Exploring the Role of Emoji Semantics in Enhancing Hate Speech Detection on Chinese Social Media

Xuhang Liu [1] **Pu Zhi** [1] **Huiyun Zhu** [1]

Abstract

Detection of hate-speech plays a crucial role in improving digital environment, where complicated and diverse emoticons are often used as the cover to evil intentions. This study delves into the interaction between textual content and emojis to understand how emojis influence the interpretation of language in online discussions. By analyzing extensive datasets of text and emojis, augmented for diversity, we aim to uncover patterns in how emojis alter the connotations and contexts of written communication. We then refine algorithms based on DeBERTa for detecting and categorizing offensive content. Ultimately, our research enhances the accuracy of detecting hate-speech on Chinese media social platforms and aspires to cultivate a digital environment characterized by inclusivity, tolerance, and mutual respect, fostering healthier online communities.

Keywords: Hate Speech, Emoji Semantics, DeBERTa Model

1. Introduction

In the rapidly evolving landscape of online communication, the proliferation of digital platforms has transformed the way individuals interact and express themselves. However, amidst the vast array of textual content exchanged in these virtual spaces, there exists a pressing concern surrounding the proliferation of hate speech and offensive language[1, 2]. Hate speech, defined as language that incites violence or discrimination against individuals or groups based on characteristics such as race, gender, religion, or sexual orientation, poses a significant threat to the inclusivity and safety of online communities.

Central to the challenge of mitigating hate speech online is the nuanced interplay between textual content and emojis. Emojis, often used to convey emotions, add context, or enhance textual communication, have increasingly become a means for individuals to disguise or obscure malicious intentions behind seemingly innocuous symbols. Understanding how emojis modify the connotations and contexts of written communication is thus crucial for effectively detecting and addressing hate speech in online discourse.

This study aims to delve into this complex interaction between textual content and emojis, focusing specifically on Chinese social media platforms, since most of current research focuses on non-Chinese media data[3]. Through the analysis of extensive datasets comprising text and emojis, augmented for diversity and comprehensiveness, we seek to uncover patterns and insights into how emojis influence the interpretation of language in online discussions. By employing advanced natural language processing techniques, particularly leveraging the DeBERTa algorithm[4], we aim to refine hate speech detection models for identifying and categorizing offensive content.

The significance of this research lies not only in its potential to enhance the accuracy of hate speech detection on Chinese social media platforms but also in its broader implications for fostering a more inclusive and respectful digital environment. By cultivating a deeper understanding of the mechanisms through which emojis modify language connotations, we aspire to contribute to the creation of online communities characterized by inclusivity, tolerance, and mutual respect.

The rest of the paper is organized as follows. Section 2 is literature review involving relevant research about hate-speech online. Section 3 gives the current limitations and methodology adopted in our research, which is further presented and validated in Section 4. Finally, Section 5 summarizes the main findings and contributions.

2. Related Work

In recent years, the effective identification and handling of inappropriate content in online texts have become a focal point in the field of natural language processing. Within this domain, researchers have explored various techniques and methodologies to address inappropriate content, including hate speech and offensive language[5]. Several significant research endeavors, particularly those investigating the role of emojis in linguistic communication and how leveraging this information can enhance the detection of inappropriate content, demonstrate potential reward and perspective.

The studies conducted by Francesco Barbieri et al.[6] shed light on the multifaceted nature of emojis, exploring their usage across different languages and contexts, as well as their potential to obscure malicious intent. Additionally, the work of Eckhard Bick [7] highlights the importance of annotating emojis in social media corpora for hate speech research, underscoring their relevance in understanding online discourse.

Moreover, Michele Corazza et al. [8] delved into the challenges posed by emojis in various NLP tasks, such as abusive language detection and multimedia retrieval, emphasizing the need for robust methodologies to handle this novel modality effectively.

These studies collectively demonstrate the growing recognition of emojis as an integral part of online communication and the imperative to incorporate them into computational models for detecting inappropriate content. However, there are still a insufficiency: the lack of Chinese hate-speech detection. Even though Jiang et al. contributed and validated a textual dataset (Weibo) stemming from Chinese social media[3], it took little of Emoji into account. The special influence of emoticons on semantics and the complexity of Chinese expression require extension of Chinese dataset containing Emojis and deeper detection. Besides, Hannah Rose Kirk et al. provided a comprehensive Emoji-based dataset and model that tackling text and Emojis[9], which lays the foundation of this article.

3. Method

3.1. Dataset Construction

To address the limitations identified in our initial study, we constructed a comprehensive dataset (Tab. 1) through following steps:

- **Original Data Preparation**: We obtain the original Chinese dataset sourced from Weibo[3], which contains 8969 hate speech text data along with labels. Another original hate speech with emoji dataset is from [9] called HATEMOJIBUILD, which contains 4728 training data and 593 test data. The HATEMOJIBUILD dataset is used to improve the performance of BERT model on the emoji dataset.
- **Test Data Preparation**: We began by modifying Chinese text data sourced from Weibo[3]. This modification involved adding synonymous emojis, replacing text with relevant emojis, and introducing emoji interference (emojis that are unrelated or antonymous to the text). Corresponding labels were adjusted to reflect these changes, and the modified data *Weibo_emoji* was designated as the test set. The annotators includes all of the group members and we check each other's annotations after finishing them. And we denote remaining Weibo data as *Weibo_text*.
- **Training Data Preparation**: To obtain a Chinese-based HATEMOJIBUILD dataset, we translated the original HATEMOJIBUILD dataset from [9] into Chinese and conducted a thorough review and correction process. This translated dataset *Emoji_build_ch* was used as the training set.

