# Automatically Detecting Cyberbullying Comments on Online Game Forums

Hanh Hong Phuc Vo[1,2,*], Hieu Trung Tran[1,2,*], and Son T. Luu[1,2,†]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University Ho Chi Minh City, Vietnam
Email: *{*18520275,18520754*}*@gm.uit.edu.vn, †*sonlt@uit.edu.vn*

*Abstract*—**Online game forums are popular to most of game players. They use it to communicate and discuss the strategy of the game, or even to make friends. However, game forums also contain abusive and harassment speech, disturbing and threatening players. Therefore, it is necessary to automatically detect and remove cyberbullying comments to keep the game forum clean and friendly. We use the Cyberbullying dataset collected from World of Warcraft (WoW) and League of Legends (LoL) forums and train classification models to automatically detect whether a comment of a player is abusive or not. The result obtains 82.69% of macro F1-score for LoL forum and 83.86% of macro F1-score for WoW forum by the Toxic-BERT model on the Cyberbullying dataset.**

*Index Terms*—**game forums, cyberbullying, machine learning, text classification, deep neural network, bert.**

## I. Introduction

Nowadays, offensive comments are common issues on gaming forums. These comments often have negative, harmful, and abusive, causing a pessimistic impression on the gaming community. To prevent offensive comments by users on forums, administrators have to manually find each comment, then hide it or ban the user with abusive comment. However, this approach seems impossible on big forums with huge user comments. Therefore, automatically detecting offensive comments on game forums is necessary to help administrators find and hide negative comments or even ban users who post abusive and harassment comments. We solve the problem of automatically detecting cyberbullying comments on the forum of League of Legends and World of Warcraft, which are popular to most of game players.

We use machine learning approaches for text classification tasks to automatically classify whether a comment from a game player is abusive or not. We also propose alternative solutions such as data pre-processing and data enhancement to increase the performances of classifiers. Then, we evaluate the empirical results to find the best models on the Cyberbullying dataset. Our approaches for machine learning include: traditional machine learning models(Logistic Regression, SVM), deep neural models (Text-CNN, GRU), and transfer learning model (Toxic BERT).

Finally, based on the empirical results, we choose the highest performance model to build a tool for helping administrators detect and hide abusive comments on game forums.

The paper is organized as follow. Section 1 introduces the topic and motivation to study the aforementioned task. Section 2 presents related work on the problem of identifying abusive comments. Section 3 describes the Cyberbullying dataset that used for training machine learning models. Section 4 and Section 5 illustrate our approaches and empirical results. Finally, Section 6 concludes our work.

## II. Related work

In 2017, Wulczyn et al. analyzed Wikipedia comments to find personally offensive comments with a dataset of over 100,000 high-quality human-labeled comments and 63M machine-labeled ones [1]. The data set achieved the best AUG result (96.59%) using multi-layer perceptrons (MLP) model on ED (empirical distribution) label with char n-gram. In the case of a model trained on ED labels, the attack score represents the predicted fraction of annotators who would consider the comment an attack.

Also in 2017, Davidson et al. conducted the study on Automated Hate Speech Detection and the Problem of Offensive Language [2]. The authors collected a dataset from tweets on Twitter then trained them with machine learning models such as Logistic Regression, Naive Bayes, Decision Trees, Random Forests, and Linear SVMs. After training, the best performance model, which is Logistic Regression, has the results of 91.00% by Precision and 90.00% by Recall and F1-scores.

In addition, in 2018, Ribeiro et al. performed characterizing and detecting hateful users on Twitter [3]. They also collected this dataset on Twitter. However, unlike other datasets, this dataset focuses on the user, not the content of the user comments. The author crawled the dataset from 100,386 users with up to 200 tweets per user. Then, they selected a subsample to be annotated. It has 4,972 comments that are manually annotated users, of which 544 users labeled as hateful. They also implement dataset models such as GraphSage, AdaBoost, GradBoost. Using both user's feature and glove properties together with the GraphSage model gives the best results. Additionally, AdaBoos delivers a good AUC score, but misclassifying many of the normal users is negative, resulting in low accuracy and F1-score.

