalized each time the generator fools it; that is, it can identify the fake data created by the generator as real.

In the GAN+transformer-based approach (Croce et al., 2020), we consider labeled and unlabeled data where unlabeled data is accumulated by removing the label. The generator and discriminator are both multilayer perceptrons with a single hidden layer of 512 neurons. The input of the generator is a randomly generated vector of 100 dimensions, and it outputs a fake transformer embedding vector for a single token. The transformer-based model (BERT, ELECTRA) feeds the input text, generating a contextualized embedding vector of the CLS token. The embedding vectors generated by both the transformer and generator are used as input for the discriminator. The input of the discriminator can be expressed by Eq. 1.

$$H_* \epsilon R^D \tag{1}$$

Where, H_* can be either H_{FAKE} or H_{CLS}. H_{FAKE} denotes the outputs of the generator and H_{CLS} is the output of the transformer model. The output of the discriminator is extended to $k+1$ classes, where k is the number of classes, and the extra class is 'REAL'. The system identifies using $k+1^{th}$ class whether the embedding encoded by the transformer-based architecture is real or not. The goal is to acquire a good discriminator in k-class classification. The discriminator and final classification are defined by Eqs. 2-3.

$$D_{logits} = MLP(x) \tag{2}$$

[3]https://huggingface.co/models

$$P_{class_i} = \frac{e^{D_{logits_i}}}{\sum_{k=1}^{k+1} e^{D_{logits_k}}} \quad (3)$$

Here D_{logits} is the output of passing the input vector 'x' through the multi-layer neural network of the discriminator. P_{class_i} denotes the probability of a text sequence belonging to a given class.

A dropout rate of 0.1 is added to both the generator and discriminator architecture to prevent overfitting. The Adam optimizer with a batch size of 16 and a learning rate of 5e-5 is used to train the models for 10 epochs. For testing, we just discard the generator and use the BERT and discriminator model to classify the input data. We mask the prediction output for the 'REAL' class in testing.

5 Results

The efficacy of the models is determined based on the macro-F1 score (MF1). However, we also consider the precision (P) and recall (R) metrics to perform the analysis. Table 2 illustrates the performance of employed models for the task. Among the ML models, SGD achieved the highest MF1 score of 65.34, while LR performed poorly on the test set. On the other hand, DL-based methods did not surpass the performance of the best ML model (MF1 score of 65.34). Low amounts of data samples might cause this, as DL models are generally data-hungry. Adding attention (Vaswani et al., 2017) to BiLSTM improved its performance by almost 3.12%. All transformer-based models outperformed the ML and DL models, with Bangla-BERT scoring the highest (68.36). Although the GAN-based transformer models improved the scores of their respective transformers, the Bangla-BERT-based standalone and GAN-based models performed almost identically. GAN+Bangla-ELECTRA outperformed all the models, achieving the highest f1-score of 74.59. With the GAN+transformer approach, the inner representation of BERT is being fine-tuned by both labeled and unlabeled data. For this reason, the inner representation of BERT is more robust towards unseen data points.

Table 2: Performance of various models on the test set

Classifier	**P**	**R**	**MF1**
LR	63.08	57.34	29.28
DT	59.89	59.72	53.11
RF	71.88	68.01	59.92
MNB	69.07	68.80	63.91
SVM	73.01	65.62	55.50
SGD	71.34	70.68	65.3
CNN	66.67	65.58	57.26
BiLSTM	67.72	66.91	60.02
BiLSTM + Attention	67.83	67.81	61.89
Bangla-ELECTRA	72.34	72.77	67.18
Bangla-BERT-1	71.88	71.92	66.45
Bangla-BERT	76.13	73.12	68.36
Bangla-BERT-2	75.25	72.97	67.05
GAN+Bangla-BERT-1	71.31	71.23	66.33
GAN+Bangla-BERT-2	75.04	74.21	69.66
GAN+Bangla-BERT	76.32	76.49	72.35
GAN+Bangla-ELECTRA	**77.98**	**77.43**	**74.59**

Table 3 shows the class-wise performance (MF1) of hybrid models. Results demonstrated that the proposed approach (GAN+Bangla-ELECTRA) attained the highest scores in all classes than the other hybrid models.

