NTCIR 17 Conference: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan
DOI: https://doi.org/10.20736/0002001301

ISLab at the NTCIR-17 QA Lab-PoliInfo-4: Models for Automatically Identifying Politicians' Stances on Bills

Guan-Yu Chen
Department of Computer Science and Information Engineering
National Kaohsiung University of Science and Technology
Kaohsiung, Taiwan, R.O.C
C109151123@nkust.edu.tw

Yu-Cheng Liu
Department of Computer Science and Information Engineering
National Kaohsiung University of Science and Technology
Kaohsiung, Taiwan, R.O.C
F110151106@nkust.edu.tw

Tao-Hsing Chang
Department of Computer Science and Information Engineering
National Kaohsiung University of Science and Technology
Kaohsiung, Taiwan, R.O.C
changth@nkust.edu.tw

Fu-Yuan Hsu
Research Center for Psychological and Educational Testing
National Taiwan Normal University
Taipei, Taiwan, R.O.C
kevin@rcpet.ntnu.edu.tw

ABSTRACT

This paper aims to design a model that can determine whether the politician's stance is approved or disapproved the bill based on the politician's utterance on a specific bill in the parliament. This study proposed two frameworks for determining the stance in utterances. The first framework involves concatenating BERT model with Bi-LSTM model to form a comprehensive decision-making model while the second framework is concatenating Curie model with ChatGPT model. This paper used the dataset provided by Stance Classification 2 task in NTCIR-17 for model training and testing, and GPT-based model this paper proposed achieved an accuracy of 0.932.

KEYWORDS

BERT, Bi-LSTM, GPT-3, ChatGPT, Curie, Stance Classification.

TEAM NAME

ISLab

SUBTASKS

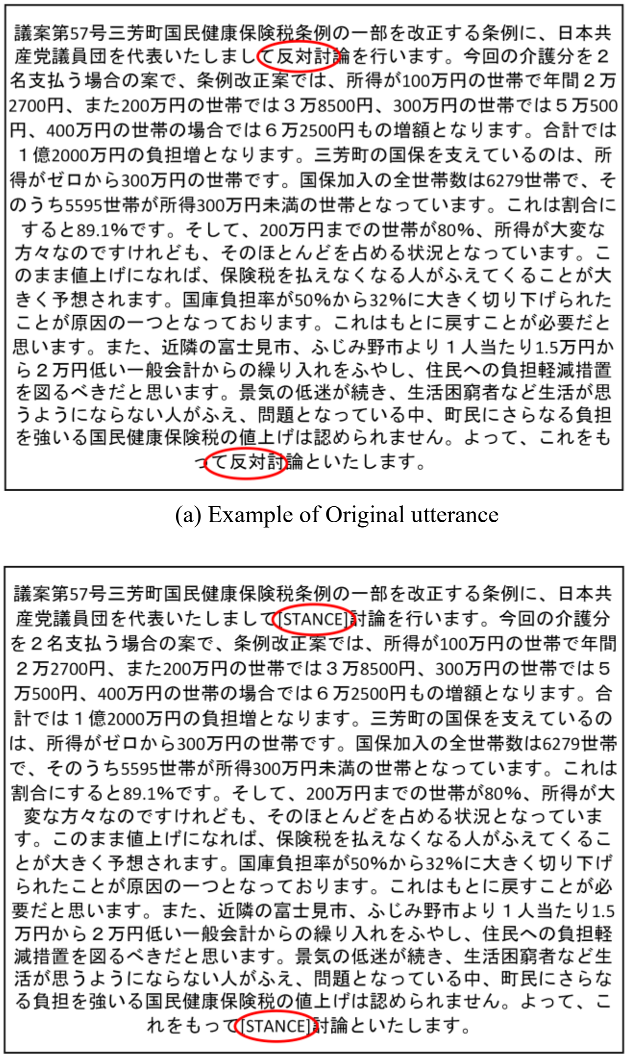
Stance Classification-2 (Japanese)