Besides, there are also many datasets for abusive comment recognition built in multilingual, such as hatEval dataset from SemEval-2019 Task 5 contest (2019) [4] on identifying offensive comments about women and migrants in English and Spanish, or CONAN dataset is used to identify comments inciting violence in the English, French and Italian languages [5].

## III. DATASET

The Cyberbullying dataset[1] [6] includes two parts: comments of players on League of Legends (LoL) and World of Warcraft (WoW) forums. The data from League of Legends forum contains 17,354 comments with 207 offensive cases, and the data from World of Warcraft forum contains 16,975 comments with 137 offensive cases. All comments are annotated manually and reviewed by experts. The dataset contains English comments on 20 different topics. There are two labels in the dataset, label 1 indicates offensive comments, and label 0 indicates non-offensive comments. Table I and Figure 1 show the data distribution of labels on two forums - League of Legends and World of Warcraft.

In addition, according to Table II, comments are often written in an informal form, using a lot of slang, abbreviations, strange characters and conversations centered on the exchanges between players on game forums. Through the Wordcloud from Figure 2, it can be seen that comments from the Cyberbullying dataset often use informal words such as abbreviations and slang, for example: "realli", "bullsbeept", and "forev". This is a common feature for comments on social media platforms [7].

TABLE I: Distribution of labels by percentage on two forums in the Cyberbullying dataset.

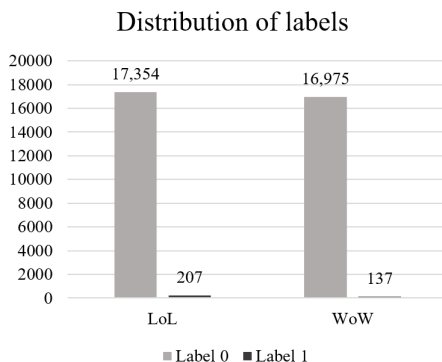| Forum | Label 0 (%) | Label 1 (%) |
|---|---|---|
| League of Legends | 98.81 | 1.19 |
| World of Warcraft | 99.19 | 0.81 |



Fig. 1: Distribution of labels in the Cyberbullying dataset.

Furthermore, as showed on Table I and Figure 1 the distribution of labels is imbalanced on both LoL and WoW

[1]http://ub-web.de/research/

forums in the dataset. Most of the comments are skewed to label 0, which account for 98.81% on the LoL forum and 99.19% on the WoW forum, respectively. This indicates that Cyberbullying dataset is imbalanced, and most of comments are non-offensive comments.

TABLE II: Samples in Cyberbullying dataset.

| No. | Comment | Label |
|---|---|---|
| | **Comments on League of Legends forum** | |
| 1 | We can only hope. | 0 |
| 2 | I hit a ping spike that maxed out at 21,000+ ping. Find it hard to believe thats on my end. | 0 |
| 3 | I have spotted signs of Photoshop in the pictures you linked. You are a lying despicable thief and you weren't hacked you tried to use that as a cover up for your friend who ****ed up your account now you want sympathy and I bet you won't call the police because you mess with the police over this **** that's not even true they might lock your ass up for wasting their time you better now go back and grind the runes again that your FRIEND you gave account access to ruined this was not A HACKER YOU LIAR AND ONLY IDIOTS BELIEVE YOU. | 1 |
| | **Comments on World of Warcraft forum** | |
| 1 | I have to be honest here I got bored right after I got level 100, at that point all my enthusiasm was gone and the first thing that came in my mind was: "" here we go again"" gearing, profession levelling. Nothing new and fun. | 0 |
| 2 | Presuming someone was boosted in CMs cause he has CM achis- moron logic. See how that works? | 0 |
| 3 | Do you give a damn one way or the other, f1eab69493d84bd8, or are you just trolling? I've seen some interesting comments on both sides of the debate from people here, but nothing from you aside from trolling. As you clearly have no interest in a Vanilla server and wouldn't play on one if Blizzard launched one, why are you even posting in this thread if not for sheer troll value? | 1 |



League of Legends



World of Warcraft

Fig. 2: Wordcloud of Cyberbullying dataset.

