Human Interaction and Emerging Technologies (IHIET 2022), Vol. 68, 2022, 241–246
https://doi.org/10.54941/ahfe1002738

Determination of Power Harassment Expressions in Conversational Speech Using Natural Language Processing

Hinari Sasaki, Yujie Li, Hideaki Kawano, and Yoshihisa Nakatoh

Kyushu Institute of Technology, 1-1, Sensui-cho, Tobata-ku, Kitakyusyu-shi, Fukuoka Prefecture, Japan

ABSTRACT

In recent years, the number of power harassment consultations is increasing, and power harassment with ambiguous criteria such as mental aggression is rampant. The purpose of this study is to notify the perpetrator when the likelihood of power harassment is judged to be high based on conversational speech. We attempt to use natural language processing to determine whether the target text constitutes power harassment, based on textual data on past precedents that have led to power harassment. The proposed method determines whether the target text constitutes power harassment or not by calculating the similarity (cos-similarity) between the target text and the text of the precedents and comparing it with a threshold value set through the experiment. The resemblance is calculated from a 768-dimensional feature vector obtained from each text's Bidirectional Encoder Representation from Transformers (BERT). The morphological analyzer is Juman++ and the BERT Japanese Pre-trained Model is used as a pre-trained model. We used two types of surveys to determine thresholds and assess accuracy. In the experiment, we determine the threshold according to the questionnaire results and obtain a high discrimination rate, which shows that our method is effective.

Keywords: Natural language processing, Power harassment, similarity

INTRODUCTION

According to a survey conducted by the Ministry of Health, Labour and Welfare on the Number of Consultations on Individual Labor Disputes (Bullying and Harassment), the number of consultations has increased to 87,570 cases in 2018. Of these, the percentage of talks related to power harassment was high at 48.2% (Ministry of Health, Labour and Welfare. 2019). Power harassment can be categorized into several types, and mental aggression is the highest at 74.5% which indicates rampant harassment with ambiguous standards (Ministry of Health, Labour and Welfare. 2021). This reveals the background of the difficulty in determining whether harassment is harassment or not. According to the Power Harassment Prevention Law, the standard for power harassment is words or actions in the workplace that are made against the background of a special relationship, that go beyond

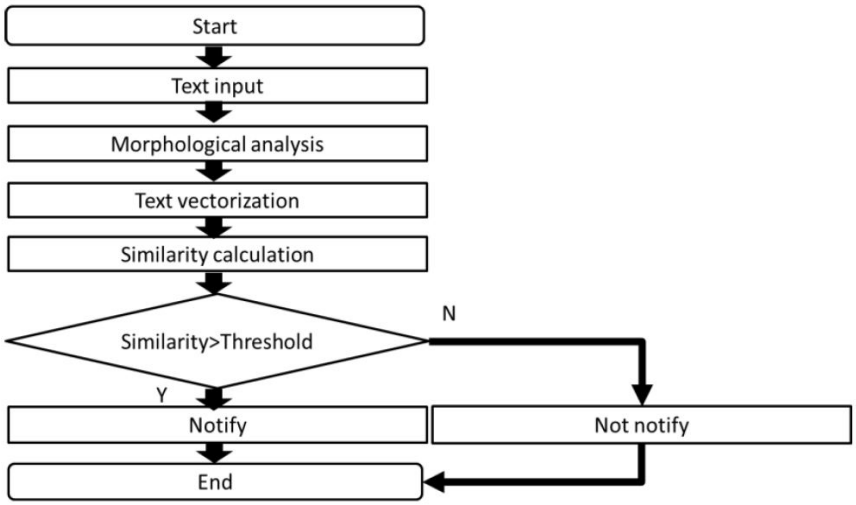

Figure 1: Proposed method.

what is necessary and reasonable in the course of work, and harm the working environment of the worker. The ambiguity of this standard makes it difficult to judge (Ministry of Health, Labour and Welfare. 2021). Against this background, there is a possibility that the harassment may become power harassment without the perpetrator being aware of it, and the issue of how to determine power harassment against the perpetrator before it occurs is considered. In addition, although many words are regarded as abusive language, it does not necessarily mean that if a comment is included, it constitutes power harassment. Therefore, it is necessary to judge whether or not it is power harassment based on the situation and context at the time. The ultimate goal of this study is to notify the perpetrator when the possibility of power harassment is judged to be high based on the conversational voice. We propose a method to determine whether the target text constitutes power harassment or not using natural language processing based on text data of past court precedents.

PROPOSED METHOD

The flow of the proposed method is shown in Figure 1. First, the text to be judged is decomposed into little words using a language model. In this case, Juman++ is used for morphological analysis. This enables analysis to take into account the semantic naturalness of the expressed sequence (KurohashiLaboratory et al. 2017). Next, we use BERT, a type of machine learning, to calculate the cos-similarity between sentences based on the word sequence information. The BERT Japanese Pre-trained Model is used as a pre-trained model (Jacob Devlin et al. 2019).

