Toxicity : detection, classification and reduction

Niharika Sojitra, Joy Christian, Dhruvin Donda, Vijay Mago
Department of Computer Science
Lakehead University
Thunder Bay, Canada
{nsojitra, jchrist8, ddonda, vmago}@lakeheadu.ca

Abstract

Due to the increasing popularity of social media platforms these days, there has been a significant increase in the use of online communication channels, resulting in a vast amount of text data where individuals express their opinions on various subjects. However, this has also led to a rise in online abuse and harassment threats, which has become a significant issue to solve with the help of Natural Language Processing. In this research, we aim is to detect, categorize, and convert such comments into neutral or non-toxic ones by using various machine learning and neural network models along with pre-trained language models. In terms of detection and classification, DistillBERT achieved the highest accuracy rate of 94.34%, while our proposed t5-paraDetox showed the best results for detoxification. Additionally, publicly available API has also been created that can convert toxic comments into their civil version .

1 Introduction

The use of the internet for various purposes like news, entertainment and socialization has increased dramatically in the present era, leading to a significant surge in users on platforms that offer these services. The social media has also become a popular medium for communication between individuals. However, this expansion has led to new challenges, such as the spread of fake news, insults, harassment, and other offensive remarks (Mazari et al., 2023). Such toxicity has emerged as a major problem in modern society, with various classifications that are used to describe it, including abusive, hateful, and offensive language. Language toxicity is primarily defined as 'any form of speech that is rude, disrespectful, or unreasonable, and which can cause someone to leave a discussion' (Li et al.). It has become increasingly difficult to identify and eradicate these threats, which are spreading in online communities, making it important for platform providers to address this issue and maintain constructive and inclusive online interaction (Pavlopoulos et al., 2019).

Language toxicity has emerged as one of the most challenging active research topic in Natural Language Processing (NLP) over the past few years. There have been numerous efforts to detect and mitigate this toxicity completely from texts (Dinakar et al., 2021; Hasanuzzaman et al., 2017; Burnap and Williams, 2015; Dadvar et al., 2013; Kwok and Wang, 2013; Wulczyn et al., 2016; Park and Fung, 2017). Initially, businesses and organizations have begun flagging inappropriate comments and blocking users who use abusive language manually in order to prevent authorized and genuine users from being exposed to such a behavior (Gambäck and Sikdar, 2017). With the increase in social media traffic, there is an acute need for automated identification and removal for this harmful textual content, which can be addressed using NLP. There have been ample amount of machine learning and deep learning techniques, along with transformer models that have been implemented till date to tackle this challenge which will be discussed further in following section. However, there is still scope for improvement in terms of results. In this paper, we aim to address this problem by identifying and classifying toxic comments with improved results, and the eventual goal of removing, reducing, or converting them into neutral or non-toxic comments by employing various machine learning, neural nets and language models. Along with this, a publicly available user API is provided which can display civil rephrase for toxic sentence.

The rest of the paper is outlined as follows: Section 2 represents the existing research on datasets and methods for toxicity detection, classification and transformation of toxic sentences into their

The source code for this research is available under: https://github.com/niharika-sojitra/Toxicity-Classification-and-detoxification

civil counterparts. Methods used in this paper are described in Section 3 while section 4 describes the results and analysis for performed methods. Section 5 discusses the research challenges, limitations and future direction for this research and section 6 concludes it.

2 Literature Review

Toxicology detection has received a lot of attention over the recent years. Although, the term "toxic" is in a general sense, it should pointed out that there are many other names for various types of toxic language in literature, such as offensive, abusive, and hateful. Hate speech was first introduced to shine the issue of racial abuse, and its social, political and psychological effects on individuals (Delgado, 1982). Since then, in addition to race there has been numerous other factors that identify hate speech, such as insult or harassment based on ethnicity, origin, gender, sexual orientation, religion, age, caste, and disability to a person or group (Nockleby, 2000; Alkomah and Ma, 2022). Nowadays, the principle source of these toxic comments are online platforms, that have been growing rapidly in the past five years due to ever evolving cyberspace. It has become utterly imperative to censor these comments as doing it manually is too expensive for internet operators and platform providers. But with the recent development in NLP , automated textual toxicity detection has become one of the most demanding research topic (Saroar Jahan and Oussalah, 2021). The literature review has been primarily split into two components based on the purpose of the research.

