A System to Correct Toxic Expression with BERT

Motonobu Yoshida[1], Kazuyuki Matsumoto[2], Minoru Yoshida[2] and Kenji Kita[2]
[1]*Tokushima University, 2-1, Minami-jousangima-cho, Tokushima-shi, Tokushima, Japan*
[2]*Tokushima University, Graduate School of Technology, Industrial and Social Sciences, 2-1, Minami-jousangima-cho, Tokushima-shi, Tokushima, Japan*

Keywords: Toxic Expression, BERT, Classification, Text Correction, Flame War.

Abstract: This paper describes a system for converting posts with toxic expression on social media, such as those containing slander and libel, into less-toxic sentences. In recent years, the number of social media users as well as the cases of online flame wars has been increasing. Therefore, to prevent flaming, we first use a prediction model based on Bidirectional Encoder Representations from Transformers (BERT) to determine whether a sentence is likely to be flamed before it is posted. The highest classification accuracy recorded 82% with the Japanese Spoken Language Field Adaptive BERT Model (Japanese Spoken BERT model) as a pre-trained model. Then, for sentences that are judged to be toxic, we propose a system that uses BERT's masked word prediction to convert toxic expressions into safe expressions, thereby converting them into sentences with mitigated aggression. In addition, the BERTScore is used to quantify whether the meaning of the converted sentence has changed in meaning compared to the original sentence and evaluate whether the modified sentence is safe while preserving the meaning of the original sentence.

1 INTRODUCTION

In recent years, more than 80% of people in Japan are using internet through smartphone and other devices. Therefore, the number of social media users is also increasing every year and subsequently the number of posts by people with low internet literacy and inadvertent posting of content is also increasing. An inadvertent content such as a joke that is meant to be a funny among acquaintances is spread and exposed among many people resulting in online flame wars and slander.

In this paper, we focus on the text posted on Social Media. First, a classifier based on the natural language processing model BERT is used to determine whether a post is a toxic expression or not. Texts judged as toxic are converted to less toxic sentences by replacing the expression judged to be toxic with another safe expression. If this system can be implemented, it will be possible to prevent the Internet flame wars caused by unintentional posting of toxic messages and slandering of others on Social Media.

This paper is organized as follows. In Section 2, related studies are discussed. The method proposed is described in Section 3. In Section 4, we describe the experimental results. Finally, the conclusions and future issues are explained in Section 5.

2 RELATED WORKS

There have been research on flame wars including the use of sentiment analysis (Takahashi and Higashi, 2017; Ozawa et al., 2016) and real-time tweet status counts and other information to make inferences (Steinberger et al., 2017; Iwasaki et al., 2013). Other research has been done on creating datasets for detecting abusive comments(Karayiğit et al., 2021; Omar et al., 2020) and on classifying and detecting hate speech and hateful expressions(Kapli and Ekbal, 2020; Watanabe et al., 2018).

Onishi et al.(Onishi et al., 2015) judges initially whether the input text is likely to be flamed or not. Then, the words that can cause flames are detected with support vector machine (SVM). For the correction of detected words, we used a model learned from the distributed representation of words by word2vec, which was learned from 50 million Japanese tweets collected from Twitter. These collected tweets are pre-processed by removing URLs and hashtags. Furthermore, the predicted replies to the corrected input

Yoshida, M., Matsumoto, K., Yoshida, M. and Kita, K.
A System to Correct Toxic Expression with BERT.
DOI: 10.5220/0011586100003335
In *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2022) - Volume 2: KEOD*, pages 92-97
ISBN: 978-989-758-614-9; ISSN: 2184-3228

text are generated using a neural network-based language model. Table 1 shows an example of the actual process. The study showed good results with an F-measure of 0.74 for flame detection. However, the corrected text may output sentences that do not make sense in Japanese is a possibility. Furthermore, another problem cited is that when the usage of a corpus is similar, words with opposite meanings may also get judged as highly similar.

Yamakoshi et al.(Yamakoshi et al., 2020) also used BERT to create a predictive model that strictly distinguishes legal terms with similar meanings and readings and calibrates them to the correct wording. The study examines three patterns of fine tuning: domain adaptation, undersampling, and classifier aggregation. The results show overall effectiveness in domain adaptation and undersampling. However, the aggregation of classifiers was effective for data with low information content.

With reference to these studies, this study uses a classifier model based on BERT to detect and transform toxic expressions in a flow similar to that of Onishi et al(Onishi et al., 2015). The difference between flaming texts and toxic texts lies in the range of genres specified. In addition to toxic expressions, flaming texts include criminal suggestions and morally offensive content. However, toxic texts are focused on slander and libel.

