Journal of Systemics, Cybernetics and Informatics (2023) 21(2), 1-6
https://doi.org/10.54808/JSCI.21.02.1

A Post-Processing Framework for Crowd Worker Responses Using Large Language Models

Ryuya ITANO, Tatsuki TAMANO, Takahiro KOITA
Graduate School of Science and Engineering, Doshisha University
Kyotanabe, Kyoto 610-0394, Japan

and

Honoka TANITSU
Faculty of Science and Engineering, Doshisha University
Kyotanabe, Kyoto 610-0394, Japan

ABSTRACT[1]

To develop quality crowdsourcing systems, aggregating responses from workers is a critical issue. However, it has been difficult to construct an automatic mechanism that flexibly aggregates worker responses in natural language. Accordingly, responses need to be collected in a standardized format, such as binary-choice or multiple categorizations, to avoid large aggregation costs. Recently, with the advent of large language models (LLMs), natural language responses can be automatically and flexibly aggregated. We propose a framework that uses LLMs to flexibly aggregate natural language responses from workers and, as a promising example, consider this framework for crime detection from surveillance cameras using crowdsourced cognitive abilities. In an experiment using subjective evaluation, our proposed framework is shown to be effective for automatically aggregating natural language responses from crowd workers.

Keywords: Crowdsourcing, Human-in-the-loop, Large Language Models (LLMs), Anomaly Detection, Crime Detection, Video Recognition

1. INTRODUCTION

Aggregating crowd workers' responses accurately is a key issue in developing quality crowdsourcing systems. Currently, to aggregate worker responses automatically, workers should answer in a fixed format, such as the binary choice format. However, when responses are collected in such a way, workers have only two choices, which may not promote proactive recognition. Furthermore, determining whether the worker is a spammer (workers who do not engage in tasks sincerely) from binary choice answers is difficult. Therefore, a natural language response format encouraging workers to enter text freely is ideal. By employing a natural language response format, workers are more likely to engage in tasks proactively, and the increased amount of information from workers is expected to help identify spammers easily. Whether a change from binary choice to natural language format improves the quality of responses still needs to be verified, which is done through a preliminary experiment in Section 3.

Aggregating responses that includes differences in expressions and various misspellings is complicated, and the aggregation must be automated for incorporation in a crowdsourcing system. When dealing with this problem as an automatic text classification problem, a sufficient number of labeled training datasets and learning costs are required. Methods to comprehensively aggregate multiple responses from crowd workers rather than classifying each response text have also been proposed. However, these methods are extractive approaches such as choosing a best response, which cannot directly incorporate all responses gathered from workers.

To automatically aggregate natural language responses from workers while considering all of the responses, we propose a new post-processing framework using large language models (LLMs). LLMs are natural language processing models such as OpenAI's ChatGPT[2], and their wide range of capabilities to process natural language makes them promising for applications in many domains. In this study, we adopt LLMs for aggregation modules in a proposed post-processing framework with the expectation that LLMs can handle differences in expressions and various misspellings of worker responses.

Our goal is to develop quality crowdsourcing systems. In this study, we focus on improving the quality of crowdsourced responses toward that goal. The main contributions of this study are as follows:

[1] The author acknowledges Mr. Ron Read for his assistance in the English editing of this paper.

[2] **ChatGPT,** https://openai.com/blog/chatgpt

- A preliminary experiment to evaluate the quality of responses by changing the response format;
- The proposal of a new post-processing framework using LLMs;
- An experiment using subjective evaluations as a way to evaluate the proposed framework.

2. RESEARCH BACKGROUND

The significance of automated natural language aggregation is based on the challenges that arose when building a system of anomaly detection from surveillance cameras utilizing crowdsourced human cognitive abilities [1] [2]. In this section, we describe the background of the building of a crowdsourced anomaly detection system and its challenges. In addition, we introduce other areas in which automated natural language aggregation is useful.

Necessity of automated crime detection
The installation of surveillance cameras is increasing because these devices are becoming cheaper with the development of IoT technology. However, this trend has required a larger workforce to monitor the surveillance camera videos. Consequently, due to insufficient monitoring, serious anomaly moments, such as times of criminal activity, may be overlooked. The later the crimes are detected, the later the police are called, giving suspects a greater chance to escape. Therefore, it is desirable to detect anomalies automatically and quickly from surveillance camera videos.

Existing deep-learning method
Although Several methods using deep-learning models have been proposed to automate crime detection, high detection quality has not been achieved [3][4]. This is because training video datasets containing moments of crime are generally unavailable, making it difficult to train models sufficiently. Moreover, the models tend to generate false alerts for movements of many people or irregular movements such as flying insects. Thus, developing an anomaly detection system that only uses deep-learning models is difficult.