Finally, Table III and Figure 3 show the average length of comments in the Cyberbullying dataset. The average length of comments is computed by word level. According to Table III, the average length of comments on label 0 and label 1 are 102.95 and 137.60 respectively on the LoL forum, and 105.42 and 165.68 respectively on the WoW forum. In general, the

average length of comments on label 1 are longer than the average length of comments on label 0 for both two forums.

TABLE III: Average length of comments in the Cyberbullying dataset.

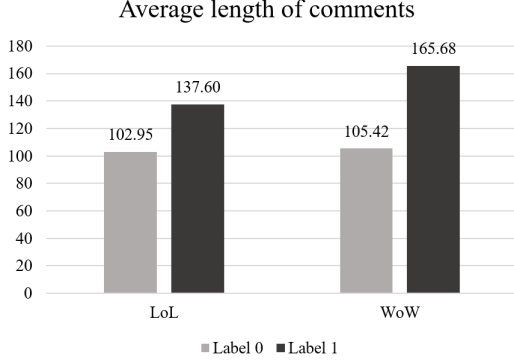| Forum | Label 0 | Label 1 |
|---|---|---|
| League of Legends | 102.95 | 137.60 |
| World of Warcraft | 105.42 | 165.58 |

Average length of comments



Fig. 3: Average length of comments in Cyberbullying dataset.

## IV. METHODOLOGIES

### A. Text classification task

Text classification is one of the most common tasks in machine learning and natural language processing. Aggarwal and Zhai (2012) defined the text classification task as given the training set D = $\{X_1, ..., X_n\}$, where each data point $X_i \in D$ is labeled with a value derived from the set of labels are numbered $\{1...k\}$. The training data is built into a classification model. With the test set given, the classification model predicts the label for each data point in the test set [8].

We apply the text classification task for our problem to classify whether a comments come from a player is toxic or non-toxic. Figure 4 describes our model overview for detecting cyberbullying comments.
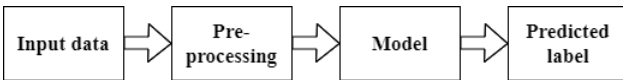


Fig. 4: Model overview for detecting cyberbullying comments.

With input comments, the model does pre-processing steps including splitting comments into tokens, removing unimportant characters, using SMOTE to generate new examples and using word embedding (GloVe, fastText) to present vectors for input data. Next, the pre-processed data are fit into classification models including traditional machine learning models, deep neural models, and transformer learning model. Finally, the output is labels that was predicted by classification models.

### B. SMOTE

Synthetic Minority Oversampling Technique (SMOTE) is a technique to increase the number of samples for the minority

class. It is widely used to handle imbalanced datasets. SMOTE works by selecting data points in the feature space and drawing a line between those points and the nearest neighbors. Then based on this line to select new data for the minority class. New data points are generated from the original data points to increase the amount of minority class. This is an effective method for solving imbalance in datasets [9].

### C. Word embedding

Word embedding is a method of word representation that used in natural language processing. This method represents a word (or phrase) by a real number vector that was mapped from a corpus of text. Using word embedding improves the performance of the predictive model [10].

In this paper, we use sets of word representation vectors such as GloVe [11] and fastText [12]. The glove.6B.300d word embedding[2] comprises 300 dimensions vectors over 400,000 vocabularies, which is built on Wikipedia and Gigaword corpora. The fastText [3] is trained on Common Crawl and Wikipedia datasets using CBOW with position-weights in 300 dimensions with 5-grams features.