2.1 Datasets

There are only a few datasets that contain the labelled toxic data in English. Hate Speech and Offensive Language Dataset (Davidson et al., 2017), Civil Comments, Jigsaw Toxic Comments Classification Dataset (PELLER), OLID (Offensive Language Identification Dataset) (Zampieri et al., 2019a), SOLID (Semi-Supervised Offensive Language Identification Dataset) (Rosenthal et al., 2020) and RealToxicityPrompts (Gehman et al., 2020) dataset are the datasets that have been most utilised for various classification tasks. Out of these, Jigsaw is only largest multi-label dataset that is derived from Civil Comments dataset collected from Wikipedia, while Hate Speech and Offensive Language Dataset, OLID, and SOLID each have less than 25000 comments with three classification labels. Other than toxicity detection and classification, only two datasets have been specifically created for the task of toxicity reduction, which has pairs of toxic and civil comments: Parallel detoxification dataset (Dementieva et al., 2021) and ParaDetox dataset (Logacheva et al., 2022).

2.2 Techniques

There are few renowned competitions that helped in advancing this research further (e.g., ALW1[1], TA-COS[2],SemEval-2019[3], 2020[4] and 2021[5] and GermEval[6]). They have shared tasks that are created along with the dataset to get benchmark results on each one.

In SemEval-2019, two tasks were on abusive language detection, Task 5, 'hatEval: Multilingual detection of hate speech against immigrants and women in Twitter' (Basile et al., 2019) and Task 6 was on OLID dataset, 'OffensEval: Identifying and Categorizing Offensive Language in Social Media' (Zampieri et al., 2019b). For Task 5, SVM model with RBF kernel obtained the highest result at macro-averaged F1-score of 0.651, while the rest in top five list used neural networks. BERT based models were used by the top 10 rank holders for task 6, surpassing every other models (Zampieri et al., 2019b; Pavlopoulos et al., 2019).

The "TOXICSPANS dataset", which was obtained from the "Civil Comments dataset" and contained solely toxic posts, was introduced for the 'Toxic Spans Detection' task of SemEval-2021. A random sample of 11,000 posts from this dataset were then manually annotated. The competitors have to determine the parts of a post that make it toxic for this assignment. The cost of hand annotations was significant, and there were a lot of mislabelled predictions on the testing data, thus this dataset does have some limitations (Pavlopoulos et al., 2021; Zhu et al., 2021). It was examined using a variety of algorithms, but the two primary ones that performed the best with accuracy values of 70.83% and 70.77% were HITSZ-HLT submission and S-NLP submission respectively. The benchmark model "SPAN-BERT-SEQ"

[1] https://sites.google.com/site/abusivelanguageworkshop2017/

[2] http://ta-cos.org/

[3] https://alt.qcri.org/semeval2019/index.php?id=tasks

[4] https://alt.qcri.org/semeval2020/index.php?id=tasks

[5] https://semeval.github.io/SemEval2021/tasks

[6] https://goo.gl/uZEerk

on The “TOXICSPANS dataset" was also provided, with a f1-score of 63 (Pavlopoulos et al., 2022). This research was extended further for toxic to civil transformation of comments, in which the authors used two toxic-to-civil transfer models, ‘CAE-T5’ and ‘SED-T5’, both based on the T5 transformer encoder-decoder architecture (Raffel et al., 2020). When comparing the toxicity scores of posts with and without explicit toxicity, they discovered a 14% difference between the two groups. A self supervised learning model called ‘CAE-T5’ on ‘Jigsaw dataset’, achieved notable accuracy rate of 75% on automatic evaluation (Laugier et al., 2021).The quantitative and qualitative analysis demonstrates that, despite the fact that many generated instances still suffer from serious semantic drift, ML systems may be able to help in moderate online discussions to some extent.

