Research on the Detection and Rephrasing of Toxic Text Based on Large-scale Pre-training Language Models

Shih-Hung Wu*[1][0000−0002−1769−0613*], TSAI Tsung Hsun[2], and Ping-Hsuan Lee[3]

Department of Computer Science and Information Engineering, Chaoyang University
shwu@cyut.edu.tw*[1] s11227607@gm.cyut.edu.tw[2] s10927057@gm.cyut.edu.tw[3]

Abstract. Offensive speech on the Internet is a harm to the people who receive it, and how to balance between freedom of speech and reducing the dissemination of malicious language is a direction that natural language processing can strive for. While using the large-scale pre-trained language models available to detect malicious language, we also tried to keep the constructive comments in the speech and rewrite them into a non-offensive narrative. This research is based on the results of traditional NLP problems such as sentiment analysis, satirical detection, and automatic assessment of conversational language quality. The goal of the system is to detect and rewrite poor quality texts on the Internet, such that we can avoid spreading inappropriate speech and not simply blocking them. Using the public available datasets, we test the toxic text detection ability of various deep learning language models on the corpus, and the generated rephrasing text is also tested with two different language models.

Keywords: harmful language detection · toxic text rephrasing · large-scale language model.

1 Introduction

In today's digital society, detecting and rewriting harmful speech is an important task. This dissertation outline aims to explore how to effectively identify and modify harmful speech on the Internet to promote a healthier communication environment. First, the definition of harmful speech and its impact on individuals and society will be introduced. Current technologies and methods for detecting harmful speech, including natural language processing (NLP) and machine learning algorithms, will then be analyzed. This article will explore the applications and challenges of these technologies in identifying different types of harmful speech, such as hate speech, cyberbullying, and fake news. Going further, we will evaluate the effectiveness of these technologies in different contexts and cultural contexts, and explore how to overcome the challenges posed by linguistic diversity. In addition, the paper will discuss how to balance the boundaries between freedom of expression and cybersecurity, and raise relevant

legal and ethical issues. In the section on rephrasing harmful speech, we will explore how harmful content can be automatically modified by algorithms while preserving the intent and emotion of the original message, which involves complex semantic understanding and sentiment analysis. Finally, the paper will look forward to future research directions, including how to improve the accuracy and sensitivity of detection systems, and how to educate netizens to identify and counter harmful speech.

The goal of this dissertation is to provide a comprehensive framework for the development and evaluation of systems for detecting and rewriting harmful speech, and to provide practical guidance and recommendations for researchers and practitioners in related fields. Through this research, we hope to contribute to the healthy development of the network environment. This is not only a technical challenge, but also part of social responsibility and ethical practices. Technology Review In the current digital age, research methods to detect and rewrite harmful speech have become even more important. This type of research often involves machine learning and natural language processing techniques to identify and modify text that may be offensive, discriminatory, or otherwise harmful. An effective detection system needs to be able to understand the nuances and context of language in order to accurately identify harmful content. In addition, the rewriting system should be able to retain the intent of the original information while removing any harmful elements. When researchers develop these systems, they use large datasets to train and test their models. These datasets may include text collected from social media platforms, forums, or other online communities. To improve the accuracy of detection systems, researchers use a variety of machine learning algorithms, such as support vector machines (SVMs), decision trees, or deep neural networks. These algorithms can learn patterns in the text and identify harmful speech based on those patterns. In addition to traditional text analysis methods, the researchers also explored the use of semantic analysis to improve the accuracy of detection. Semantic analysis involves understanding the meaning of words and phrases and their use in a particular context. This can help the system better understand metaphors, sarcasm, and other rhetorical devices that may be used in harmful speech.

Rewriting harmful speech is a more challenging task, as it requires the system not only to identify harmful content, but also to be able to generate new, harmless text to replace the original text. This often requires a sophisticated generative model that is able to understand the intent of the original text and produce the corresponding alternative expressions. These models may make use of transformational neural networks, such as GPT (generative pre-trained transformers) or BERT (bidirectional encoder representation transformers), which are important advances in the field of natural language processing in recent years. To further improve the effectiveness of these systems, researchers are also exploring ways to combine multiple technologies. For example, they might combine a machine learning algorithm with a rule-based system to take advantage of both.