3.2. Experiments based on Weibo dataset

Firstly, referring to the method and dataset from Jiang[3], we used *Weibo_text* to train the DeBERTa model and obtained the result. Then we used the same method by using *Weibo_emoji* and the accuracy (Fig. 2) decreases significantly compared to text data. Then we develop a new method based on [9] to improve the performance of DeBERTa on emoji dataset in section 3.3 and 3.4.

3.3. Model Training

We employed a DeBERTa-Large model, leveraging a four-round training strategy to balance the emphasis between text and emoji information:

- **Round 1 (R1)**: We utilized one-quarter of the *Emoji_build_ch* dataset to initialize the model training.
- **Rounds 2-4 (R2-R4)**: Subsequent training rounds incorporated text data from *Weibo_text* and the remaining *Emoji_build_ch* dataset. This progressive training approach ensured the model's accuracy in predicting pure text while emphasizing the significance of emojis.

3.4. Model Evaluation

The newly trained model was evaluated using the test set constructed through the *Weibo_emoji* described earlier. This evaluation was crucial in assessing the model's performance and its ability to interpret and integrate emoji information within the context of Chinese text data.

4. Validation

The accuracy across the training process is demonstrated in Fig. 1. In the first round, the accuracy on emoji dataset is only around 52%, which in line with our expectations due to the limitation of BERT model on dataset with emoji. From R1 to R2, the accuracy increases from 52% to approximate 80%, verifying our method and the impact of hybrid text. After R2, the accuracy and F1 score has become stable, which around 80% and 75%. Because emoji is usually regarded and tokenized as special text, the text data actually improve model's ability to understand emojis without ignoring main textual expression. Also benefiting

from the advantage of BERT on tackling complicated text, the model reaches stability quickly with high accuracy.

In addition to the training accuracy, we tested the performance of the *Weibo_emoji* dataset on the newly developed model, which derived from the training process R1-R4. As shown in Fig. 2, this model demonstrates superior performance in predicting Chinese text containing emojis compared to the original Weibo data. Specifically, the metrics for *Weibo_emoji* are consistently higher across all measured categories. The accuracy for *Weibo_emoji* reaches approximately 82.3%, significantly outperforming the original *Weibo_emoji_prev* datasets. This illustrates that the model not only maintains its performance on conventional text but also enhances its predictive capability when emojis are involved.

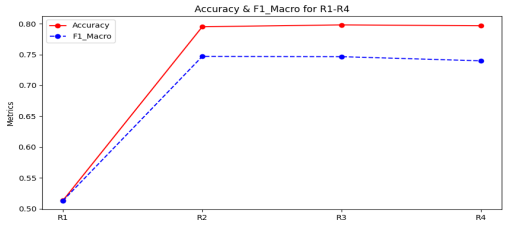

Figure 1. ACCURACY FROM R1 TO R4

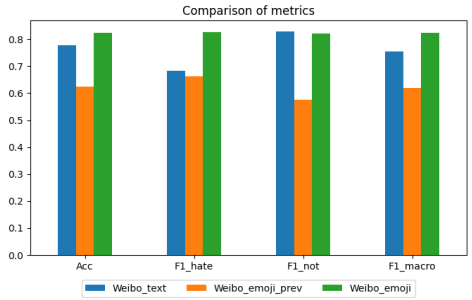

Figure 2. TEST RESULTS

DATA SET	R1	R2	R3	R4	TEST
WEIBO_EMOJI	0	0	0	0	3000
WEIBO_TEXT	0	5968	5968	5968	0
EMOJI_BUILD_CH	1329	1329	1329	1334	0

Table 1. DATASET IN TRAINING AND TEST

5. Conclusion

In this project, we built a new dataset with Chinese text and Emoji, based on which we develop an efficient model based on R1-R4 training process with *Emoji_build_ch* data to identify hatespeech on Chinese social media platform. This DeBERTa-based model and round training strategy show satisfying stability and efficiency. Compared to previous model only performing well on Chinese text dataset, our model ensure high accuracy on sentences with and without Emojis.

References

[1] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings of the 25th international conference on world wide web*, pp. 145–153, 2016.

[2] E. W. Pamungkas, V. Basile, and V. Patti, "A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection," *Information Processing & Management*, vol. 58, no. 4, p. 102544, 2021.

[3] A. Jiang, X. Yang, Y. Liu, and A. Zubiaga, "Swsr: A chinese dataset and lexicon for online sexism detection," *Online Social Networks and Media*, vol. 27, p. 100182, 2022.

[4] Z. Ma, K. Ethayarajh, T. Thrush, S. Jain, L. Wu, R. Jia, C. Potts, A. Williams, and D. Kiela, "Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking," 2021.

[5] P. Fortuna, J. Soler-Company, and L. Wanner, "How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?," *Information Processing & Management*, vol. 58, no. 3, p. 102524, 2021.

[6] F. Barbieri, G. Kruszewski, F. Ronzano, and H. Saggion, "How cosmopolitan are emojis? exploring emojis usage and meaning over different languages with distributional semantics," in *Proceedings of the 24th ACM international conference on Multimedia*, pp. 531–535, 2016.

[7] E. Bick, "Annotating emoticons and emojis in a german-danish social media corpus for hate speech research," *RASK–International journal of language and communication*, vol. 52, pp. 1–20, 2020.

[8] M. Corazza, S. Menini, E. Cabrio, S. Tonelli, and S. Villata, "Hybrid emoji-based masked language models for zero-shot abusive language detection," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 943–949, 2020.

[9] H. R. Kirk, B. Vidgen, P. Röttger, T. Thrush, and S. A. Hale, "Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate," *arXiv preprint arXiv:2108.05921*, 2021.