3 PROPOSED METHOD

The flow of the system procedure of the proposed method is shown in Figure 1. Each of the proposed systems will be explained.

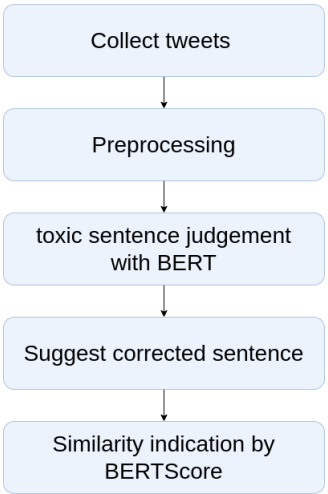

Figure 1: Procedure of the proposed model.

3.1 Collecting Tweets

In this study, we collected tweets with Twitter API[1]. The first step was to collect approximately 300,000 tweets without specifying initial conditions. Then, the tweets were matched with a list of words that had toxic impressions with regular expressions and the corresponding tweets were selected. The 2000 tweets collected in this way were tagged as “safe” ,“toxic” and “spam” by four workers.

3.2 Preprocessing of Tweets

In the pre-processing of tweets, “TweetL[2]” is partially modified for pre-processing. TweetL allows you to:

- Compress multiple spaces into a single space
- removal of Hashtag
- removal of URL
- removal of Pictogram(emoji)
- removal of mension
- Removal of link strings such as images
- Convert all to lowercase
- Converts kana to full-width characters and numbers and ASCII characters to half-width characters.
- Converts numbers to zeros
- Removal of retweets
- Normalization process with neologdn

Of the above functions, the process of converting numbers to zeros is excluded. Also, “!” , “?” , “w” are preprocessed by adding a process that combines them into a single character if there are two or more consecutive characters. Morphological analysis is performed on these preprocessed sentences with MeCab. The dictionary for morphological analysis is “mecab-ipadic-NEologd”. As this dictionary is updated twice a week, it is suitable for analyzing tweets in Social Media, where peculiar expressions, new words, and coined words are frequently used.

3.3 Creating a BERT Classifier

Creating a BERT classifier using the preprocessed tweet data. BERT is a language model proposed by Jacob et al.(Jacob et al., 2018) Using unlabeled data for pre-training and labeled data for fine-tuning has yielded suitable results for many tasks. In this study,

[1]Twitter API, https://developer.twitter.com/en/products/twitter-api

[2]TweetL, https://github.com/deepblue-ts/Tweetl

Table 1: An example of detecting and correcting flaming words system.

	input text	flaming words	corrected text	predicted reply
S1	情弱とキモオタは死ね！ (Die, the low information people and the creepy nerds.)	情弱, キモオタ, 死ね (low information people, creepy nerds, die)	情強とオタクは苦しめ！ (Suffer, information powerhouses and geeks!)	笑わす中もらうわ寝るわ (Sleep in the midst of laughter.)
S2	飲酒運転なう (Drink driving now.)	飲酒運転 (drink driving)	飲酒なう (I'm drinking alcohol.)	何かのことね ww (That is something lol.)

the "Japanese Spoken BERT model" developed by Katsumata et al.(Katsumata and Sakata, 2021) in collaboration with the National Institute for Japanese Language and Linguistics is used as a pre-learning model. This model was trained by adding a spoken Japanese corpus to WikipediaBERT [3], which was developed by the Inui Lab at Tohoku University. To this model, 80% of the 2,000 tweets that were collected and tagged were used as training data for fine tuning as Figure 2. The model is then used as a BERT classifier.

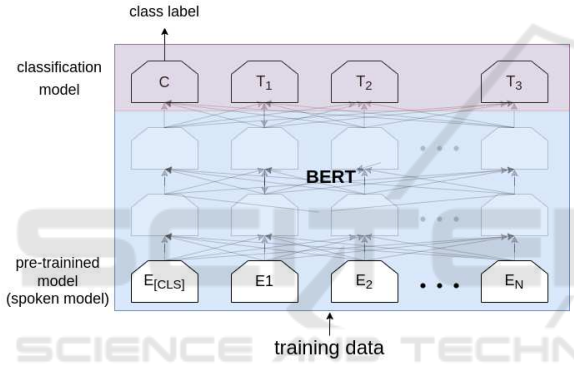

Figure 2: Fine-tuning by BERT.