Crowd-aided Method
In our previous research, we proposed a crowd-aided anomaly detection method combining deep-learning models with crowdsourced human cognitive abilities [2]. In the crowd-aided method, chunks of surveillance videos are first input into a deep-learning model, and the model calculates an anomaly score for each video chunk. Video chunks with an anomaly score higher than a certain value are then recognized by crowd workers. The crowd workers watch these video chunks and determine whether they contain an anomaly moment. Each of the video chunks is given a new anomaly score based on several crowd workers' answers. Finally, the anomaly detection system determines whether to generate an alert based on the updated anomaly score. In this way, high detection quality can be expected due to the partial incorporation of human recognition. This method is also expected to cost less than hiring a surveillance video monitor because it is combined with deep-learning models. The quality evaluation of the crowd-aided method involves the use of frame-level AUC (Area Under the Curve), a binary-classification evaluation index created for classifying video frames into binary values, i.e., anomaly or not. The crowd-aided method produced a higher frame-level AUC (72.94%) than the deep-learning method (60.99%), but there remains room for improving quality. This is because many crowd workers tend to answer "yes" (this video has an anomaly moment) for video chunks without any anomaly moments, resulting in quality loss. To address this quality loss, we attempted approaches such as changing task descriptions given to workers or changing the rewards of tasks, but these did not improve quality. Therefore, we focus on the response format and consider changing it from a conventional binary choice format to a more complex natural language format that prompts crowd workers to engage in tasks more proactively. When aggregating natural language responses from workers, doing this process automatically is a key issue.

Automated aggregation in other areas
Automated aggregation of natural language responses could be useful in areas other than anomaly detection systems. VizWiz [5] is a mobile application that assists the blind through crowdsourcing. Its blind users take a picture of what's in front of them with their smartphones and ask crowd workers for identification. Multiple workers respond in natural language, and the users listen to the answers being read and determine what's in front of them. Users sometimes struggle because they must aggregate all of the responses to make a decision. Zensors [6] uses crowdsourcing as a smart sensor, for example, showing surveillance camera video to workers and asking them for such subjective judgments as "is the line orderly?" and "how messy is the countertop?". However, mechanically aggregating such subjective judgments is difficult.

3. PRELIMINARY EXPERIMENT

In this section, we evaluate the quality of worker responses by changing the response format from binary choice format to natural language format, with the aim of exploring whether the latter really improves quality. Here, the experiment is limited to the crime of shoplifting as a type of anomaly.

Experimental conditions
We prepared two response formats: binary choice and natural language. In the binary choice format, workers watch a video and answer whether the video has a moment of crime by selection with a YES/NO button. In the natural language format, workers answer with the type of anomaly freely by entering text. If there are no anomalies in the video, workers are asked to enter the text of "None". We

also prepared two videos: one containing a shoplifting scene (with crime) and the other containing no crime scene (without crime). Each video was created by manually clipping 10 seconds each of the crime and non-crime moments from a single video in the UCF-Crime Dataset[3]. Consequently, we prepared four conditions: two types of response formats for two types of videos.

Experimental protocol

We collected 50 responses from crowd workers for each of the four conditions described in the previous subsection, i.e., a total of 200 responses were collected. Crowd workers were recruited from Amazon Mechanical Turk[4]. After collecting responses from workers, the correct answer rate was calculated for each condition. Here, the correct answer rate is the percentage of correct responses out of all responses. The videos have two labels: "with crime" and "without crime", so we treat "YES" as "with crime" and "NO" as "without crime" in the binary choice responses; however, in natural language responses, if a crime type is entered, we mechanically treat it as "with crime," and any other input is treated as "without crime". In this experiment, the aggregation of natural language responses is done manually.

Results and discussion

Table 1 shows the correct answer rate for each condition. In the without-crime condition, the correct answer rate of the natural language format was 30% higher than that of the binary choice format. As we expected, the natural language format was able to eliminate the tendency of crowd workers to answer "YES" (crime is happening) for videos without crime. On the other hand, in the with-crime condition, the correct answer rate of natural language format was 8% lower than that of binary choice format. Perhaps certain natural language responses whose mechanical discrimination of crime-related activity is difficult lowered the quality, such as "lifting", "shooting", and "the very bad". This demonstrates the need for flexible aggregation of natural language responses that does not use a mechanical approach.

Table 1: Correct answer rate for each condition

	Binary choice	Natural language
With crime	**100%**	92%
Without Crime	54%	**84%**

4. RELATED WORK

In this section, we introduce existing research related to natural language aggregation, including text classification, methods using embeddings, and the introduction of LLMs.