The cos similarity indicates how similar the two vectors are. The formula for calculating the cos similarity is shown in Equation 1.

$$\cos\ \text{similarity} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

Table 1. Representative text and meaning.

Representative Text	Meaning
sigoto ga deki nai n daxtu tara haya ku ya mere ba ii	Quit. I can't do the job.
yo no naka name ten zya nee yo , bakayarou	Idiot. Disparagement.
o mae ha nanisama no tumori na n da	Thoughtless words. Disparagement.
o mae ha i te mo muda da kara ko naku te ii	Denial of character. Don't come to the office.
si n de simae ba ii	Die.
buxtu korosu zo	Kill.

A_i and B_i are 768-dimensional feature vectors of the target text obtained from the hidden layer of BERT and the representative text received from the precedents respectively. The reasons for selecting representative texts are to reduce processing time and prevent unintended measurement results of similarity. By comparing the calculated similarity with the threshold value, a determination is made as to whether or not power harassment is present.

EXPERIMENTAL METHOD

Approximately 1,000 texts were collected from power harassment court cases publicly available from the courts and the Ministry of Health, Labor, and Welfare. We selected and used 100 texts. They were classified into six categories with similar meanings, and representative texts were chosen for each type. The reasons for choosing representative texts are to reduce processing time and prevent unintended measurement results of similarity. The selection method was to select texts with direct expressions such that similarity could be calculated appropriately. Table 1 summarizes the implications and representative texts. Here, representative texts are Japanese sentences converted into romaji.

Two types of surveys were developed and administered in this experiment. Survey 1 was used to determine threshold values. Survey 2 was used to evaluate accuracy. The questionnaire was administered to 15 participants. The number of texts was 100 texts each, 50 texts from precedents, and 50 texts based on daily conversation, which the researcher created. The case texts were distributed in equal proportions among six representative texts and texts with the same meaning, and all readers were statements made by a supervisor to a subordinate.

Participants were asked to respond on a five-point scale of "is power harassment," "rather power harassment," "can't say either," "rather not power harassment," and "not power harassment". We given responses of $[+2, +1, 0, -1, -2]$ respectively. If the sum of the 15 responses was more significant than 0, it was considered power harassment; otherwise, it was not considered power harassment.

From the results of Survey 1, the threshold with the highest F-measure was calculated, and that value was determined as the threshold value, which was then used to evaluate the results of Survey 2.

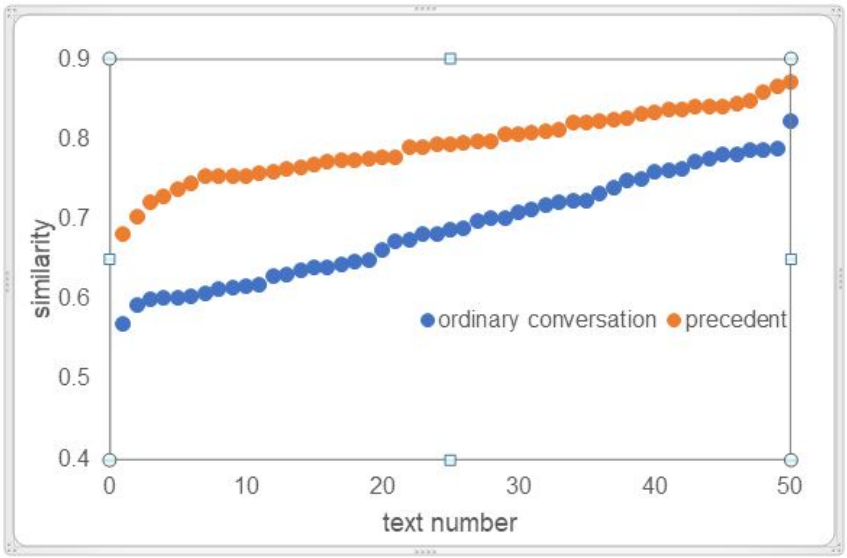

Figure 2: The similarity of each text (Survey1).

Table 2. Result when the threshold is 0.7296.

Survey	Accuracy	Recall	Precision	F-measure
Survey1	0.7800	0.6935	0.9348	0.7963
Survey2	0.8100	0.7678	0.8776	0.8190

EXPERIMENTAL RESULTS

The experimental results showed that the processing time for a single text was about 3 seconds.

A graph summarizing the similarity of the Survey1 texts in ascending order is shown below in Figure 2.

The results of Survey 1 show that four of the texts in the case law text are not power harassment, according to the survey results. From the results of Survey 1, the threshold for the highest F-measure was 0.7296.