Rule-based systems can identify harmful speech based on predefined rules, while machine learning algorithms can learn patterns in text, which combine to provide more comprehensive detection. In addition, researchers must consider the ethical implications of detecting and rewriting systems. For example, these systems must be very careful when dealing with sensitive topics to avoid misjudgment or over-censorship. They also need to ensure that the system is not unfair to certain groups or individuals because of bias.

Finally, in order for these systems to work effectively in the real world, researchers need to conduct extensive testing and adjustments. This includes testing the system in different text types and contexts, as well as continuously optimizing the system's performance based on feedback. Through these efforts, we can expect more advanced tools to help us create a safer and more inclusive online environment in the future. To ensure that rewriting system-generated text is harmless, researchers and developers have taken a variety of approaches. First, they design algorithms to identify and filter out language that is offensive, discriminatory, or otherwise harmful. These algorithms are often based on natural language processing techniques and are able to understand the context and meaning of the text. Second, the rephrasing system is trained to avoid using words and expressions that may cause misunderstanding or offense. In addition, these systems learn from a large number of datasets that contain text in a variety of contexts, allowing the system to better understand language use in different contexts. Further, the rephrasing system undergoes extensive testing to ensure that the text it generates is appropriate for different cultural and social contexts. This includes ongoing monitoring and updating of the system in response to constant language changes and new harmful expressions. To improve the reliability of the rewrite system, developers may use a combination of rule-based methods and machine learning models. Rule-based approaches can identify harmful content based on predefined rules, while machine learning models can learn and identify patterns of harmful speech from large amounts of text. In addition, some systems employ the step of human review, where a human reviewer checks and confirms the rewritten text to ensure its accuracy and harmlessness.

2 Methodology

2.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) [1] has a context-understanding language representation model, and its model structure is based on the Transformer mentioned in a paper published by Google in 2017: Attention Is All You Need [2], which can be divided into two blocksThe main purpose of Encoder is to convert input text into vectors, and Decoder is to convert vectors back into text, that is, a Seq2Seq (Sequence-to-sequence) framework, but the difference with Seq2Seq is that it is improved on RNN (Recurrent Neural Network) (as shown in Figure 3), which will lead to the problem of inability to parallel operation when using RNN, in Attention In the Is All You

Need paper, the self-attention mechanism is proposed with reference to the attention mechanism, and this mechanism solves the problem that could not be parallelized.

2.2 RoBERTa

Robustly Optimized BERT Pretraining Approach (RoBERTa) [3], the authors of RoBERTa spend more training time and increase the training batch, and use longer sequences for training, the biggest difference with BERT is the shielding part, the shielding of BERT is static shielding, and the shielding is used in the step of data processing. The authors of RoBERTa switched to a dynamic shielding method for this part, which is different from static in that the shielding is no longer fixed, and the shielding part will be selected when it is ready to start training, and different regions will be selected for shielding each time it is trained, which makes the model learn better than the original BERT, and learn faster.

2.3 ChatGPT

After the rise of the BERT model, many natural language processing models have successively performed well in various natural language processing tasks. For example, the BERT model, the GPT model [5], the Transformer XL model [6], the XLNet model [7][8][9][10][11][12], and so on. Among them, the GPT model family organized by OpenAI is particularly suitable for dialogue generation. The GPT model is a pre-trained model designed based on the Generative Pre-trained Transformer (GPT) model [13], which learns how to generate conversational text through training. GPT has very strong language comprehension skills, is able to communicate in natural language, and in many cases exhibits human-like intelligence. It can also be used in different applications, including text generation, text classification, and text summarization. We can use the GPT model to generate smooth and reasonable responses. When the GPT model operates, it only outputs one token at a time, and every time a new word is generated, this word will be added to the back of the previously generated word sequence, and this sequence will become the next new input of the model, this mechanism is called auto-regression, and it is also the most special mechanism of the RNN model. The GPT model is designed to use unsupervised pre-trained models to do supervised tasks. Because of the temporal nature of text data, an output sequence can be plotted as the product of a series of conditional probabilities (1):