### D. Traditional machine learning models

**Logistic Regression** is one of the basic and well-known classification algorithms commonly used for binary classification problems. For classification problems, we have to convert texts to vectors before fitting them into the training model [13].

**Support Vector Machine (SVM)** is a classification algorithm based on the concept of hyper-planes, which divide data into different classes. This algorithm is commonly used for classification and regression problems. Similar to Logistic Regression, SVM converts text data into vector form before fitting into the training model [14].

### E. Deep neural models

**Text-CNN** is a multi-layered neural network architecture developed for the classification problem. CNN model uses Convolutional layers (CONV layer) to extract valuable information from data [15].

**GRU** is an improved version of standard recurrent neural network(RNN). To solve the vanishing gradient problem of a standard RNN, GRU uses special gates including update gate and reset gate. Basically, these are two vectors which decide what information should be passed to the output. The special thing about them is that they can be trained to keep information from long ago, without washing it through time or remove information which is relevant to the prediction [16].

### F. Transformer model

BERT is a transformer architecture created by Devlin et al. [17] and become popular to many of natural language processing tasks. With the pre-trained models on large and unsupervised data, BERT can be fine-tuned later in specific

---

[2]https://nlp.stanford.edu/projects/glove/
[3]https://fasttext.cc/docs/en/crawl-vectors.html

downstream tasks such as text classification and question answering. Toxic-BERT (also called as Detoxify) [18] is a model based on BERT architecture that used for toxic comments classification task. Toxic-BERT are trained on three different toxic dataset comes from three Jigsaw challenges. We implement the Toxic-BERT model on the Cyberbullying dataset for detecting cyberbullying comments from players.

## V. EMPIRICAL RESULTS

### A. Model parameters

**Logistic Regression**: We implement the Logistic Regression with random_state equal to 0, C equal to 1, and max_feature equal to 13,000.

**SVM**: We implement the SVM with linear kernel, C equal to 1, and max_feature equalt to 13,000.

**Text-CNN:** We implement the Text-CNN with 5 epochs, batch_size equal to 512, sequence_length equal to 300, conv_layer_size equal to 5, 128 units, and using sigmoid activation function.

**GRU:** We implement the GRU with 5 epochs, batch_size equal to 512, sequence_length equal to 300, 128 units, and dropout equal to 0.1.

**Toxic-BERT:** We implement the Toxic-BERT with 5 epochs, train_batch_size equal to 16, and test_batch_size equal to 8. We used the Simple transformer[4] for implementing the Toxic-BERT model.

### B. Data preparation

Before running the model, we do pre-processing steps as follows: (1) we remove most of the special characters in the comments (except for the "*") and keep only the letters. The characters "*" represent encoded offensive words. These are handled by changing to the word "beep". (2) Then splitting comments into tokens using TweetTokenizer of NLTK library. In addition, all comments are converted to lowercase, and (3) we use stop words to get rid of less important words. Stop words are words like "the", "in", "a", "an" they have less meaning in the sentence.

We use the k-fold cross-validation with the parameter k = 5 on the Cyberbullying dataset. The dataset is randomly divided into 5 equal parts with proportion 8:2 for train set and test set respectively. The final result is the mean of 5 folds.

### C. Empirical results

Table IV presents the results of the models on two forums, LoL and WoW of the dataset. According to Table IV, the results of the models on the two forums are similar.

For traditional machine learning models, the macro F1-score and accuracy are quite high disparities. The reason is the imbalanced distribution of labels in the dataset as described in Section III.

For deep neural models, the Text-CNN model with the GloVe word embedding gives the best results by macro F1-score, which are 80.68% on the LoL forum and 83.10%

[4]https://simpletransformers.ai/

on the WoW forum, respectively. On the other hand, there is a discrepancy between Accuracy and macro F1-score on deep neural models due to unbalanced data. However, this difference is not as much as in traditional machine learning models.

Among the models, Toxic-BERT gives the highest results according to the macro F1-score on both forums, which are 82.69% on LoL forum and 83.86% on the WoW forum, respectively according to Table IV.