Automatic text classification

Automatic text classification is a well-researched issue [7]. Much research has adopted machine-learning or deep-learning to achieve automatic text categorization. A commonly used way to create a text classification model is to fine-tune base models of natural language processing represented by Bidirectional Encoder Representations from Transformers (BERT) [8]. However, fine-tuning these models requires a sufficient number of labeled training datasets and learning costs. Furthermore, these fine-tuned models are domain specific, so models must be trained for each domain. For example, training datasets differ if you want to discriminate only a certain type of crime (e.g., shoplifting-related text or not) or any type of crime (i.e., crime-related text or not).

Selective aggregation using embeddings

Several studies have attempted to comprehensively aggregate multiple responses from crowd workers rather than classifying each response text. The method proposed by Chai et al. [9] represented the natural language responses from workers with multiple embeddings and estimated the true answer from these embeddings. Finally, it selects the response that is closest to the estimated answer. Li et al.'s method [10] also estimated the true answer by representing the worker responses with embeddings and weighted the embeddings according to each worker's reliability. The above methods are effective for selecting the best response. However, since they are selective aggregation approaches, they cannot directly incorporate all responses from crowd workers. Therefore, these methods may not be effective for considering all responses and making a decision like issuing an alert.

Large Language Models

In the area of natural language processing, large language models (LLMs) have been attracting attention. LLMs are deep-learning models with a larger number of trainable parameters than conventional language models, achieving human-like natural language generation by learning a large number of sentences. Among current LLMs, models such as GPT-3.5 and GPT-4, available in OpenAI's ChatGPT, are fine-tuned to perform any task including translation, text summarization, sentiment analysis, and document classification [11]. LLMs can perform these tasks due to their zero-shot ability, which outputs answers without exemplary answers [12]. The user can obtain the desired output from an LLM by entering a phrase called a "prompt" (e.g., "Summarize the following statement..." or "Classify the following sentence as positive or negative."). Although the advent of LLMs is still very recent, their applications are being considered in a variety of areas. This study assumes that LLMs can automatically aggregate natural

[3] **UCF-Crime Dataset**, https://www.crcv.ucf.edu/projects/real-world/

[4] **Amazon Mechanical Turk,** https://www.mturk.com

language responses from crowd workers by entering the appropriate prompts in the LLMs.

5. PROPOSED FRAMEWORK

To automatically aggregate natural language responses from workers, while considering the differences in expressions and misspellings, we propose a post-processing aggregation framework using LLMs. **Figure 1** outlines the proposed framework. This is an example of a crowdsourcing task for answering what crime is occurring in a video. Worker 1 responded "shoplifting", worker 2 misspelled it as "sshoplifting", worker 3 responded with the different expression "theft", and worker 4 responded "no". Under such circumstances with differences in expressions and misspellings, the aggregation module can assemble all of the responses and finally produce output in any format, including a single word, a decision to alert or not, or an anomaly score. The proposed framework does not have to fine-tune models to obtain the desired formats of outputs but only to change the prompts, so labeled training datasets and learning costs are not needed. Moreover, the aggregation of the proposed framework is not a selective approach, permitting the direct incorporation of all responses from the crowd workers.

6. EXPERIMENT

We conducted an experiment using subjective evaluations to test how well our proposed framework operates.

Experimental conditions
We prepared three conditions of worker responses: with crime, without crime, and an artificial fifty-fifty condition. In with- and without-crime conditions, we collected natural language responses from ten crowd workers to both with- and without-crime videos. These videos were the same as those prepared in the preliminary experiment in Section 3. In addition, we prepared the artificial fifty-fifty condition in which 5 out of 10 responses were "Shoplifting" and the rest were "none"; this condition was used to examine the behavior of LLMs when the responses are split fifty-fifty between crime-related and not crime-related.

Prompt settings
We used a GPT-3.5 model of ChatGPT as the LLM in the aggregating module and created a prompt template as shown in **Figure 2**. The "{response_XX}" parts of the prompt template are replaced by the responses shown in **Table 2** and entered in the LLM. The prompt was designed to ask LLMs to output whether a crime has occurred with "YES" or "NO" to determine whether to generate an alert. The prompt also asks LLMs for clear explanations to justify the decision to ensure explainability.