$$p(\mathbf{z}) = \prod_{i=1}^{n} p(S_n \mid S_1, \ldots, S_{n-1}) \tag{1}$$

Equation (1) means to predict the unknown output $= \{S_k, \ldots, S_k\}$ based on the known input $= \{S_1, S_2, \ldots, S_{n-1}\}$. So the model can be expressed as (output|input, task). In decaNLP [14][13], the proposed MQAN model can unify 10

types of tasks, such as machine translation, semantic analysis, relationship extraction, and natural language reasoning, into one classification task, and there is no need to design a separate model for each subtask. When using the GPT model, if you want to avoid the generation of indecent words, short words (less than 5 words) or a large number of repetitive words in the sentence, you can adjust the model internally, and if the GPT model generates the above problems, let the GPT model be regenerated again. ChatGPT [15] is a dialogue generation model based on GPT-3 (Generative Pretrained Transformer-3), trained by OpenAI. It uses Transformer's neural network architecture to understand the syntax and semantics of the language by learning large amounts of existing text data. In addition to the Transformer architecture, ChatGPT uses a number of advanced technologies, including conversational context and model resuscitation. These techniques can help models better understand different concepts in conversations and generate more human responses. ChatGPT is able to accomplish these tasks by relying on its huge dataset totaling 175 billion data and 45 terabytes of training data, not because it really understands the meaning of these statements, but because it chooses the most likely and most human-preferred way from its past training data. Currently, ChatGPT is a highly advanced language model capable of generating high-quality text and offers many different applications, such as chatbots, voice assistants, and natural language generation systems.

	comment_text	toxic	severe_toxic
6	COCKSUCKER BEFORE YOU PISS AROUND ON MY WORK	1	1
12	Hey... what is it..\n@ | talk .\nWhat is it.....	1	0
16	Bye! \n\nDon't look, come or think of comming ...	1	0
42	You are gay or antisemmitian? \n\nArchangel WH...	1	0
43	FUCK YOUR FILTHY MOTHER IN THE ASS, DRY!	1	0
...	...	...	...
159494	"\n\n our previous conversation \n\nyou fuckin...	1	0
159514	YOU ARE A MISCHIEVIOUS PUBIC HAIR	1	0
159541	Your absurd edits \n\nYour absurd edits on gre...	1	0
159546	"\n\nHey listen don't you ever!!!! Delete my e...	1	0
159554	and i'm going to keep posting the stuff u dele...	1	0

15294 rows × 3 columns

Fig. 1. The dataset examples

3 Dataset

The goal of our research is to protect users from receiving offensive remarks, however, it is almost impossible to know where everyone receives information. We will first develop the system using publicly available training data. The dataset, Toxic Comment Classification Challenge Identify and classify toxic online comments, published on Kaggle is a collection of corpora through the Internet and manually labeled whether it is maliciously offensive or not.To assess the ability of LLM on hate speech detection and rewriting, we use the Toxic Comment Classification Challenge data set.

This collection is from the Toxic Comment Classification Challenge on Kaggle. The Conversation AI team, founded by Jigsaw and Google (both subsidiaries of Alphabet), is developing tools to improve online conversations. The focus of research was on negative online behaviors, such as harmful comments (i.e. responses in a conversation that were rude, disrespectful, or otherwise likely to take someone away from the discussion). So far, they've built a range of publicly available models through the Perspective API, including those with harmful speech. But current models are still fallible, and they don't allow users to choose the type of harmful content they're interested in discovering (e.g., some platforms may be comfortable with profanity, but not others). A more correct model is needed. Because the Internet is full of all kinds of harassment and threatening speech. If platform operators fail to effectively facilitate conversations, it will result in many online communities restricting or closing user comments altogether.

In this contest dataset, different types of harmfulness such as threats, obscenity, insults and identity-based hatred will be detected. The text comes from a conversation dataset edited by a Wikipedia talk page. Improvements to the current model are expected to help the discussion become more productive and respectful. Of course, the provider reminds the disclaimer that the dataset contains text that could be considered profane, vulgar, or offensive. Figure 1 shows examples of corpus annotation. The content of the sentence is a reply to a statement taken from an edit of the Wikipedia talk page. It is marked with whether it is harmful speech and whether it is harmful speech of severity. This corpus has a total of 15294 records. Among them, 144277 were harmless and 15,294 were harmful. It can be used as a litmus test for us. In this study, we will use the natural language processing technology of a publicly available pre-trained deep language model to improve the detection ability. At the same time, the use of different data labels as sub-topics also predicts various aspects of inappropriate speech, and also tries to rewrite it into a non-offensive narrative. The distribution of the labels are listed in the following table I.