To handle the imbalanced in the dataset, we apply the SMOTE method to increasing the data in minor label. After using the SMOTE method, on traditional machine learning models, the results by macro F1-score increased significantly. For Logistic Regression, the macro F1-score increases 10.49% and 11.41% on the LoL forum and WoW forums, respectively after using SMOTE. Same as Logistic Regression, the results by macro F1-score on the SVM increases 8.95% and 11.41% on the LoL forum and WoW forums, respectively after using SMOTE. However, SMOTE does not augments the performance of deep neural models on both forums.

TABLE IV: Empirical results of Cyberbullying dataset.

| Model | SMOTE | | NOT SMOTE | |
|---|---|---|---|---|
| | Acc(%) | F1-sc(%) | Acc(%) | F1-sc(%) |
| **League of Legends Forum** | | | | |
| Logistic Regression | 97.02 | 60.19 | 98.81 | 49.70 |
| SVM | 98.20 | 58.65 | 98.81 | 49.70 |
| Text-CNN + fastText | 97.87 | 75.30 | 99.29 | 75.67 |
| Text-CNN+GloVe | 97.06 | 73.19 | 99.37 | 80.68 |
| GRU + fastText | 96.69 | 66.37 | 99.18 | 78.46 |
| GRU + GloVe | 95.54 | 66.50 | 99.32 | 75.53 |
| **Toxic-BERT** | - | - | **99.38** | **82.69** |
| **World of Warcraft Forum** | | | | |
| Logistic Regression | 98.11 | 63.50 | 99.19 | 49.79 |
| SVM | 98.74 | 61.20 | 99.19 | 49.79 |
| Text-CNN + fastText | 97.43 | 73.07 | 99.48 | 75.37 |
| Text-CNN+GloVe | 97.56 | 74.40 | 99.62 | 83.10 |
| GRU + fastText | 96.27 | 66.19 | 99.49 | 74.44 |
| GRU + GloVe | 97.77 | 71.14 | 99.47 | 72.99 |
| **Toxic-BERT** | - | - | **99.58** | **83.86** |

### D. Error analysis

Since the results of the models on the two forums are similar according to Table IV, we analyze the error prediction based on the LoL forum in the Cyberbullying dataset.

For the traditional machine learning model, Figure 5 shows the confusion matrix of the Logistic Regression model before and after the SMOTE technique is applied. As shown in Figure 5, SMOTE increases the predictability of the Logistic Regression model on label 1. However, there is still a lot of data that is mispredicted from label 1 to label 0.

Next, Figure 6 shows the confusion matrix of the Text-CNN model with the GloVe embedding before and after enhancing the data using the SMOTE technique. After the data enhancement, the amount of data correctly predicted in label 1 is increased by 8.2%. However, the number of labels 0 correctly predicted decreased by 2.4%. Therefore, SMOTE does not increase the performance of deep neural models.
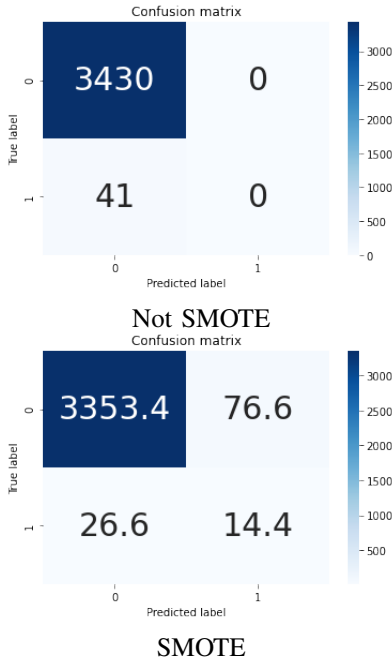
Not SMOTE