Additionally, two novel unsupervised methods were also proposed to eliminate toxicity using large pre-trained neural networks from text on Jigsaw dataset (Dale et al., 2021). The first method, ‘ParaGeDi’ was combination of two ideas: (1) the use of small style-conditional language models to guide the generation process, and (2) the use of paraphrasing models for style transfer. The second technique, "CondBERT" employs BERT to substitute toxic words with their less offensive alternatives. Although ParaGeDi exhibited better performance than all other models, including CondBERT, with an accuracy of 0.81, it is still not entirely dependable since its precision cannot be compared to that of humans, who can generate sentences manually, presumably with a hundred percent accuracy.

In this paper, we tried to extend the current study to obtain better results by using following methodology.

3 Methodology

The primary aim of this study was to detect and classify toxicity, and then convert toxic statements into more polite versions. To achieve this, multiple machine learning and deep learning models were employed to identify and categorize toxicity, and the results were compared and analyzed in depth. Additionally, various detoxification models were tested to find the most effective way to transform toxic comments into civil language. A detailed description of the entire methodology can be found in Figure 1.

3.1 Data Description

This study utilized the "Jigsaw Dataset" that was derived from the ‘Kaggle toxic comment classification challenge’[7]. The dataset includes a vast number of Wikipedia comments that have been labeled as toxic or non-toxic and further classified by human raters into six types of toxic behavior, namely toxic, severe toxic, identity hate, obscene, insult, and threat. The dataset has four files: ‘train.csv’ and ‘test.csv’ containing comments with binary labels for training and testing respectively, ‘sample_submission.csv’ with 15,000 records of all six toxicity levels and their IDs, and ‘test_labels.csv’ containing labels for the test data. The word cloud for the dataset according to each toxicity feature is presented in Figure 2 [8]. All of the six features provided in the dataset were used in this research to train and test the classification models which are depicted below.

Figure 2: Word Cloud for the Jigsaw dataset according to all six features[8].

3.2 Data Preprocessing

The Jigsaw dataset, which is a large multilingual dataset with 1.8 million comments, had only 12% toxic comments, and out of those, only 17,000 comments were in English. This data is highly imbalanced according to toxicity and its labels, with most comments falling under the "Toxic" category and very few in the "Threat" category. This makes it

[7]Toxic Comment Classification Challenge link is available under : https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data

[8]Disclaimer : This data has been extracted from the Jigsaw Dataset and do not represent the opinions of authors, contributors, supervisors, or anyone else connected to this research.