4 Experimental Results and Discussion

We selected 25,000 harmless replies and 15,294 harmless replies from the original dataset to form our preliminary experimental dataset, and randomly divided the

Table 1. Data Distribution

Dataset	Labels					
	Non-toxic Count	Toxic Count	Non-threat Count	Threat Count	Non-severe Toxic Count	Severe Toxic Count
Training	115,418	12,238	127,252	404	126,382	1,274
Validation	14,421	1,536	15,920	37	15,798	159
Test	14,438	1,520	15,921	37	15,796	162
Percentage	90.42%	9.58%	99.70%	0.30%	99.00%	1.00%
	Non-obscene Count	Obscene Count	Non-insult Count	Insult Count	Non-identity Hate Count	Identity Hate Count
Training	120,922	6,734	121,393	6,263	126,545	1,111
Validation	15,098	859	15,151	806	15,801	156
Test	15,102	856	15,150	808	15,820	138
Percentage	94.71%	5.29%	95.06%	4.94%	99.12%	0.88%

data into 8:1:1, training set, development set, and test set. Fine-tune 3 epochs using the pretrained model. The resulting loss function trend is shown in Figure 1 below. The preliminary experimental results are shown in Table II and III. As we can see in the table, the accuracy is very high for each class, this is due to the fact that the data distribution is no balanced, most of the sentences are not harmful. The detection of toxic text is shown in Fig 2(a) and 2(b).

4.1 Harmfulness Detection by BERT

The detection performance of the BERT model is listed in the following table II.

Table 2. Detection Performance by BERT

	Accuracy	**Precision**	**Recall**	**F1**
toxic	0.9684	0.8437	0.8204	0.8319
threat	0.9977	0.5128	0.5405	0.5263
severe_toxic	0.9902	0.5319	0.3086	0.3906
obscene	0.9815	0.8243	0.8329	0.8286
insult	0.9762	0.7477	0.7995	0.7727
identity_hate	0.9925	0.5841	0.4783	0.5259

4.2 Harmfulness Detection by RoBERTa

The detection performance of the RoBERTa model is listed in the following table III.

