Dataset Augmentation for Counteracting Bias in Toxic Comment Classification

Senhao Cheng *

Department of Control Science and Engineering, Zhejiang University, Hangzhou, China

* Corresponding Author Email: 3200102892@zju.edu.com

Abstract. Toxic comments are a prevalent issue on online social media and networking platforms. These comments contain offensive, malicious, hate speech, or other harmful content that negatively impacts audiences and communities. Effectively detecting and categorizing toxic comments is essential for maintaining order in the online environ ment, protecting user safety, and enhancing user experience. This is despite the fact that researchers and companies have developed various models to recognize toxicity in online chats and comments, achieving some success. However, many of the currently used models incorrectly classify non-toxic comments that contain certain identity terms as potentially toxic. This misclassification hinders the ability to accurately identify categorized comments. In this paper, the detection and classification of toxic comments were implemented using Term Frequency-Inverse Document Frequency (TF-IDF) and machine learning techniques. Additionally, two dataset-specific optimizations were proposed to mitigate the impact of bias on text classification by expanding the number of datasets. Comparative analysis of bias evaluation metrics demonstrates that this approach can effectively mitigate bias while maintaining the accuracy of the original model as much as possible.

Keywords: Toxic Comments, Bias, CTRL, Machine Learning.

1. Introduction

When considering the realm of online communication and social media platforms, the prevalence of toxic comments has emerged as a significant concern. These toxic comments have the potential to include insults, discriminatory language, hate speech, and other forms of inappropriate content that have a detrimental effect on the user experience and can even result in psychological harm. To tackle this issue, a great number of efforts have been made to the development of automated tools, such as toxic comment classifiers, with the aim of identifying and eliminating these detrimental comments [1].

The issue of classifying toxic comments involves the utilization of machine learning and natural language processing methodologies to automatically categorize comment texts as either "toxic" or "non-toxic" [2]. Such classifiers have the potential to aid social media platforms in efficiently detecting and resolving toxic comments, thereby safeguarding users from instances of harassment and abuse. Despite notable progress in the development of automated classifiers for identifying toxic comments, there remain inherent challenges in the classification process, specifically pertaining to the presence of bias [3].

Bias is a significant consideration in the categorization of malicious comments. Bias may potentially exist within the training data, and if the training data includes a substantial number of biased statements pertaining to a specific group (such as blacks, gays, etc.), it can result in the classifier demonstrating unfair behavior towards those particular groups or topics [4]. This can lead to the misclassification of ordinary statements from that group as malicious comments. For instance, the statement "I identify as a black woman" may be erroneously categorized as a toxic comment due to the mention of the individual's racial identity. Bias can also arise from the imbalanced distribution of the training sample, such as when the majority of toxic comments in the training set are characterized by brevity and single-word expressions. These biases have the potential to result in unjust content filtering and censorship, thereby restricting the fundamental principles of free speech and equitable treatment.

Therefore, it is imperative to address bias as a critical objective in the research pertaining to the classification of toxic comments. Researchers have extensively investigated different strategies to alleviate the influence of bias. These strategies include utilizing diverse training datasets, fine-tuning algorithm and model parameters, and implementing post-processing techniques to rectify bias in classification outcomes. However, the issue of bias continues to be a multifaceted and demanding challenge, necessitating continuous research and endeavors to develop more equitable and precise classifiers for toxic comments.

The objective of this study is to decrease the occurrence of toxic comments being misclassified as non-toxic by addressing the issue of bias in text categorization algorithms. The main contributions of this study are summarized below:

1) Statistical analysis and training of the dataset, utilizing techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) and machine learning, revealed that the uneven distribution of data containing identity-specific terms was a significant factor contributing to unintended bias in the classification of toxic comments.

2) Mitigating the effect of bias on toxic comment categorization by artificially augmenting the dataset. Two forms are proposed: one involves introducing labeled data that contains identity-specific information, and the other involves introducing data generated by Conditional Transformer Language Model (CTRL) that also contains identity-specific information.

3) Utilize a metric that specifically evaluates model bias and demonstrate, through a before-and-after comparison of this metric, that the proposed approach reduces unintended bias while maintaining the overall quality of the model.

2. Method

2.1. Dataset Preparation

This study examines a dataset that Jigsaw published for the Kaggle Toxic Comment Classification Challenge [5]. The dataset used in this study consists of 19,470 training samples and 7,537 test samples. Each training sample represents a comment that is evaluated for the presence of the following six attributes: toxic, severe toxic, obscene, threat, insult, and identity hate. These attributes are binary, with a value of 0 indicating the absence of the attribute and a value of 1 indicating the presence of the attribute. A toxic comment is typically harsh, offensive, or illogical, and it might prevent important conversations from happening or deter other people from voicing their ideas [6].