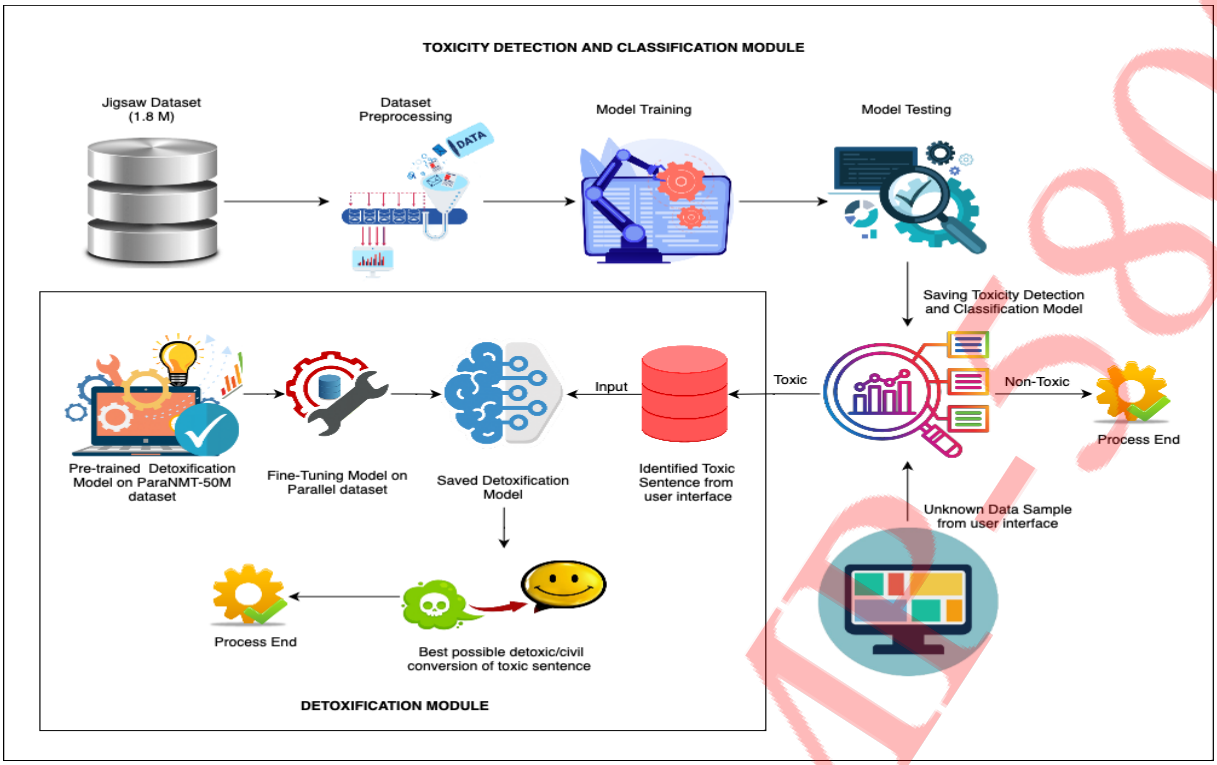

Figure 1: Flow diagram for methodology

challenging to identify each category labels present in unknown test samples. To address this imbalance, the Random over-sampling method was used to add more comments to the toxic data. Figure 3 illustrates the flow for dataset pre-processing.

3.3 Toxicity Detection and Classification

The initial stage of the study, as depicted in Figure 1, involved identifying and categorizing any potential harmful content in the comments. The term "toxicity detection" in this context refers to determining whether a comment is toxic or not, which is indicated by a binary label. If a comment is found to be toxic, it is then classified into multiple categories based on its specific characteristics. To achieve this, various machine learning and neural network algorithms were utilized, which are explained in more detail below.

3.3.1 Machine Learning Approaches

Various baseline machine learning models like Naïve Bayes, Linear Support Vector Classifier, and Logistic Regression, were utilized for multi-label classification. Moreover, a machine learning pipeline was established to automate the workflow, allowing significant data manipulation and transformation for training each classifier. In a multi-label classification scenario, OneVsRestClassifiers,

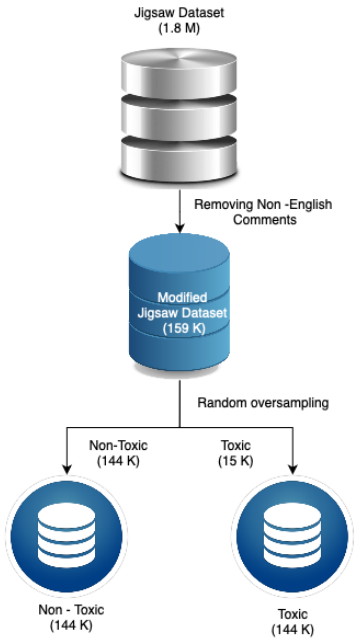

Figure 3: Data Pre-processing using Random Oversampling

a method that accepts a binary mask over multiple labels, was utilized with these models. The prediction outcome in OneVSRestClassifier produced an array of 0s and 1s, indicating the applicable class labels for each input sample row. Label-wise accuracy for each classifier was calculated for comparison.

3.3.2 Neural Net Approaches

Even though the label-wise accuracy for machine learning models is quite high, applying these classifiers to new sets of unknown data gives false predictions as it can not determine if there are multiple labels present in a single comment. Thus it was determined that basic machine learning models might not be the best way to achieve the objective of study. To overcome this deficiency, neural net classifiers were employed, such as 'Recurrent Neural Network' (RNN) that can recognize the sequential characteristics of data and use those patterns to predict the next likely scenario. Jigsaw dataset was used for this model which was divided into 8:2 ratio to train RNN, along with this the Glove Embedding was used to create an embedded vector. Adam optimizer has been used with a learning rate of 1e-3. Another neural net model that was implemented for this task was a sequential 'Bi-direction LSTM model', with pre-processing of the Jigsaw dataset through text vectorization to feed the model. 'BERT' and 'DistilBERT' - pre-trained Bi-directional Transformers were also implemented for Language Understanding and were further fine-tuned to classify the toxic sentences.