SMOTE

Fig. 5: Confusion matrix of the Logistic Regression model on the LoL forum.
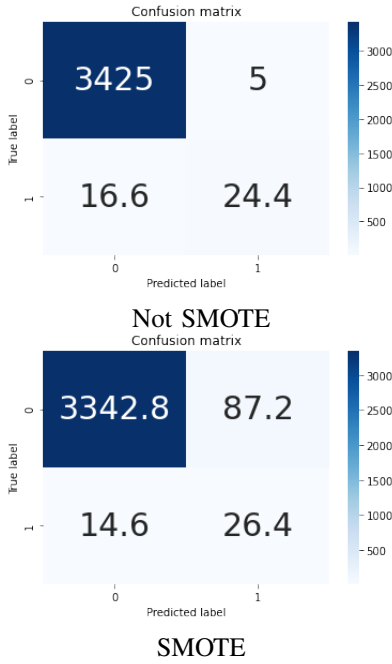


Not SMOTE



SMOTE

Fig. 6: Confusion matrix of the Text-CNN with GloVe word embedding on the LoL forum.

Besides, Figure 7 shows the confusion matrix of the Text-CNN model on two word embeddings including GloVe and fastText without using SMOTE technique. As shown in Figure 7, the predictive accuracy on label 1 of the Text CNN model on GloVe word embedding is better than fastText word embedding. For label 0, the prediction results on both models are similar. Therefore, the performance of the Text-CNN model

with the GloVe word embedding is better than fastText word embedding.
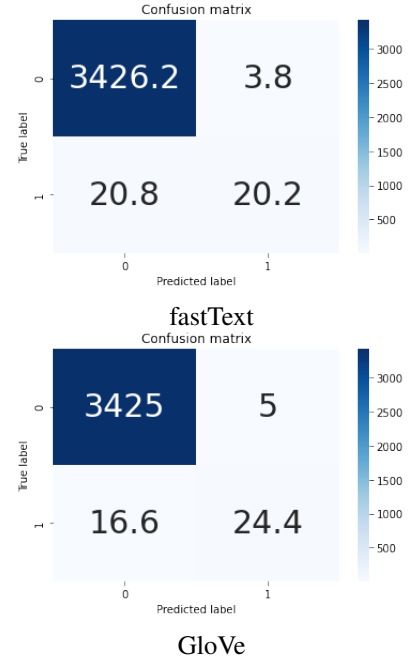


fastText



GloVe

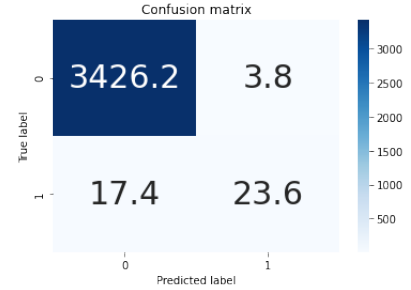Fig. 7: Confusion matrix of the Text-CNN model on the LoL forum (without SMOTE).



Fig. 8: Confusion matrix of the Toxic-BERT model on the LoL forum.

Inaddition, Figure 8 shows the confusion matrix of the Toxic BERT model. According to confusion matrix in Figure 8, the Toxic-BERT model has a higher number of correct predictions on label 1 than the number of incorrect predictions. Comparing with other models, the Toxic-BERT model has a higher total percentage of correct prediction on label 0 and label 1 than other models. Therefore, Toxic-BERT has the highest accuracy on the dataset, even without the data enhancement method.

Finally, according to Table V, many comments of players on the forum use acronyms such as "plz" (roughly meaning: please) in comment No #1, or "u" (roughly meaning: you) in comment No #4. In addition, the syntax and grammar in player comments are often inaccurate compared to formal texts, for example, the word "previou" (roughly meaning: previous) in comment No #3, and "minut" (roughly meaning: minute) in

comment No #4. To improve the predictability of the model, it is necessary to apply lexicon based approaches to normalize the comments into formal texts.