In the training data, identity-related words were affected by false positive bias, and these words were used disproportionately in toxic comments compared to overall comments. For example, the word "black" has a probability of 0.77% in toxic comments, compared to 0.06% in overall comments. Similarly, the word "gay" appeared in 0.52% of the toxic comments, compared to 0.05% of the overall comments. This disproportionate frequency of occurrence suggests that identity-related words are more likely to be targeted in toxic comments. Table 1 illustrates the difference between the probability of seeing a given identity in a toxic comment and its overall probability.

Table 1. Frequency of identity terms in toxic and all comments

Term	Toxic	All
black	0.77%	0.06%
gay	0.52%	0.05%
fat	0.46%	0.04%
admin	0.17%	0.07%
mother	0.11%	0.02%
mexicans	0.09%	0.01%
transgender	0.003%	0.001%
heterosexual	0.001%	0.001%

The presence of this bias may cause the model to incorrectly associate specific identity words with toxicity, resulting in biased categorization or judgments of specific groups. To minimize this bias, the current study proposes two approaches to improve the fairness and accuracy of the results in Section 2.3

2.2. TF-IDF and Comparison Methods

In toxic comment classification, the first step is to adopt the TF-IDF method. This method uses TF-IDF features as input feature vectors. The TF-IDF features aid the model in identifying words that have a significant impact on the classification decision, thereby improving the accuracy of classification. The TF-IDF vector of each comment text is used as the feature vector for a sample, while the label of the sample indicates whether the comment is toxic or not. This study implemented training using four models including Random Forest Classifier, Linear SVC, Multinomial Naive Bayes, and Logistic Regression.

2.2.1. TF-IDF

TF-IDF is a widely used method for extracting text features in tasks such as text categorization. It evaluates the importance of words by calculating the product of word frequency and inverse document frequency [7]. When addressing toxic comments, TF-IDF can assist in identifying keywords that are commonly found in toxic comments but are less prevalent in other comments. These keywords can subsequently serve as features to distinguish between toxic and non-toxic comments. The formulas of TF-IDF are shown in equation (1), equation (2), equation (3). The closer the score is to 0, the more common the word is.

$$TF(TermFrequency) = \frac{\#of\ repetition\ of\ word\ in\ sentence}{Total\#\ words\ in\ sentence} \tag{1}$$

$$DF(InverseDenseFrequency) = \frac{\#\ of\ sentences}{Total\#\ words\ in\ sentence} \tag{2}$$

$$ScoreMatrix = TF \times IDF \tag{3}$$

2.2.2. Random Forest Classifier

The Random Forest method is an ensemble learning algorithm that performs categorization by constructing multiple decision trees and aggregating their results [8]. Random Forest can effectively handle text features and address category imbalance for the classification of toxic comments by leveraging the integration of decision trees. Random Forests can reduce the risk of overfitting and improve the robustness of classification by randomly selecting a subset of features and samples to construct decision trees.

2.2.3. Liner SVC

Linear SVC is a classification algorithm based on support vector machines, which is also widely used for text classification [9]. Linear support vector machines divide toxic and non-toxic comments by constructing a linear hyperplane. It maximizes the separation of samples between different categories by finding the optimal segmentation hyperplane. Linear Support Vector Machines have better performance in dealing with high-dimensional text features and can handle complex decision boundaries.

2.2.4. Multinomial NB

Multinomial Naive Bayes is a classification algorithm based on Bayes' theorem for features that have a polynomial distribution [10]. When dealing with the classification of toxic comments, Multinomial Plain Bayes assumes that the features are independent of each other and classifies them by calculating posterior probabilities. It estimates the probabilities by counting the frequency of occurrence of different words in the text and uses Bayesian inference for classification.

2.2.5. Logistic Regression

Logistic regression is a classical linear model that is widely used in text categorization [11]. Logistic regression makes classification predictions by weighting and summing input features and applying a logistic function, such as the sigmoid function. In the task of classifying toxic comments, logistic regression can determine whether a comment is toxic or not by learning the weights of the features. It can handle both binary and multiclass classification tasks and has good interpretability for text classification problems.

2.3. Bias Mitigation

To mitigate the problem of bias resulting from data imbalance, this study adopted the following approach for the training dataset. In the dataset, there is a significant imbalance between non-toxic and toxic comments, as well as an uneven distribution of comments containing different identity terms, both within the toxic and non-toxic categories. To address this issue, this study included more data and primarily focused on incorporating non-toxic comments that specifically targeted identity terms with the most imbalanced distribution. The following two main approaches were used:

1. Extracting new data from published articles and websites, such as Wikipedia, can augment the quantity of non-toxic comments.