Table 3: Class-wise violence inciting text detection performance on the test set

Class	NV	PV	DV
GAN+Bangla-BERT-1	0.79	0.62	0.58
GAN+Bangla-BERT-2	0.81	0.68	0.60
GAN+Bangla-BERT	0.82	0.70	0.65
GAN+Bangla-ELECTRA	**0.82**	**0.73**	**0.69**

5.1 Error Analysis

A detailed error analysis is performed quantitatively and qualitatively to provide in-depth insights into the performance of the proposed model.

Quantitative Analysis: A quantitative error analysis of the best-performed model is done using the confusion matrix (Fig. 3). The proposed GAN+Bangla-ELECTRA classified a total of 1561 samples correctly out of 2016 samples in the test dataset. The model did comparatively better results in the NV class. The model identified 910 instances of the NV class correctly. It incorrectly classified 171 samples as NV class of which 150 data samples were originally from PV and 21 data samples were originally from DV. The model becomes more confused between NV and PV as it misclassified a total of 311 instances between the two classes, whereas the instances that were misclassified as DV and the DV true instances that were misclassified as NV or PV total only 144. This may happen because a regular discussion with one person might be a derogatory or abusive use of language to another, as some words can be used for both peaceful and violent discussions.

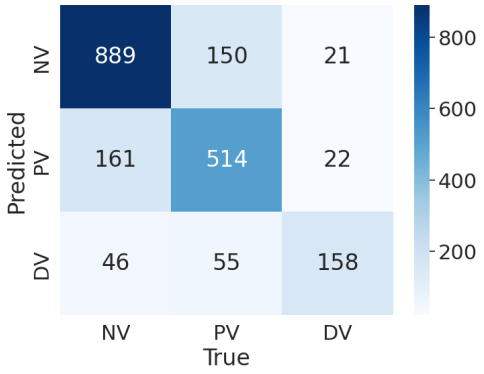

Figure 3: Confusion matrix of the proposed model ((GAN+Bangla-ELECTRA)

Qualitative Analysis: Figure 4 illustrates some predicted outcomes by the proposed model. The

Text sample	Actual	Predicted
Sample1. 'সাংবাদিকতা করতে এসে ঠান্ডা সন্ত্রাসী বন্ধ করো।' (Stop being a cold terrorist when you come to do journalism.)	PV	PV
Sample2. 'ভারতীয় দালাল নির্মূল এখন সময়ের দাবি, ভারত হলো বাংলাদেশের সকল সমস্যার মূল হোতা।' (Eradication of Indian brokers is the need of the hour, India is the root of all problems in Bangladesh.)	PV	DV
Sample3. 'এরশাদ ভালো না হইলেও ওর ভাই জিএম কাদের ভালো আছে।' (Even if Ershad is not good, his brother GM Kader is good.)	NV	NV
Sample4. 'সত্য কথা বলার জন্য ধন্যবাদ।' (Thanks for speaking the truth.)	NV	NV
Sample5. 'মোসল্লীরা ইটপাটকেল ছুড়তে দেখলাম না মিথ্যা বলছেন কেন?' (I didn't see Muslims throwing bricks, why are they lying?)	NV	PV

Figure 4: Few examples of predicted outputs by the proposed (GAN+Bangla-ELECTRA) model

proposed model correctly predicts text samples 1, 3, and 4, whereas text samples 2 and 5 are not predicted correctly. Text sample 2 is wrongly predicted as DV, whereas the actual class is PV. Similarly, text sample 5 is incorrectly predicted as PV instead of actual class (NV). The class imbalance issue might be the reason for wrong predictions, as a few instances of the DV class (201 samples) are available in the dataset. This scarcity of samples may be inadequate for the model to learn. Another reason might be that the words used in DV do not often overlap with the largest class (i.e., NV).