TABLE V: Some predicted results are incorrect in Cyberbullying dataset.

| No. | Input | True | Predicted |
|---|---|---|---|
| 1 | tri argu common sens girl that problem sexist bigot plz never show face realiti kthnx vote oblivion pl x. | 1 | 0 |
| 2 | stop give account info idiot absolut true still think account deserv rune back. | 0 | 1 |
| 3 | refund get kill wait oh idiot rescind previou statement trust memori. | 0 | 1 |
| 4 | post extra two minut travel much stupid u think min travel. | 1 | 0 |
| 5 | post well hug piss stain guy guess happen smell like piss make stupid guy smell annnd sinc hug straight suggest go take shower unless like smell rub | 0 | 1 |
| 6 | leav guy troll wont play stupid card fun get attent world wide ignor troll idiot. | 0 | 1 |

## VI. CONCLUSION

In this paper, we introduce a solution to identify offensive comments on game forums by machine learning approaches using text classification models such as Logistic Regression, SVM, Text-CNN, GRU, and Toxic-BERT models. The results are quite optimistic. The Toxic-BERT model got the best results on the Cyberbullying dataset. The weakness of the Cyberbullying dataset is the imbalance between label 1 and label 0, thus leading to too much wrong prediction of label 1. To solve this problem, we used SMOTE for traditional machine learning models and deep neural models to improve the data imbalance, however, results do not improved significantly on deep neural models. From error analysis, we found that the incorrect prediction caused by the informal texts used by users on game forums such as acronyms, slangs, and wrong syntax sentences. Therefore, it is necessary to normalize the comments of users to increase the performance of classification models.

In future, we are going to apply the Easy Data Augmentation techniques (EDA) [19] to enhance data on minor label, and implement a dictionary to standardize irregular words. The Hurlex [20] and the abusive lexicon provided by Wiegand et al. [21] are two potential abusive lexicon sets used to increase the performance of classification models. Besides, based on the results obtained in this paper, we plan to build a module to automatically detect offensive comments on game forums in order to help moderators for keeping the clean and friendly space for discussion among game players.

## REFERENCES

[1] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," in *Proceedings of the 26th International Conference on World Wide Web*, 2017.

[2] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2017.

[3] M. Ribeiro, P. Calais, Y. Santos, V. Almeida, and W. Meira Jr, "Characterizing and detecting hateful users on twitter," in *Proceedings of the International AAAI Conference on Web and Social Media*, 2018.

[4] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti, "SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter," in *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019.

[5] Y.-L. Chung, E. Kuzmenko, S. S. Tekiroglu, and M. Guerini, "CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019.

[6] U. Bretschneider and R. Peters, "Detecting cyberbullying in online communities," 2016.

[7] W. Fan and M. D. Gordon, "The power of social media analytics," *Communications of the ACM*, 2014.

[8] C. C. Aggarwal and C. Zhai, "A survey of text classification algorithms," in *Mining text data*. Springer, 2012.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, 2002.

[10] V. Character-Level, "End-to-end recurrent neural network models for vietnamese named entity recognition: Word-level," in *Computational Linguistics: 15th International Conference of the Pacific Association for Computational Linguistics, PACLING 2017, Yangon, Myanmar, August 16–18, 2017, Revised Selected Papers*. Springer, 2018.

[11] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014.

[12] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018.

[13] K. Nigam, J. Lafferty, and A. McCallum, "Using maximum entropy for text classification," in *IJCAI-99 workshop on machine learning for information filtering*. Stockholom, Sweden, 1999.

[14] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, 2011.

[15] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751.

[16] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019.

[18] L. Hanu and Unitary team, "Detoxify," Github. https://github.com/unitaryai/detoxify, 2020.

[19] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6382–6388.

[20] E. Bassignana, V. Basile, and V. Patti, "Hurtlex: A multilingual lexicon of words to hurt," in *5th Italian Conference on Computational Linguistics, CLiC-it 2018*. CEUR-WS, 2018.

[21] M. Wiegand, J. Ruppenhofer, A. Schmidt, and C. Greenberg, "Inducing a lexicon of abusive words – a feature-based approach," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018.