2. Utilizing the available toxic and non-toxic data and leveraging a pre-trained language model to generate comments that are non-toxic and inclusive of identity-specific terms.

2.3.1. Extracting New Data from Published Articles and Websites

Extracting new data from published articles on “Wikipedia Talk Labels: Toxicity”, the dataset posted on 2017-02-23. This study only collected review data that is non-toxic and includes the specific identifying information required. From the acquired sample, 2,000 comments were selected as a test sample, and 99.3% of these comments were identified as non-toxic.

By collecting these comments and incorporating them into the training dataset, the number of non-toxic comments can be increased. This will ensure that the ratio of toxic and non-toxic comments in the entire dataset aligns with the predetermined distribution.

2.3.2. CTRL

In this study, CTRL was utilized, a pre-trained language model developed by OpenAI [12]. CTRL is specifically designed to generate text based on a given condition. The CTRL model can generate non-toxic comments related to a specific identity term when provided as a conditional input.

The underlying layer of CTRL is also based on the Transformer, specifically utilizing its Encoder component. The fundamental structure of the model's underlying layer remains largely unchanged. The previous model calculates the likelihood of the next word, word n, based on the preceding n-1 words in the word sequence as follows:

$$p(x) = \prod_{i=1}^{n} p(x_i \mid x < i) \tag{4}$$

CTRL added a new condition, c, which represents the control information of the article, such as the type. This condition is now taken into account when calculating the probability. The specific operation is to include the type description before the specific content of each sequence. This ensures that during the calculation of Attention, the type is linked to all the elements in the sequence. The process is as follows:

$$p(x \mid c) = \prod_{i=1}^{n} p(x_i \mid x < i, c) \tag{5}$$

OpenAI utilizes the "gpt-3.5-turbo" model, an enhanced version of the GPT-3 architecture, to generate non-toxic comments that include identity-specific terms. Using the OpenAI API, the identity-specific terms are inputted into the model, and the resulting comments free of drug references are obtained as output. Table 2 shows some of the comments containing specific identities that were generated using CTRL.

Table 2. Some comments that were generated using CTRL. IDENTITY can be 'gay', 'mexicans', 'mother' etc.

Examples	Label
I am a <IDENTITY> person	*Non-Toxic*
<IDENTITY> should be treated equally	*Non-Toxic*
We need to show <IDENTITY> some respect and understanding	*Non-Toxic*
Even though I'm <IDENTITY>, everyone is welcome to come and talk to me	*Non-Toxic*
As a <IDENTITY>, I deeply appreciate the power of faith and the joy it brings. Christian doctrine teaches us love and forgiveness and genuine relationships with others. Faith is not just a personal choice, but a guide to values and lifestyle.	*Non-Toxic*
I am a <IDENTITY> person, I also realize that our strength lies in unity and supporting each other. We need to continue to fight for equal rights for ourselves, while also building partnerships with other communities to advance progress and equality in society.	*Non-Toxic*
Examples	Label

2.4. Evaluation Metrics

In terms of the most basic categorical assessment of toxicity detection, the Area Under Curve (AUC), precision, and F1-scores were calculated on the test set. Predictive comments with toxicity scores greater than or equal to 0.5 were considered positive (toxic), and vice versa. Also, this study used the Generalized Mean of Bias AUCs as a measure exclusively for model reduction bias.

Generalized Mean of Bias AUCs An overall metric is computed from the subgroup AUC using the following formula:

$$M_p(m_s) = \left(\frac{1}{N}\sum_{s=1}^{N} m_p^s\right)^{\frac{1}{p}} \tag{6}$$

Where $\mathbf{M_p}$ is the $\boldsymbol{p}$th power-mean function ($\boldsymbol{p}$ value = -5), $\mathbf{m_s}$ is the bias metric $\boldsymbol{m}$ calculated for subgroup $\boldsymbol{s}$, and $\mathbf{N}$ is the number of identity subgroups [13].

3. Results and Discussion

The dichotomous scores for all baseline models are shown in Table 3. Among the four models, Linear SVC is optimal in terms of performance. Therefore, in the subsequent task of bias mitigation, Linear SVC is chosen as the foundational model for improvement. After incorporating non-toxic comments from Wikipedia and non-toxic comments with specific identities generated by CTRL into the training set, the updated scores of each model are presented in Table 4.