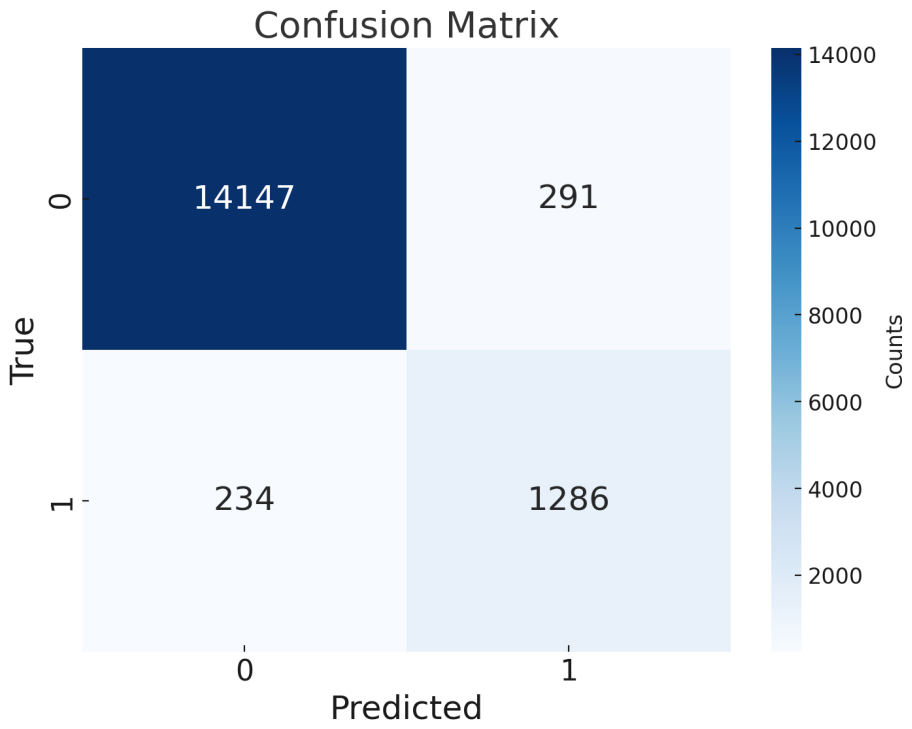

Fig. 2. Confusion matrix of the Toxic label by BERT

Table 3. Detection Performance by RoBERTa

	Accuracy	**Precision**	**Recall**	**F1**
Toxic	0.9671	0.8155	0.8461	0.8305
Threat	0.9972	0.42	0.5676	0.4828
Severe Toxic	0.9907	0.5507	0.4691	0.5067
Obscene	0.9805	0.8093	0.8329	0.821
Insult	0.9757	0.7336	0.8181	0.7736
Identity Hate	0.9916	0.5185	0.4058	0.4553

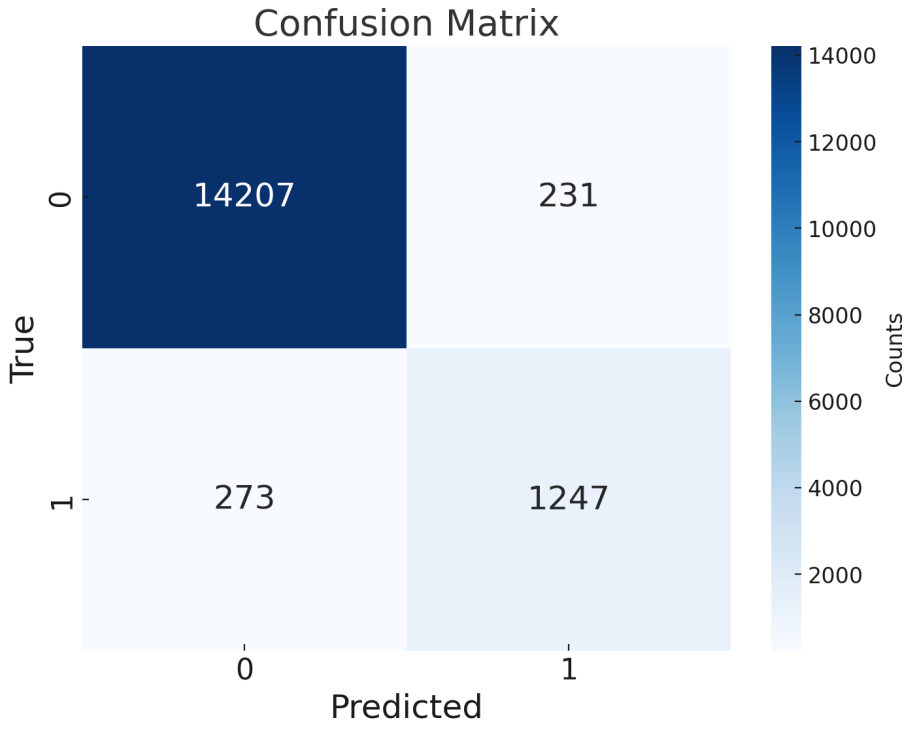

Fig. 3. Confusion matrix of the Toxic label by RoBERTa

4.3 Harmfulness Detection by ChatGPT

Due to the limitation of resource, the test set for ChatGPT contains only 997 sentences. The detection result is shown in the following table IV is just for reference before the system rewrites the text. Since there is no training process, and size of the test set is much smaller, the result should not compare to the results of BERT and RoBERTa.