In addition, for the testing phase, text filters for slang detection, abbreviation expansion and emoji to text conversion were also carried out to improve the results. For these tasks, publicly available datasets from 'Kaggle' were used which are 'abbreviations.csv'[9], 'full_emoji.csv'[10] and 'twitterSlang.csv'[11] to identify their meaning and replace it with the correct word sequences. After applying those filters with random orders, it was determined that putting slang, abbreviations and emojis in the same order, as order of this does not affect the results. Moreover, we have added a publicly grammar correction library called 'Caribe'[12] to further enhance the grammar and performance.

3.4 Detoxification

Another task that was undertaken in this research is detoxification, rephrasing toxic sentences into their civil versions. To achieve this, various methods have been examined, which includes 'SED-T5', 'CAE-T5', 'CondBert', 'ParaGeDi', and our own model, 't5-paraDetox'. Each model has been modified according to the need of the task and dataset, to further refine the outcomes.

SED-T5 is a supervised encoder decoder based on a text-to-text transformer model that has been fine-tuned on a parallel dataset with toxic comments and their civil version as pairs (Pavlopoulos et al., 2022). CondBERT is another detoxification model, a conditional BERT that replaces toxic words with their civil form and reforms the sentence again by using a transformer (Dale et al., 2021). Even though these two models do show changes in original toxic comments, it failed to remove toxicity completely from it.

Therefore, for detoxification, a fine tune pre-trained ParaGeDi, a generative discriminator guided sequence generation model was used as it showed some promising results (Dale et al., 2021). The authors used heuristic from the original GeDI model proposed by Krause et.al (Krause et al., 2020). To train the model effectively, two types of losses are used in combination: the generative loss L(G) that is commonly used in Language Model (LM) training and the discriminative loss L(d) that aims to increase the distance between different classes. The model is also improved through the use of several inference heuristics that enhance content preservation, improve style transfer accuracy, and increase conditional LM probability.

The most effective outcome was achieved through the use of 't5-paraDetox', which is a modified version of the publicly accessible Huggingface model called 't5-paranmt-detox' (Wieting and Gimpel, 2018). This is large Pre-trained model on ParaNMT-50M[13] dataset and was fine-tuned on the Parallel dataset, where 'toxic_comments' were provided as input and 'civil_comments' were the target. The data was split into a 9:1 ratio for training and testing purposes.

3.5 User Interface

As shown in Figure 1, a user interface was used to get an unknown sample for testing the different

[9] https://www.kaggle.com/datasets/rizdelhi/socialmediaabbrevations

[10] https://www.kaggle.com/datasets/subinium/emojiimage-dataset

[11] https://www.kaggle.com/datasets/gogylogy/twitterslang

[12] https://pypi.org/project/Caribe/

[13] https://www.cs.cmu.edu/~jwieting/

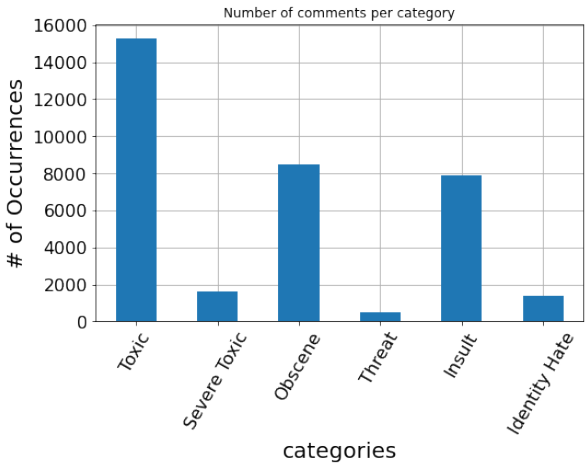

Figure 4: Toxic comments counts according to its labels

models. This user interface was also served as a final deployment for this entire study to produce the results. To create an interface, a publicly available API called Gradio [14] was utilised. This API creates a public link to a web interface that can be shared with anyone and is easily integrated in Python to deploy machine learning model.

4 Results and Analysis

4.1 Dataset Preprocessing

While visualizing the data, it was discovered that 'Jigsaw Dataset' is highly disproportionate according to class as well as its labels. Even-though it is a fairly large dataset (1.8 M total comments), only 230K comments are toxic. Also it can be seen from Figure 4 that labels are distributed unevenly, 'Toxic' class having the most comments at 15294 while only 478 comments are in 'Threat' class. This provides quite a challenge in identifying all labels that are present in toxic data.