1 INTRODUCTION

This study, as a subtask of the QA Lab-PoliInfo-4 [1] task of the NTCIR-17 conference under the “Stance Classification 2” subtask, aims to design a model that can determine whether the politician's stance is approved or disapproved the bill based on the politician's utterance on a specific bill in the parliament. This subtask is similar to the subtask in NTCIR-15 [2], but the types of processing texts subtask are different; Figure 1 illustrates the difference between these subtasks. Figure 1(a) shows the original text of an utterance, which is the form of the text processed in NTCIR-15. This utterance focuses on a specific politician's opinion on Proposal 57. The politician believes that amendments to the bill will increase people’s medical expenses before eventually becoming unaffordable and hence is disapproving the proposal.

Figure 1(b) shows a new text formed by masking words related to the politician's stance in Figure 1(a) with a [STANCE] marker. As shown in Figure 1(a), when the word “反対(disapproval)” appears in rows 2 and 19, as marked in red, it is replaced with [STANCE], which generates the text in Figure 1(b), i.e., the text processed in the subtask "Stance Classification 2." As explicit stance-indicative words are replaced with the same marker, the stance-determining model must rely on other information within the text to determine the conveyed stance of the given utterance.

We proposed two model frameworks for determining the stance in utterances. The first framework involves concatenating a semantic space model with a sequential data processing model to form a comprehensive decision-making model (hereafter referred to as the bidirectional encoder representations from transformers (BERT) [3]–long short-term memory (LSTM) [4] model). Our idea is to first transform the text through a language space model into a semantic space coordinate representing the text, anticipating that the text of approved and disapproved stances would form separate clusters in the semantic coordinate space. Therefore, by training a simple classifier to differentiate stances, we can convert the utterance to be judged into a semantic coordinate and then use the classifier to determine the stance. This framework has been widely applied to various classification issues with considerable success.



議案第57号三芳町国民健康保険税条例の一部を改正する条例に、日本共産党議員団を代表いたしまして反対討論を行います。今回の介護分を２名支払う場合の案で、条例改正案では、所得が100万円の世帯で年間２万2700円、また200万円の世帯では３万8500円、300万円の世帯では５万500円、400万円の世帯の場合では６万2500円もの増額となります。合計では１億2000万円の負担増となります。三芳町の国保を支えているのは、所得がゼロから300万円の世帯です。国保加入の全世帯数は6279世帯で、そのうち5595世帯が所得300万円未満の世帯となっています。これは割合にすると89.1%です。そして、200万円までの世帯が80%、所得が大変な方々なのですけれども、そのほとんどを占める状況となっています。このまま値上げになれば、保険税を払えなくなる人がふえてくることが大きく予想されます。国庫負担率が50%から32%に大きく切り下げられたことが原因の一つとなっております。これはもとに戻すことが必要だと思います。また、近隣の富士見市、ふじみ野市より１人当たり1.5万円から２万円低い一般会計からの繰り入れをふやし、住民への負担軽減措置を図るべきだと思います。景気の低迷が続き、生活困窮者など生活が思うようにならない人がふえ、問題となっている中、町民にさらなる負担を強いる国民健康保険税の値上げは認められません。よって、これをもって反対討論といたします。