Table 4. Detection Performance by ChatGPT

	Accuracy	**Precision**	**Recall**	**F1**
toxic	0.75	0.11	0.28	0.16
severe_toxic	0.97	0	0	0
obscene	0.99	0.14	0.14	0.15
threat	0.96	0	0	0
insult	0.8	0.04	0.15	0.07
identity_hate	0.96	0.03	0.2	0.05

4.4 Harmfulness Detection by BERT and RoBERTa after the test sentence is rewriten by ChatGPT

The detection performance of the BERT and RoBERTa models shown in Fig. 3. Almost all the rewritten sentences are not toxic, according to the detector that we developed in the subsection A and B. This means that ChatGPT can generate text without toxic with suitable prompt. This is a quite different case from the text generation with other prompts [16]. Samples of original and rewritten text are shown in Table V. The prompt that we used in our system is listed here:

1. You're asking me to analyze the emotional content of English sentences, marking them as either 1 (indicating presence) or 0 (indicating absence) for categories like *toxic, severe_ toxic, obscene, threat, insult, and identity_ hate*. I'll need to make nuanced judgments, leaning towards 1 for even slight indications and 0 for absence of any such indications.
2. Each sentence should be treated as a whole, without breaking it apart for analysis.
3. Here's a sample template for responses:
   (Sequential numbering starting from 1, the sentence itself isn't displayed here)

```
------------------------------(This dividing line should also be retained)
```
Original Sentence:
Emotions:
1.toxic: 0

2.severe_toxic: 0
3.obscene: 0
4.threat: 0
5.insult: 0
6.identity_hate: 0
`------------------------------(This dividing line should also be retained)`

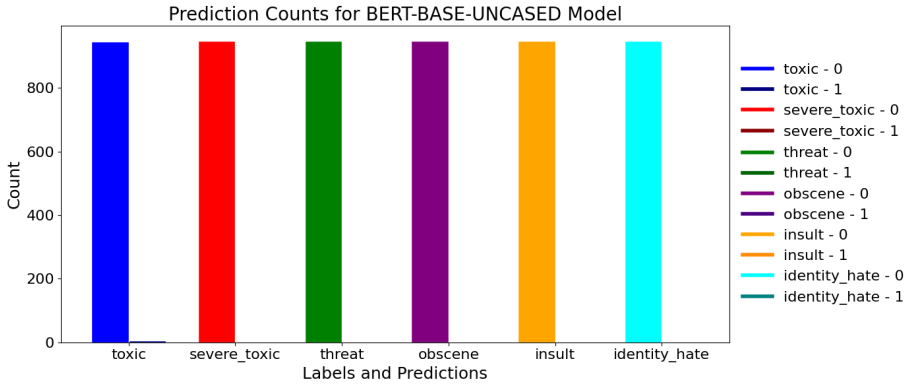

Fig. 4. Detection of all labels by BERT

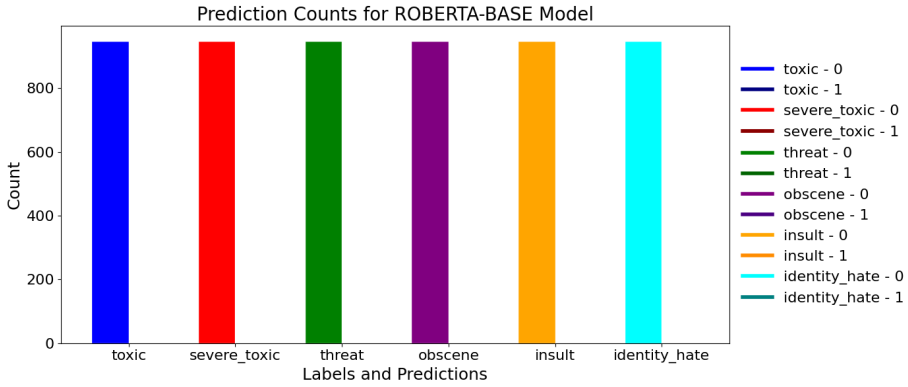

Fig. 5. Detection of all labels by RoBERTa

Here are some actual examples for you to analyze.:

1.Original Sentence:Actually, Jarlaxle is right; that shit was done by me. Alex, you're a cunt. 'Nuff said.