Many data balancing approaches, such as SMOTE under-sampling, SMOTE oversampling, Random under-sampling, and Random oversampling, were tried in an effort to balance the dataset. Because of the very limited data, the under-sampling methods were underfitting the model. Whereas SMOTE oversampling added comments based on the proportion of subcategories, increasing the data imbalance further, that nullified the actual purpose. As a result, Random over-sampling method is used, which adds the substantial portion to toxic comments at random which is described in Figure 5. Dataset balancing helped in increasing the classification results of each model with minor

[14] https://3f559185551b5009.gradio.app/

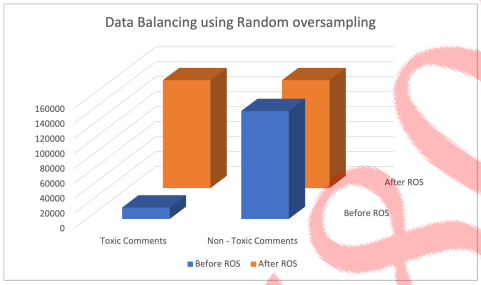

Figure 5: Balancing the data using Random oversampling(ROS)

difference.

4.2 Toxicity detection and classification

This experiment comprises two phases namely toxicity detection and detoxification. Neural nets and basic machine learning models were both used in the first phase to detect and classify the toxic content. Naive Bayes, Logistic Regression, and Support Vector Classification(SVC) are used as fundamental machine learning models that provide label-wise accuracy. The results showed that SVC gave the highest average accuracy of 0.919. Although the machine learning models exhibited high accuracy rates and precision for each label, the recall and F1-scores were relatively low. The previous models were not suitable for the next phase of research as they could only predict one label at a time, which led to incorrect predictions when dealing with unknown test samples with multiple labels.

Therefore, neural networks were utilized for further research, and four models, namely RNN, Bi-LSTM, BERT, and DistilBERT, were implemented. Among these models, DistilBERT achieved the highest Precision,F1-score and an accuracy rate of 0.8497, 0.8318, 94.34%, respectively, while BERT exhibited the highest Recall for classification at 0.8346. It can be verified from these results that language models based on BERT perform better on this particular data. All the results obtained by each model are presented in Table 1.

4.3 Detoxification

Various techniques like SED-T5, CondBert and ParaGeDi were used to detoxify the comments. Our detoxification model, named "t5-paraDetox", was compared with the aforementioned methods. In order to evaluate the outcomes of each detoxification model, two types of similarity scores were computed between the original and the generated

	Model	Precision	Recall	F1 Score	Accuracy
ML Models(Average of Lable-wise Results)	**NB**	0.99	0.11	0.19	0.8986
	LR	0.89	0.52	0.66	0.9181
	SVC	0.84	0.62	0.72	0.9197
Neural Net Models	**RNN**	0.7227	0.6144	0.64	0.8487
	Bi-LSTM	0.7774	0.7343	0.7553	0.7589
	BERT	0.8257	**0.8346**	0.8301	0.9402
	DistilBERT	**0.8497**	0.8164	**0.8318**	**0.9434**

Table 1: Evaluation metric for Machine Learning, Neural nets and Language models for classification

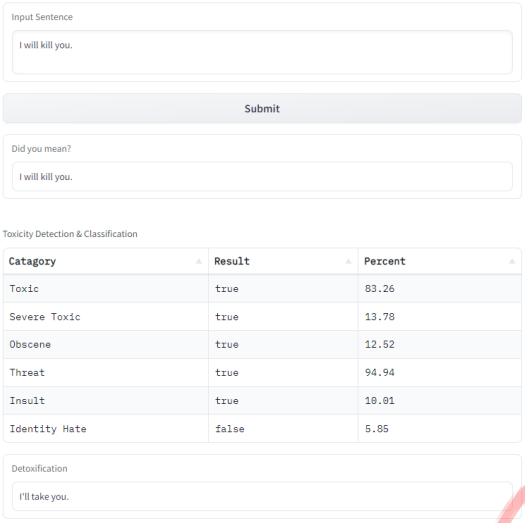

Figure 6: Result for an User Interface when applied toxic sentence.

sentences - namely BLEU and cosine similarity.

