The 6th International Conference on Interdisciplinary research on Computer science, Psychology, and Education (ICICPE' 2022)
December 26-28, 2022. Pattaya, Thailand.

Efficient Way for Constructing Hate Speech-Counter Narrative Dataset

Seungyoon Lee[1] , Suhyune Son[2], Dahyun Jung[3], Chanjun Park[4], Yuna Hur[5*]
[1] Chung-ang University, [2] Korea University, [3] Hanguk University of Foreign Studies,
[4] Upstage, [5] Human-inspired AI Research

dltmddbs100@cau.ac.kr, dhaabb14@gmail.com, chanjun.park@upstage.ai,
{ssh5131, yj72722}@korea.ac.kr

Abstract. Hate speech that occurs online is one of the major problems in society. To alleviate this problem, recent studies using deep leaning to use a pair of Counter Narrative with the purpose of rehabilitating utterances are being conducted. However, constructing suitable counter speech for each hate speech requires many experts, so it takes a lot of time and money to construct data. Also, generating a corresponding utterance is considered a difficult task. To alleviate this problem, we propose an efficient way to construct hate speech-counter narrative dataset using semantic search with existing dataset. In addition, we present a baseline that can simultaneously solve hate speech classification and counter narrative generation through multi-task learning with pre-trained model.

Keywords: Hate Speech, Counter Narrative, Text Generation, Deep Learning

1 Introduction

In modern society, people are talking and communicating based on anonymity in virtual space. As a result, various social problems have been raised, and among them, Hate speech (HS) is considered to be the main problem affecting people's perception. To solve this problem, various studies using deep learning have been conducted [1]. However, most studies focus only on the detection of HS, so they only make a dataset or model. Since the detection of HS focuses only on whether or not it is HS or what type of expression it is, it is difficult to fundamentally reduce the total amount of HS, so cannot contribute to eliminating or reducing the cause of HS [2]. To alleviate these limitations, the research field of counter narrative (CN) generation against HS has been activated [3].

CN is considered an effective way to induce the improvement of the speaker's erroneous hate perception based on objective facts. CN with reliable evidence can contribute to the mitigation of HS by pointing out the wrong information and suppressing the abuse of the HS by the speaker. However, it is costly to construct a pair of CN for each HS. In general, HS is collected through crawling on SNS, but a

* Corresponding Author

pair of CN suitable for each requires the knowledge and inspection of experts, so it has a disadvantage in that it costs a lot to build. In this work, we present a simple data construction method for generating CN using semantic search using pre-built datasets. Additionally, we use a single model to perform HS classification and CN generation at the same time through multi-task learning in the experiment. Through this, we find that CN generation has a major effect on classification as well.

2 Dataset Construction

The proposed data construction process is as follows. We use the previously constructed HS dataset as an input sentence (implicit HS of [1]) and the explicit HS of the CN dataset (Multitarget CONAN [7]) as candidate sentences do. Through semantic search using Sentence-Transformer [4], candidate sentences similar to the input sentence are searched and the sentence with the highest cosine similarity is extracted. Through this process, a total of 4,947 HS-CN pairs are constructed. An example of the constructed data is shown in Table 1.

Table 1. Examples of matched utterances through semantic search.

HS and CN matched with the input sentence
HS: Islam must be banned from all civilized societies. **Matched:** Islam is incompatible with Western civilization and should be banned. **CN:** This is just another example of the 'us versus them' mentality that pervades society. We as a civilization and as a people must stop giving into fearmongering and scapegoating. It's time to get real.
HS: U know the uk is in europe and europeans are white right? **Matched:** Only White people can be truly 'European'. **CN:** It is a myth that only white people are allowed to be 'European'. Everyone who obtains a citizenship can be called "European".

Additionally, we integrate the non-hate speech by generating a response using BlenderBot [5] so that the model can generate a general response even for the non-hate speech. Finally, for the experiment, we construct data by merging implicit HS-CN, explicit hate Multitarget CONAN, and non-hate speech sentences. The entire data was divided at a ratio of 8:1:1 and used as train, validation, and evaluation set, respectively.

3 Experimental Results

We used the sequence-to-sequence structure BART-base \cite{lewis2019bart} provided by Huggingface [6] for HS classification and CN generation. Multi-task learning is applied by classifying implicit HS, explicit HS, and non-hate expression through the last hidden state of the encoder and generating an appropriate CN or general response through the decoder. In particular, in order to find an appropriate

weight between two tasks, weight rate alpha is applied from 0.1 to 0.9 to find the alpha that shows the most balanced performance. When training the model, the weight with the lowest loss is stored for 10 epochs based on the validation data. When generating, greedy search is applied. BLEU and ROUGE-L, which are generally used in generation, are used as evaluation metrics. For classification, we use accuracy, precision, recall, and F1 score. We use single RTX-8000 GPU for the experiment.

Table 2. HS classification and CN generation performance.

alpha	Classification				Generation			
	P	R	F1	Acc	BELU-1	BLEU-3	BLEU-4	ROUGE-L
0.3	83.11	82.74	82.45	82.93	25.36	11.7	9.76	21.25
0.5	**83.29**	**83.76**	**83.12**	**83.16**	25.47	11.87	9.9	21.48
0.7	83.15	83.68	82.92	82.97	25.74	12.05	10.03	21.73
0.9	79.17	71.6	67.98	77.44	**26.32**	**12.82**	**10.78**	**22.62**

The experimental results are shown in Table 2. As a result of HS classification and CN generation, when alpha is set low, more weight is given to classification, so the classification performance is generally high and the generation performance is low. In contrast, when alpha is high, overall generation performance is high because it focuses on CN generation. However, in the case of alpha is 0.9, which is too high, the accuracy of the model during classification is significantly lowered to 77.44. Also, comparing the case where the alpha is 0.3 and 0.5, the former has lower classification performance than the latter, even though the former gives more weight to the classification. In other words, an excessively low weight given to the generation task shows a result that the classification performance is rather degraded. It means that CN generation has a positive effect on HS classification in the process of multi-task learning of the model.

4 Conclusion

In this paper, we present an efficient method for building HS-CN pairs. The limitation of data construction can be relaxed by using the existing dataset. Generation of CN and the simultaneous execution of HS classification are tested through multi-task learning using a pre-trained model. The classification of HS also has a positive effect on the generation, confirming that the appropriate weight is important. In future studies, we intend to improve the generation and the construction of CN data for various HS.

Acknowledgments. This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques). This research was supported by Basic Science

Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425).

References

1. ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., & Yang, D. (2021). Latent hatred: A benchmark for understanding implicit hate speech. arXiv preprint arXiv:2109.05322.
2. Del Vigna12, F., Cimino23, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook. In Proceedings of the First Italian Conference on Cybersecurity (ITASEC17) (pp. 86-95).
3. Chung, Y. L., Kuzmenko, E., Tekiroglu, S. S., & Guerini, M. (2019). CONAN--COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. arXiv preprint arXiv:1910.03270.
4. Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.
5. Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., ... & Weston, J. (2020). Recipes for building an open-domain chatbot. arXiv preprint arXiv:2004.13637.
6. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.
7. Fanton, M., Bonaldi, H., Tekiroglu, S. S., & Guerini, M. (2021). Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. arXiv preprint arXiv:2107.08720.