Emotions:
1.toxic: 1
2.severe$_t oxic$: 0
3.$obscene$: 1
4.$threat$: 0
5.$insult$: 1
6.$identity_h ate$: 0

2.Original Sentence:Plus you are all gay for questioning this bull, get a life.
Emotions:
1.toxic: 1
2.severe$_t oxic$: 0
3.$obscene$: 1
4.$threat$: 0
5.$insult$: 1
6.$identity_h ate$: 1

3.Original Sentence:you are a prat, and my revenge will be sweet! hahahahahaahahahahahahahahahahahahaha.
Emotions:
1.toxic: 1
2.severe$_t oxic$: 0
3.$obscene$: 1
4.$threat$: 1
5.$insult$: 1
6.$identity_h ate$: 0

4.Original Sentence:Shut up asshole Shut up you asshole I fucking hate you, Kids CBC IS stupid so why keep reverting my edits?
Emotions:
1.toxic: 1
2.severe$_t oxic$: 1
3.$obscene$: 1
4.$threat$: 0
5.$insult$: 1
6.$identity_h ate$: 0

5.Original Sentence::RSPW Poster is a sick gay freak
Emotions:
1.toxic: 1
2.severe$_t oxic$: 0
3.$obscene$: 1
4.$threat$: 0
5.$insult$: 1

$6.identity_hate : 1$

5 Conclusion and Future Works

In this paper we report our finding on toxic text detection and rephrasing. The detection precision, recall, and F1-score are 0.8155, 0.8461, and 0.8305 respectively. The detection ability is quite high but no perfect. On the other hand, the rephrasing performance is perfect. The sentences which are rewritten by ChatGPT under suitable prompt contain no toxic text. Our system cannot detect any toxic text at all.

In the future, the system can be further extended to be a multilingual system. The rephrasing system's approach to multilingualism and cross-cultural problems is multifaceted and involves advanced natural language processing (NLP) techniques and deep learning. When it comes to cross-cultural issues, the rephrasing system must be able to identify and adapt the expressions and taboos of different cultures. This often requires an in-depth knowledge of specific cultures, as well as a sensitivity to how language is used in those cultures. To improve cross-cultural communication, these systems need to understand their own cultural assumptions, biases, and preferences, and learn how to appropriately adjust between the norms and values of different cultures.

Acknowledgment This study was supported by the National Science and Technology Council under the grant number NSTC 113-2221-E-324-009.

References

1. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019.
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I.: Attention Is All You Need. arXiv:1706.03762v5, 6 Dec 2017.
3. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692, 2019.
4. Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., Hu, G.: Revisiting Pre-trained Models for Chinese Natural Language Processing. arXiv:2004.13922v2, 2 Nov 2020.
5. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Dhariwal, P., Luan, D., Sutskever, I.: Generative pretraining from pixels. arXiv:2006.08437, 2020.
6. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., Salakhutdinov, R.: Transformer-xl: Attentive language models beyond a fixed-length context. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.

7. Burtsev, M. S., Sapunov, G. V.: Revisiting PreTrained Models for Chinese Natural Language Processing, Memory transformer. arXiv:1909.12571, 2019.
8. Parisotto, E., Song, H. F., Rae, J. W., Pascanu, R., Gulcehre, C., Jayakumar, S. M., Jaderberg, M., Kaufman, R. L., Clark, A., Noury, S., Botvinick, M. M., Heess, N., Hadsell, R.: Stabilizing transformers for reinforcement learning. arXiv:1910.06764, 2019.
9. Shen, Z., Zhang, M., Zhao, H., Yi, S., Li, H.: Efficient attention: Attention with linear complexities. arXiv:1812.01243, 2018.
10. Sukhbaatar, S., Grave, E., Lample, G., Jégou, H., Joulin, A.: Augmenting self-attention with persistent memory. arXiv:1907.01470, 2019.
11. Vecoven, N., Ernst, D., Drion, G.: A bio-inspired bistable recurrent cell allows for long-lasting memory. arXiv:2003.06147, 2020.
12. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., Rush, A. M.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 38–45.
13. McCann, B., Keskar, N. S., Xiong, C., Socher, R.: The Natural Language Decathlon: Multitask Learning as Question Answering. arXiv:1806.08730, 2018.
14. OpenAI: ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt/, accessed 30 Jul 2023.
15. Gehman, S., Gururangan, S., Sap, M., Choi, Y., Smith, N. A.: RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 3356–3369, 2020.