3.4 MASK Processing Conversion with BERT

The method used for MASK processing conversion by BERT is explained. This process is performed on text that is deemed toxic.

First, we perform the MASK processing transformation with Attention values. The Attention values are taken from the final layer of Transformer (Ashish et al., 2017) in the BERT model and then normalizes so that the maximum value is 1 with min-max normalization. Then, conversion is performed to MASK for words above a certain threshold(up to 0.82). The Figure 3 outlines the process of listing words that are scheduled to be converted. After this process, correction of listed words is performed by MASK predictive transform with BeamSearch.

After this process, the words in the list of toxic

[3]WikipediaBERT, https://github.com/cl-tohoku/bert-japanese

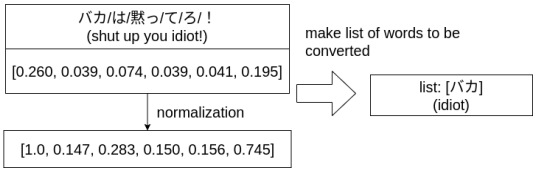

Figure 3: Conversion word extraction with Attention values.

impressions are converted to MASK in the same way. Then, the correction is performed by MASK predictive transformation with BeamSearch.

In practice, conversion prediction is performed using BeamSearch as shown in Figure 4. The beam width is set to 3, the top three appropriate scores as a sentence are saved, and the highest final score is adopted as the modified sentence. As an example of an actual sentence, as shown in Table 2, we do not perform MASK conversions on several parts of the sentence at once but perform predictive conversions one at a time. We try to use this approach multiple times to reduce the number of sentences deemed toxic.

3.5 Similarity Evaluation with BERTScore

BERTScore is a method proposed by Zhang et al.(Zhang et al., 2020) that measures the similarity between sentences with a vector representation for two sentences and the cosine similarity between each token. If one of the texts to be compared is $x = < x_1, \cdots, x_k >$(Reference) and its vector is $<x_1, \cdots, x_k>$, and the other text is $\hat{x} = < \hat{x}_1, \cdots, \hat{x}_k >$(Candidate) and its vector is $<\hat{x}_1, \cdots, \hat{x}_k >$, the reproduction and fit rates are obtained as in equation 1

$$\begin{aligned} R_{BERT} &= \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x_i^T \hat{x}_j} \\ P_{BERT} &= \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x_i^T \hat{x}_j} \end{aligned} \tag{1}$$

From equation1, the F-measure is calculated, and the formula is equation 2.

$$F_{BERT} = 2\frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}} \tag{2}$$

Table 2: MASK conversion example.

マーチ関関同立は f ラン，低学歴でしょうに
(The mid-sized private college group would be the least educated that anyone could get into.)
↓
マーチ関関同立は [MASK]，低学歴でしょうに
(Convert “anyone could get into” to [MASK])
↓
マーチ関関同立は当然，低学歴でしょうに
(The mid-sized private college group, of course, would be less educated.)
↓
マーチ関関同立は当然，[MASK] でしょうに
(Convert “less educated” to [MASK])
↓
マーチ関関同立は当然，不可能でしょうに
(The mid-sized private college group, of course, would be impossible.)

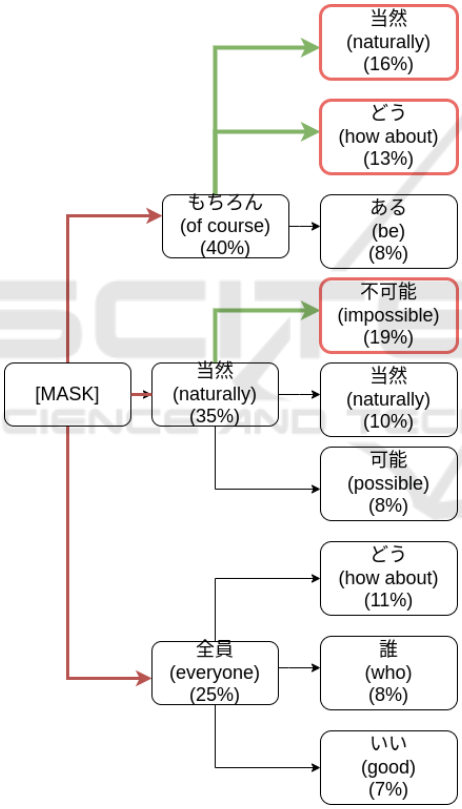

Figure 4: Inference by beamsearch.

4 EXPERIMENTAL RESULTS

In this section, we describe the datasets used and performance of the classifier. The breakdown of the data used in the experiments is given in Table 3.