During the analysis, it was discovered that when both the BLEU Score and Cosine similarity exceed 50%, it signifies that the sentence's essence is preserved while also making it more polite with minimal omissions. However, extremely high scores indicate an unchanged sentence without any meaningful refinement, while very low scores suggest a complete alteration of the original sentence. Therefore, a threshold value of 0.5 was established, and the results were determined by subtracting the threshold from the original value. Based on our research, the sentence with the lowest positive score is considered the best according to this method.

The results presented in Table 2 indicate that our detoxification model successfully maintained the meaning of the sentence while detoxifying an unidentified sample, as evidenced by the least positive values of Δ BLEU Score and Δ Cosine similarity. However, other detoxification models such as ParaGedi significantly changed the meaning of the sentence, as reflected by their low Δ BLEU Score and Δ Cosine similarity values. The SED-T5 model produced an output that was identical to the input text, as demonstrated by the high values of Δ BLEU Score and Δ Cosine similarity. On the other hand, CondBERT replaced toxic words with blank spaces, which resulted in better Δ BLEU and Δ cosine similarity scores, but some words were missing from the sentences. Additionally, a human-based response was included in the analysis, which had a lower scores compared to the other models due to sentence variation.

Figure 6 shows the result of an User Interface when applied and input toxic text.

5 Limitations and Future work

During the research process, several challenges were encountered. The first issue was the presence of imbalanced data, which made it challenging to accurately identify certain labels even after employing balancing techniques. Another difficulty was the lack of publicly available resources for text detoxification. The only two limited datasets that could be used for model training were not adequate for achieving satisfactory results. Additionally, the process of manually adding new data for detoxification was time-consuming and required significant resources.

The research's future course was influenced by these factors. We had limited time to try out alternative methods for balancing the data, which could have improved the results. However, this aspect can be further explored in future studies for better outcomes. Furthermore, all of the toxic comments in the Jigsaw dataset can be converted to civil ones, making it easier to train detoxification models. Currently, only one toxic sentence can be transformed into a civil sentence using this technique, but multiple sentences can be included as an additional adjustment.

Model	Output Text	Δ BLEU	Δ Cosine Similarity
ParaGEDI	I've been a big fan of shootings in the school.	-0.3976	-0.2743
CondBERT	i am " " " of this school shootings by " people.	0.4788	0.2092
SED_T5	I am tired of this school shootings by black people.	0.4902	0.3975
Human Response	The frequent school shootings are getting to me.	-0.443	-0.1813
t5-paraDetox	I'm tired of this school shootings.	**0.1547**	**0.0803**

Table 2: Evaluation metric for each Detoxification models for toxic input text : "I am tired of this school shootings by black people."

6 Conclusion

In this paper, we have analysed the various detection, classification and mitigation techniques regarding the textual toxicity present in online comments. For this task, we have chosen the Jigsaw dataset which has 1.8M multilingual comments from Wikipedia civil comments. This is highly imbalanced dataset with just 12% of total comments that are toxic. As only English comments were considered as our primary data, we had very less toxic content to train the models. To overcome this, we have applied Random Oversampling as balancing technique. Furthermore, we have experimented on two kind of models, baseline ML models, Neural Net and Language models to detect and classify the toxicity in text. In baseline models, we used Naive Bayes, Logistic Regression and Support Vector Classifier, from which Support Vector Classifier performed the best. Since these models only predicted label-wise results, when presented with unknown sample with multiple labels, the predicted result of the models were incorrect. So to solve this issue, we implemented Neural Net and Language models: Recurrent Neural Network, Bidirectional Long Short Term Memory (BiLSTM), BERT and DistilBERT, from which DistilBERT gave the best performance in classifying the toxic data. In addition to this, we have also run several models in order to purge the toxic comments. For this task, SED-T5', 'CAE-T5', 'CondBert', 'ParaGeDi' and 't5-paraDetox' were used. Each model performed differently to obtain the results, not all of them were successful in eliminating the toxic comment. Our model 't5-paraDetox' was the best model to generate civil versions of the toxic data when presenting unknown sample from user.