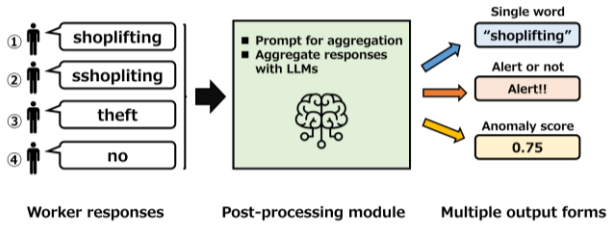

Figure 1 Proposed framework

Subjective evaluation items
To evaluate outputs from LLMs, we set the following three subjective evaluation items:

- Ability to capture misspellings and other differences in expression;
- Whether the rationale for a decision is justified;
- Whether hallucinations occur.

"Ability to capture misspellings and other differences in expression" is the most important question in this study, so we included this question in the subjective evaluation items. "Whether the rationale for a decision is justified" is included because ensuring the explainability of the decision to generate an alert or not is considered necessary in developing a reliable anomaly detection system. Furthermore, LLMs are known to have a propensity called hallucination. Hallucination is the behavior of LLMs to output false knowledge as if it were accurate. Therefore, we prepared the item: "Whether hallucinations occur" to explore whether the LLM hallucinated even in our basic tasks.

Results
Table 2 shows the natural language responses for each condition. In the with-crime condition, differences in expression between "shoplifting" and "theft" as well as misspellings such as "teft" and "tief" were observed. **Figure 3** shows the GPT-3.5 response in the with-crime condition. The decision of GPT-3.5 is "YES", and the process leading to the decision is detailed. Responses such as "shop lifting" and "theft" are consolidated into a single broader concept of theft-related activities. In addition, responses such as "teft" and "tief", considered misspellings, are included in the theft-related activities. On the other hand, the number of theft-related responses, including misspellings, that we can identify are five: response_02("teft"), 05("shop lifting"), 07("tief"), 08("Theft"), and 09("theft"), but GPT-3.5 incorrectly counts six theft-related responses. **Figure 4** shows the GPT-3.5 responses in the without-crime condition. The decision of GPT-3.5 is "NO", and the process leading to the decision is also detailed in without-crime condition. The answering format of GPT-3.5 in without-crime condition is slightly different from that in with-crime condition, and no counting of responses is conducted. **Figure 5** shows the GPT-3.5 response in the artificial fifty-fifty condition. Although the decision of GPT-3.5 is inconclusive, the reason for the decision is detailed, and

the count of responses is conducted effectively. The inconclusive output of GPT-3.5 is generally correct, but if we require a YES/NO decision, we should predefine a threshold for making an alert in the prompt.

Discussion

Here, we discuss the predetermined subjective evaluation items. Regarding the item "Ability to capture misspellings and other differences in expression", LLMs have the abilities to consider differences in expression and misspellings and to aggregate responses with similar meanings into a single broader concept. Regarding the item "Whether the rationale for a decision is justified", LLMs can make a decision logically through a detailed derivation process. Therefore, we can say that LLMs with a performance of at least GPT-3.5 can aggregate natural language responses considering differences in expression and misspellings. On the other hand, regarding the item "Whether hallucinations occur", LLMs sometimes make mistakes when counting the responses with similar meanings. This means that LLMs sometimes hallucinate, but this impact is unknown. Therefore, we should carry out a quantitative evaluation of the proposed framework. Although we did not conduct a trial-and-error process of prompt adjustment (generally called prompt engineering) in this study, we believe that prompt engineering should be applied to reduce hallucinations and unify the output format.

7. CONCLUSION AND FUTURE WORK

We identified the key problem of aggregating natural language responses from crowd workers and proposed an aggregation framework that solves this with LLMs. In an experiment using a shoplifting-detection task, our framework flexibly processed natural language responses from workers. In contrast to existing studies, our framework demonstrated the capability to consider all of the responses from workers without additional training of the model.

Our future work includes prompt engineering and quantitative evaluations. Prompt engineering requires standardized output to reduce hallucinations. In quantitative evaluations, we will experimentally compare the existing methods using binary classification evaluation metrics, such as precision, recall, and F-measure, after setting baselines and detailed experimental protocols. Beyond that, we are considering using LLMs to detect and eliminate spammers by leveraging the advantages of collecting responses in natural language.

8. REFERENCES

[1] R. Itano and T. Koita, **Evaluation of the effectiveness of a crowdsourcing-based crime detection system,** IEICE Communications Express, 11(9):607–611, 2022.

[2] R. Itano, T. Nohara and T. Koita, **Crowd-Aided Anomaly Detection in Surveillance Videos,** IEEE International Conference on Big Data (Big Data), pp. 3992-3994, 2022.

[3] W. Sultani, C. Chena and M. Shah, **Real-world anomaly detection in surveillance videos,** IEEE conference on computer vision and pattern recognition, pp. 6479–6488, 2018.