A graph summarizing the similarity of the Survey 2 texts in ascending order is shown below in Figure 3.

The results of Survey 2 show that one of the texts of the case law is not power harassment, as a result of the survey.

The evaluation method of this study was evaluated using Accuracy, Recall, Precision, and F-measure, which are used in machine learning prediction performance. The summary is shown in Table 2 and is presented below.

Table 2 shows that the F-measure was generally good at 0.8190 for Survey 2. However, it can be seen that recall was low in the two surveys' results. According to the result, it can be said that many texts were predicted not to be power harassment based on the questionnaire results, but the similarity exceeded the threshold value.

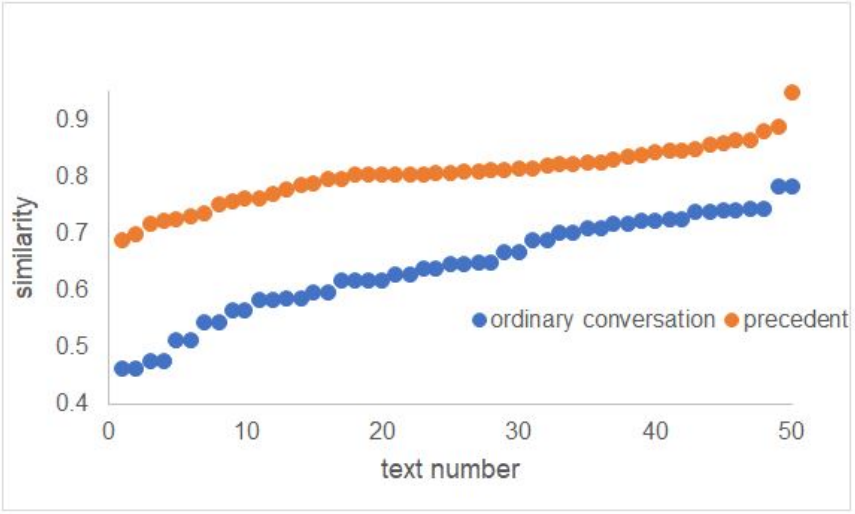

Figure 3: The similarity of each text (Survey2).

Table 3. Accuracy of each representative text.

Representative Text	Accuracy 1	Accuracy 2
sigoto ga deki nai n daxtu tara haya ku ya mere ba ii	95.0%	6.7%
yo no naka name ten zya nee yo , bakayarou	93.9%	7.6%
o mae ha nanisama no tumori na n da	90.0%	0.0%
o mae ha i te mo muda da kara ko naku te ii	82.8%	7.6%
si n de simae ba ii	62.5%	3.8%
buxtu korosu zo	100%	1.9%

We also evaluated the accuracy of each representative text. They are summarized in Table 3 and presented below. Accuracy 1 is the percentage of the assumed representative texts measured correctly, and Accuracy 2 is the percentage of readers predicted not to be power harassment that the program misjudged as power harassment. The higher the accuracy 1, the better the accuracy, and the lower the Accuracy 2.

The results indicate that the accuracy of representative texts such as "You're not here to stay, so don't come." and "You should be dead." was low. The reasons for this are that there were probably problems in the classification of texts involving subjective aspects and in the selection method and number of representative texts.

CONCLUSION

In this paper, we proposed a method that used natural language processing to determine whether a target text constitutes power harassment based on textual data of past power harassment cases, and obtained high accuracy. However, there are some problems with low recall and the precision of specific representative texts. Future issues include evaluating changes to representative text with low accuracy and resolution of topics such as privacy

and noise in systemizing the system. In addition, to further improve accuracy, we are studying a method to determine whether a text is power harassment or not by applying emotion estimation of the text, one of the techniques of natural language processing.

REFERENCES

Court, (2021) Database of precedents. https://www.courts.go.jp/app/hanrei_jp/search1

Jacob Devlin , Ming-Wei Chang , Kenton Lee & Kristina Toutanava. (2019) "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL-HLT.

Kurohashi & Murawaki Laboratory, (2017) Japanese Morphological Grammar System Juman++.

Ministry of Health, Labour and Welfare, (2018) Major Court Decisions Related to Power Harassment. https://www.mhlw.go.jp/content/11909500/000548187.pdf

Ministry of Health, Labour and Welfare, (2019) Harassment in data. https://www.no-harassment.mhlw.go.jp/foundation/statistics/

Ministry of Health, Labour and Welfare, (2021) Survey on Harassment in the Workplace. https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000165756.html

Rin Hirakawa, Yoshihisa Nakatoh, (2018) "Study on Visit Sales Detection using Similarity of Paragraph Vector", Proceedings of the 6th IIAE International Conference on Intelligent Systems and Image Processing.

Y. Kenda, H. Kikuta, K. Tanida, (2017) Introduction to Machine Learning with Free Libraries 150–152.