Table 3: Tweet Breakdown.

Tag	Safe	Toxic	Spam
All	1,064	765	171
training data	638	459	103
validation data	213	153	34
test data	213	153	34

4.1 Classification Accuracy

First, we experimented on the improvement of accuracy when the number of epochs was increased. The number of epochs is number of times the data is trained, and it is important to train an appropriate number of times. The result of the most accuracy was 100 epoch. The accuracy did not change significantly with number of times the training data was turned.

Next, the two pre-trained models, “Wikipedia-BERT” and “Japanese Spoken BERT Model” were fine-tuned with the same training data. By comparing these two models, we examine the importance of the impact of pre-trained model. At this time, to maintain fairness and compare accuracies, both are turned at 100 Epochs. The result is shown in Table 4 and Table 5. The additionally trained “Japanese Spoken BERT Model” improved identification accuracy improved overall.

Table 4: Accuracy of WikipediaBERT.

Label	Recall(%)	Precision(%)	F-measure(%)
safe	79	81	80
toxic	79	75	77
spam	68	77	72

We also compared the accuracy of the existing models, SVM and logistic regression. The Table 6 shows the results of that comparison. Undersampling

Table 5: Accuracy of Japanese Spoken BERT Model.

Label	Recall(%)	Precision(%)	F-measure(%)
safe	87	81	84
toxic	77	81	79
spam	71	89	79

(US) adjusts to the lowest number of data(spam:103), while oversampling (OS) adjusts to 1000 for each label data. The oversampling method used Wordnet to convert nouns and adjectives to create sentences. It can be confirmed that BERT is superior to other models.

Table 6: Compare to other model.

Model	Recall(%)	Precision(%)	F-measure(%)
SVM(US)	41	48	43
SVM(OS)	40	51	43
Logestic(US)	56	66	56
Logestic(OS)	51	60	52
BERT	82	82	82
BERT(OS)	80	82	80

4.2 Result of MASK Conversion by BERT

An example of the actual converted text is shown in Table 7. Compared to the pre-conversion text, the target of attack has become more ambiguous and less aggressive, however, the problem of changing the meaning has had limited improvement. In addition, the object of attention has a word or phrase that does not seem to be toxic.

4.3 Result of BERTScore

The Figure 5 summarizes the distribution of similarity of the 174 cases judged to be toxic in the test data, comparing the text before and after the conversion. In more than 80% of the sentences, the meaning is judged to be similar and in more than 85% of the BERTScore values. However, visual judgments often showed a change in meaning rather than the similarity shown.

In addition, the 174 cases that were determined to be toxic were run through the model again to determine if they were toxic. Then if it was determined to be toxic, word conversion was performed in the same way. The result is shown in the following Figure 6. A third run result is shown in Figure 7. The final number judged to be toxic is shown in Table 8.

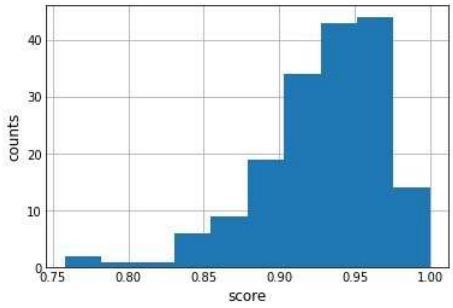

Figure 5: First time distribution of text similarity by BERTScore.

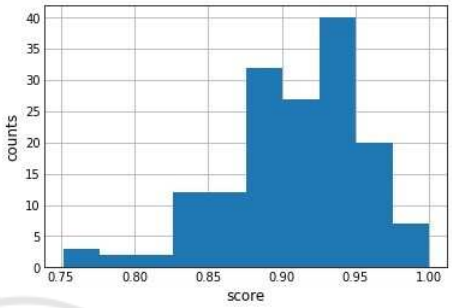

Figure 6: Second time distribution of text similarity by BERTScore.

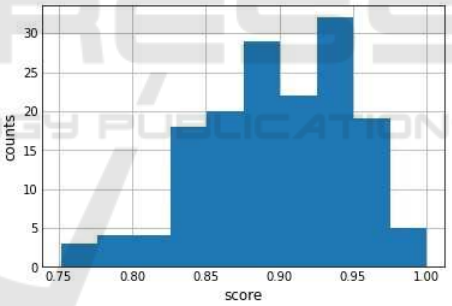

Figure 7: Third time distribution of text similarity by BERTScore.