Table 3. Binary classification performance of different models for toxic comments

Model	Generalized Mean of Bias AUCs	AUC	Precision	F1
Linear SVC	0.9151	0.9628	0.90	0.62
Logistic Regression	0.9001	0.9545	0.89	0.64
Multinomial NB	0.8897	0.9275	0.88	0.62
RF Classifier	0.7794	0.8516	0.78	0.61

Table 4. Performance of the Linear SVC model on the task of binary classification of toxic comments after introducing new data in various ways

Model	Generalized Mean of Bias AUCs	AUC	Precision	F1
Original	0.9151	0.9628	0.90	0.62
Add data from Wikipedia	0.9318	0.9597	0.89	0.63
Add data from CTRL	0.9424	0.9478	0.85	0.59

Compared to the performance of the initial Linear SVC model, the two methods resulted in an increase in the model's Generalized Mean of Bias AUCs performance by 1.67% and 2.73%, respectively. This indicates that the methods used in this study are effective for mitigating bias and performance improvement is significant. After incorporating data from Wikipedia, the accuracy of the model remained unchanged. However, the addition of CTRL-generated data resulted in a decrease in all other performance metrics of the model. The reason for analyzing the decrease in performance could be as follows:

1. There is a lack of diversity in CTRL-generated data. CTRL tends to generate samples that are similar to the training data. Lack of diversity in the training data may cause the model to overfit the training set.

2. The quality of the data generated by CTRL is difficult to guarantee. Generated data may contain syntax errors, logical inconsistencies, or may not accurately reflect the real world. These erroneous data may have a negative impact on the model's training, resulting in a decrease in accuracy.

During the processing of the dataset, it was discovered that some of the comments were in the form of pictures made up of symbols, not specific text, and that these comments should have been toxic. However, these comments could not be accurately categorized as toxic because the characters used were non-toxic. In future endeavors, the further study aims to discover a method for faster and more efficient processing of this data type, reducing the need for manual intervention. Meanwhile, the CTRL-generated data were handled carefully to mitigate model bias and uphold the accuracy of the original model.

4. Conclusion

In this task, the binary classification of toxic comments was first implemented using TF-IDF and machine learning. In the data processing process, a bias phenomenon was discovered. To address this, this study proposed two methods to improve the dataset's bias. These methods involve introducing non-toxic comments with specific identity information in different ways. It was shown that both the inclusion of pre-labeled data (from Wikipedia) and the utilization of non-toxic comments generated by CTRL, which included specific identities, enhanced the specialized bias measurement. However, using CTRL-generated data leads to a decrease in the accuracy of the original model. In contrast, only manual methods can be employed to process the graphical information in the data, which is composed of characters.

References

[1] Gorwa R, Binns R, Katzenbach C. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 2020, 7 (1): 2053951719897945.

[2] Georgakopoulos S V, Tasoulis S K, Vrahatis A G, et al. Convolutional neural networks for toxic comment classification. Proceedings of the 10th hellenic conference on artificial intelligence. 2018: 1 - 6.

[3] Liu Y, Han T, Ma S, et al. Summary of ChatGPT-related research and perspective towards the future of large language models. Meta-Radiology, 2023: 100017.

[4] Kohavi R, Wolpert D H. Bias plus variance decomposition for zero-one loss functions. ICML. 1996, 96: 275 - 283.

[5] Kaggle, Jigsaw toxic comment classification challenge, 2017, https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge.

[6] Vaidya A, Mai F, Ning Y. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. Proceedings of the International AAAI Conference on Web and social media. 2020, 14: 683 - 693.

[7] Ramos J. Using tf-idf to determine word relevance in document queries. Proceedings of the first instructional conference on machine learning. 2003, 242 (1): 29 - 48.

[8] Pal M. Random Forest classifier for remote sensing classification. International journal of remote sensing, 2005, 26 (1): 217 - 222.

[9] Awad M, Khanna R, Awad M, et al. Support vector machines for classification. Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers, 2015: 39 - 66.

[10] Rennie J D, Shih L, Teevan J, et al. Tackling the poor assumptions of naive bayes text classifiers. Proceedings of the 20th international conference on machine learning (ICML-03). 2003: 616 - 623.

[11] Genkin A, Lewis D D, Madigan D. Large-scale Bayesian logistic regression for text categorization. technometrics, 2007, 49 (3): 291 - 304.

[12] Keskar N S, McCann B, Varshney L R, et al. Ctrl: A conditional transformer language model for controllable generation. arXiv preprint arXiv: 1909. 05858, 2019.

[13] Raza S, Reji D J, Ding C. Dbias: detecting biases and ensuring fairness in news articles. International Journal of Data Science and Analytics, 2022: 1 - 21.