Acknowledgements

We are extremely grateful to our supervisor, Dr. Vijay Mago for his invaluable guidance in this research. Additionally, we would like to extend our sincere thanks to the teaching assistants, Andrew Fisher and Akriti Jindal. We would also like to appreciate the support from our class mates who had helped us in this achievement. Lastly, I would like to thank Jade Goodall, a writing assistant at University Library who helped in eradicating errors in writing of this paper.

Author's Contribution

The research project was a collaborative effort involving Niharika Sojitra, Joy Christian, and Dhruvin Donda. We all played a significant role in designing and implementing the research, analyzing the results, and writing the manuscript.

References

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6).

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. 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*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy Internet*, 7(2):223–242.

M Dadvar, D Trieschnigg, R Ordelman, and F De Jong. 2013. Improving cyberbullying detec- tion with user context. In *European Conference on Information Retrieval*, pages 693–696.

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models. In *Proceedings*

of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7979–7996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Richard Delgado. 1982. Words that wound: A tort action for racial insults, epithets, and Name-Calling, 17 harv. *L. Rev*, 133.

Daryna Dementieva, Sergey Ustyantsev, David Dale, Olga Kozlova, Nikita Semenov, Alexander Panchenko, and Varvara Logacheva. 2021. Crowdsourcing of parallel corpora: the case of style transfer for detoxification. In *Proceedings of the 2nd Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale co-located with 47th International Conference on Very Large Data Bases (VLDB 2021 (https://vldb.org/2021/))*, pages 35–49, Copenhagen, Denmark. CEUR Workshop Proceedings.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2021. Modeling the detection of textual cyberbullying. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(3):11–17.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90, Vancouver, BC, Canada. Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Mohammed Hasanuzzaman, Gaël Dias, and Andy Way. 2017. Demographic word embeddings for racism detection on Twitter. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 926–936, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq R. Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. Gedi: Generative discriminator guided sequence generation. *CoRR*, abs/2009.06367.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. *Proc. Conf. AAAI Artif. Intell.*, 27(1):1621–1622.

Léo Laugier, John Pavlopoulos, Jeffrey Sorensen, and Lucas Dixon. 2021. Civil rephrases of toxic texts with self-supervised transformers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1442–1461, Online. Association for Computational Linguistics.

Hao Li, Weiquan Mao, and Hanyuan Liu. Toxic comment detection and classification.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. ParaDetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland. Association for Computational Linguistics.

Ahmed Cherif Mazari, Nesrine Boudoukhani, and Abdelhamid Djeffal. 2023. BERT-based ensemble learning for multi-aspect hate speech detection. *Cluster Comput.*

J T Nockleby. 2000. Hate speech. *Encyclopedia of the American constitution*, 3:1277–1279.

J H Park and P Fung. 2017. One-step and two-step classification for abusive language detection on twitter. In *1st Workshop on Abusive Language Online*, pages 41–45.

John Pavlopoulos, Leo Laugier, Alexandros Xenos, Jeffrey Sorensen, and Ion Androutsopoulos. 2022. From the detection of toxic spans in online discussions to the analysis of toxic-to-civil transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3721–3734, Dublin, Ireland. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

John Pavlopoulos, Nithum Thain, Lucas Dixon, and Ion Androutsopoulos. 2019. ConvAI at SemEval-2019 task 6: Offensive language identification and categorization with perspective and BERT. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 571–576, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

JULIÁN PELLER. Jigsaw toxic comment classification dataset.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2020. Solid: A large-scale semi-supervised dataset for offensive language identification. In *Findings*.

Md Saroar Jahan and Mourad Oussalah. 2021. A systematic review of hate speech automatic detection using natural language processing. *arXiv e-prints*, pages arXiv–2106.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2016. Ex machina: Personal attacks seen at scale.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. *CoRR*, abs/1902.09666.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Qinglin Zhu, Zijie Lin, Yice Zhang, Jingyi Sun, Xiang Li, Qihui Lin, Yixue Dang, and Ruifeng Xu. 2021. HITSZ-HLT at SemEval-2021 task 5: Ensemble sequence labeling and span boundary detection for toxic span detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 521–526, Online. Association for Computational Linguistics.