5 CONCLUSIONS

In this study, we constructed a system for correcting toxic sentences with BERT. Although it is a 3-value classification, the accuracy of discriminating toxic sentences is 79%, with overall accuracy of 82%. In addition, we were able to construct a transformation system to mitigate the aggressiveness of texts judged to be toxic. We also checked whether the BERTScore is effective in examining semantic similarity.

As most of the training data used in this study were tagged as safe, the classification accuracy for safety was considered to have been improved. Therefore, it is important to prepare more data on toxic and

Table 7: Text correction method.

Before	Toxic words	Attention words	After
鉄オタは新左翼と同じくらい内ゲバが好き (Railroad geeks love internal strife as much as the left-wing.)	新左翼 (the left-wing)	鉄オタ・内ゲバ・好き (Railroad geeks, internal strife, love)	鉄人は人間と同じくらい数が多い (Ironmen are as numerous as humans.)
左翼は頭いかれたやつしかおらんのか? (Are the leftists all crazy?)	左翼 (leftists)	頭 (your thinking)	これはもういかれたやつしかおらんのか? (Is this all the crazy ones?)

Table 8: Number of toxic sentence.

times	first	second	third
count	173	157	156

spam contents for comparison. In addition, there are many areas in the MASK conversion process where the intended meaning changes drastically and it is necessary to devise a conversion method that takes part-of-speech into consideration. Although we used BERTScore to calculate the semantic similarity in this case, it may be necessary to compare it with other methods to clarify its reliability.

ACKNOWLEDGEMENTS

This work was supported by the 2022 SCAT Research Grant and JSPS KAKENHI Grant Number JP20K12027, JP21K12141.

REFERENCES

Ashish, V., Noam, S., Niki, P., Jakob, U., Llion, J., Aidan, G., Lukasz, K., and Illia, P. (2017). Attention is all you need. *arXiv:1706.03762*.

Iwasaki, Y., Orihara, R., Sei, Y., Nakagawa, H., Tahara, Y., and Ohsuga, A. (2013). Analysis of flaming and its applications in cgm. *Journal of Japanese Society for Artificial Intelligence*, pp152-160(30(1)).

Jacob, D., Ming-Wei, C., Kenton, L., and Kristina, T. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kapli, P. and Ekbal, A. (2020). A deep neural network based multi-task learning approach to hate speech detection. *Knowledge-Based Systems*, 10.1016/j.knosys.2020.106458(106458).

Karayiğit, H., Aci, C., and Akdagli, A. (2021). Detecting abusive instagram comments in turkish using convolutional neural network and machine learning methods. *Expert Systems with Applications*, 10.1016/j.eswa.2021.114802(17415(114802)).

Katsumata, S. and Sakata, H. (2021). Creation of spoken japanese bert with corpus of spontaneous japanese. *The 27th Annual Conference of the associTheation for Natural Language Pro-cessing*.

Omar, A., Mahmoud, T., and Abd-El-Hafeez, T. (2020). Comparative performance of machine learning and deep learning algorithms for arabic hate speech detection in osns. *Proceedings of the 1st International Conference on Artificial Intelligence and Computer Visions*, 10.1007/978-3-030-44289-7_2(247-257).

Onishi, M., Sawai, Y., Komai, M., Sakai, K., and Shindo, H. (2015). Building a comprehensive system for preventing flaming on twitter. *The 29th Annual Conference of the Japanese Society for Artificial Intelligence*, 3O1-3in.

Ozawa, S., Yoshida, S., Kitazono, J., Sugawara, T., and Haga, T. (2016). A sentiment polarity prediction model using transfer learning and its application to sns flaming event detection. *Proceedings of IEEE Symposium Series on Computational Intelligence*, 10.1109/SSCI.2016.7849868.

Steinberger, J., Brychcin, T., Hercig, T., and Krejzl, P. (2017). Cross-lingual flames detection in news discussions. *Proceedings of International Conference Recent Advances in Natural Language Processing*.

Takahashi, N. and Higashi, Y. (2017). Flaming detection and analysis using emotion analysis on twitter. *The Institute of Electronics, Information and Communication Engineers technical report*, pp135-140.

Watanabe, H., Bouazizi, M., and Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access: Practical Innovations*, 10.1109/ACCESS.2018.2806394(13825-13835).

Yamakoshi, T., Komamizu, T., Ogawa, Y., and Toyama, K. (2020). Japanese legal term correction using bert pretrained model. *The 34th Annual Conference of the Japanese Society for Artificial Intelligence*, 4P3-OS-8-05.

Zhang, T., Kishore, V., Wu, F., Wein-berger, K., and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. *ICLR*.