6 Conclusion

This work addresses the challenge of fine-grained classification of texts inciting violence in Bangla. We developed a solution by leveraging a benchmark dataset known as VITD. In this paper, we systematically investigated and compared 17 different baseline models, spanning various machine learning (ML), deep learning (DL), transformer, and generative adversarial network (GAN) architectures. The experimentation revealed that integrating GANs with transformers resulted in improved task performance. Specifically, the combination of GAN and Bangla-ELECTRA demonstrated the highest macro F1-score (74.59) among all the models we employed, surpassing their performance. We intend to enhance our solution by leveraging ensemble techniques in future research endeavors. Additionally, we will delve into the impact of resampling strategies on model performance, mainly as our dataset exhibits imbalance issues.

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.

Abhik Bhattacharjee, Tahmid Hasan, Kazi Samin, M Sohel Rahman, Anindya Iqbal, and Rifat Shahriyar. 2021. Banglabert: Combating embedding barrier for low-resource language understanding. *arXiv preprint arXiv:2101.00204*.

Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 2020. Gan-bert: Generative adversarial learning for robust text classification with a bunch of labeled examples.

Amit Kumar Das, Abdullah Al Asif, Anik Paul, and Md Nur Hossain. 2021. Bangla hate speech detection on social media using attention-based recurrent neural network. *Journal of Intelligent Systems*, 30(1):578–591.

Shahnoor C Eshan and Mohammad S Hasan. 2017. An application of machine learning to detect abusive bengali text. In *2017 20th International Conference of Computer and Information Technology (ICCIT)*, pages 1–6. IEEE.

Nikhil Ghanghor, Parameswari Krishnamurthy, Sajeetha Thavareesan, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2021. Iiitk@ dravidianlangtech-eacl2021: Offensive language identification and meme classification in tamil, malayalam and kannada. In *Proceedings of the first workshop on speech and language technologies for Dravidian languages*, pages 222–229.

Viktor Golem, Mladen Karan, and Jan Šnajder. 2018. Combining shallow and deep learning for aggressive text detection. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 188–198.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative

adversarial networks. *Communications of the ACM*, 63(11):139–144.

Alvi Md Ishmam and Sadia Sharmin. 2019. Hateful speech detection in public facebook pages for the bengali language. In *2019 18th IEEE International Conference on machine learning and Applications (ICMLA)*, pages 555–560. IEEE.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Md Azam Hossain, and Stefan Decker. 2021. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE.

Md 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. 2020. Evaluating aggression identification in social media. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 1–5.

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

Desmond Upton Patton, Jun Sung Hong, Megan Ranney, Sadiq Patel, Caitlin Kelley, Rob Eschmann, and Tyreasa Washington. 2014. Social media as a vector for youth violence: A review of the literature. *Computers in Human Behavior*, 35:548–553.

Julian Risch and Ralf Krestel. 2020. Bagging bert models for robust aggression identification. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 55–61.

Sourav Saha, Jahedul Alam Junaed, Maryam Saleki, Arnab Sen Sharma, Mohammad Rashidujjaman Rifat, Mohamed Rahout, Syed Ishtiaque Ahmed, Nabeel Mohammad, and Mohammad Ruhul Amin. 2023. Vio-lens: A novel dataset of annotated social network posts leading to different forms of communal violence and its evaluation. In *Proceedings of the 1st International Workshop on Bangla Language Processing (BLP-2023)*, Singapore. Association for Computational Linguistics.

Sagor Sarker. 2020. Banglabert: Bengali mask language model for bengali language understanding.

Omar Sharif and Mohammed Moshiul Hoque. 2022. Tackling cyber-aggression: Identification and fine-grained categorization of aggressive texts on social media using weighted ensemble of transformers. *Neurocomputing*, 490:462–481.

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2021. Nlp-cuet@ dravidianlangtech-eacl2021: Offensive language detection from multilingual code-mixed text using transformers. *arXiv preprint arXiv:2103.00455*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Shashank H Yadav and Pratik M Manwatkar. 2015. An approach for offensive text detection and prevention in social networks. In *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, pages 1–4. IEEE.

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). *arXiv preprint arXiv:1903.08983*.