[4] H. Lv, C. Zhou, Z. Cui, C. Xu, Y. Li, and J. Yang, **Localizing Anomalies From Weakly-Labeled Videos,** IEEE Transactions on Image Processing, vol. 30, pp. 4505-4515, 2021.

[5] JP. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, RC. Miller, R. Miller, A. Tatarowicz, B. White, S. White, and T. Yeh, **VizWiz: nearly real-time answers to visual questions,** Proceedings of the 23rd annual ACM symposium on User interface software and technology, pp. 333-342, 2010.

[6] G. Laput, WS. Lasecki, J. Wiese, R. Xiao, JP. Bigham, and C. Harrison, **Zensors: Adaptive, Rapidly Deployable, Human-Intelligent Sensor Feeds,** Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 1935-1944, 2015.

[7] A. Dhar, H. Mukherjee, NS. Dash, and K. Roy, **Text categorization: past and present,** Artificial Intelligence Review Vol. 54, pp. 3007-3054, 2021.

[8] J. Devlin, MW. Chang, K. Lee, and K. Toutanova, **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,** Proceedings of naacL-HLT, vol. 1, pp. 4171-4186, 2019.

[9] L. Chai, H. Sun, and Z. Wang, **An error consistency based approach to answer aggregation in open-ended crowdsourcing,** Information Sciences, vol. 608, pp. 1029-1044, 2022.

[10] J. Li, and F. Fukumoto, **A Dataset of Crowdsourced Word Sequences: Collections and Answer Aggregation for Ground Truth Creation,** Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP, pp. 24-28, 2019.

[11] L. Ouyang, J. Wu, X. Jiang, D. Almeida, CL. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., **Training language models to follow instructions with human feedback,** Advances in Neural Information Processing Systems, vol. 35, pp. 27730-27744, 2022.

[12] J. Wei, M. Bosma, VY. Zhao, K. Guu, AW. Yu, B. Lester, N. Du, AM. Dai, and QV. Le, **Finetuned Language Models Are Zero-Shot Learners,** The International Conference on Learning Representations, 2022.

Table 2: Responses for each condition

	With crime	Without crime	Artificial fifty-fifty
response_01	none	none	Shoplifting
response_02	teft	nonenoe	Shoplifting
response_03	shopping	No	Shoplifting
response_04	None	no body shopping	Shoplifting
response_05	shop lifting	NONE	Shoplifting
response_06	Shopping	NONE	none
response_07	tief	none	none
response_08	Theft	no	none
response_09	theft	none	none
response_10	none	none	none

```
Below are results of ten crowd workers who watched surveillance camera footage and identified a crime as occurring
or "none" if it did not. Summarize these results, determine whether a crime occurred in the video, and state "YES"
or "NO" in the text, followed by a reason.

Worker 1: {response_01}
Worker 2: {response_02}
Worker 3: {response_03}
Worker 4: {response_04}
Worker 5: {response_05}
Worker 6: {response_06}
Worker 7: {response_07}
Worker 8: {response_08}
Worker 9: {response_09}
Worker 10: {response_10}
```

Figure 2: Prompt template

Based on the results provided by crowd workers, a crime is apparently occurring in the surveillance camera footage. The majority of the workers (6 out of 10) mentioned some form of theft-related activity, such as "teft," "shop lifting," "tief," "Theft," and "theft." Although a few workers mentioned "shopping," which does not explicitly indicate a crime, the consistent references to theft suggest that a crime is indeed taking place.

Therefore, the conclusion:

YES, a crime is occurring in the video.

Reason: **The majority of the crowd workers mentioned theft-related activities, indicating the presence of a crime.**

Figure 3: GPT-3.5 response in with-crime condition

Based on the responses of ten crowd workers, none of them apparently identified a crime in the surveillance camera footage. The responses include variations of "none," "No," "no body shopping," and "no." Therefore, the summary of the results indicates that the majority consensus among the workers is that no crime is occurring in the video.

Result: **NO**

Reason: **None of the ten crowd workers identified any criminal activity in the surveillance camera footage.**

Figure 4: GPT-3.5 response in without-crime condition

Based on the results of the ten crowd workers who watched the surveillance camera footage, it appears that five workers identified a "shoplifting" as the crime, while the other five workers labeled it as "none," indicating no crime.

Therefore, the result is **inconclusive** as **there is an equal number of workers labeling it as a crime and as no crime.** To determine whether a crime is occurring in the video, further investigation or additional expert analysis may be needed to make a definitive determination.

Figure 5: GPT-3.5 response in artificial fifty-fifty condition