(a) Example of Original utterance

議案第57号三芳町国民健康保険税条例の一部を改正する条例に、日本共産党議員団を代表いたしまして[STANCE]討論を行います。今回の介護分を２名支払う場合の案で、条例改正案では、所得が100万円の世帯で年間２万2700円、また200万円の世帯では３万8500円、300万円の世帯では５万500円、400万円の世帯の場合では６万2500円もの増額となります。合計では１億2000万円の負担増となります。三芳町の国保を支えているのは、所得がゼロから300万円の世帯です。国保加入の全世帯数は6279世帯で、そのうち5595世帯が所得300万円未満の世帯となっています。これは割合にすると89.1%です。そして、200万円までの世帯が80%、所得が大変な方々なのですけれども、そのほとんどを占める状況となっています。このまま値上げになれば、保険税を払えなくなる人がふえてくることが大きく予想されます。国庫負担率が50%から32%に大きく切り下げられたことが原因の一つとなっております。これはもとに戻すことが必要だと思います。また、近隣の富士見市、ふじみ野市より１人当たり1.5万円から２万円低い一般会計からの繰り入れをふやし、住民への負担軽減措置を図るべきだと思います。景気の低迷が続き、生活困窮者など生活が思うようにならない人がふえ、問題となっている中、町民にさらなる負担を強いる国民健康保険税の値上げは認められません。よって、これをもって[STANCE]討論といたします。

(b) Text in Figure 1(a) after masking

Figure 1: Example of Text Processed in this Study

The second framework is concatenating GPT-3's [5] two sub-models, i.e., Curie[6], with the summarization model of ChatGPT[7] (hereafter referred to as the GPT-based model). Large language models (LLMs), also known as foundation models, can be fine-tuned to meet the requirements of downstream tasks. We treated the "Stance Classification 2" subtask as a question-answering (QA) issue. In other words, we provided the model with an utterance and asked it to determine the stance of the utterance. Therefore, we used ChatGPT to summarize the utterance before fine-tuning the QA model Curie with the summarized text, thus enabling the model to determine the stance of the given text.

The remaining sections of this paper are organized as follows: Section 2 reviews various methods proposed by participants in the previous “Stance Classification” subtask. Section 3 explains the architectures and details of the two stance-determination frameworks proposed in this study. Section 4 demonstrates the performance of the proposed methods on the test dataset of the "Stance Classification 2" subtask. Finally, Section 5 discusses the properties and limitations of the proposed methods and potential directions for future works.

2 RELATED WORKS

In the subtask "Stance Classification" of NTCIR-15, several studies have proposed different models and methods to process training datasets, aiming to create better stance classification capabilities than the baseline model. Given the typical distinct formats of politicians' utterances and the common usage of certain words indicative of their stances on bills, several studies employed rule-based approaches for stance determination. For instance, [8] and [9] used rule-based methods to determine stance; moreover, the latter devised a neural network–based machine learning model. However, the rule-based method was more accurate than the machine learning model.

[10] proposed a model that concatenated rule-based methods and machine learning models. They first used the rule-based method to extract the unprocessed target bill from an utterance, which involved identifying the targeted bill number from the utterance and treating the following bill number as the endpoint to describe the content of the target bill. Next, they determined the values of three features in the target bill. The first feature adopted the rule-based method. If there was a clear description of a stance, such as "賛成"(approval) or "反対"(disapproval) in the bill, the stance was taken as the value of the first feature. The second feature involved using the principal component analysis (PCA) [11] model concatenated with BERT. This model converted the semantics of the target bill into a five-dimensional vector, which served as the second feature. The third feature used the sentiment polarity dictionary to calculate the sentimental value of the target bill to determine the value of the feature. Finally, they adopted the lightGBM algorithm [12] to synthesize the three feature values to determine the stance of the utterance.

[13] proposed a two-stage method. This study divided politicians’ utterance content into opinion and debate statements. The first stage determined whether opinion statements contain explicit content approving or disapproving the target bill. They adopted a three-step method to determine the stance of the speaking politicians. The first stage is to identify the affiliation of the bill proposer and politician to assess whether their parties were in a cooperative or oppositional relationship; this information was further used to infer whether the politician approved or disapproved the bill. If there was insufficient information to determine the politician's stance in the first stage, the process advanced to the second stage, which involved inputting debate statements into a BERT-based classification model. Based on the analysis of the content of the debate statements, it attempted to classify the utterances as either approved or disapproved to the target bill.

As the corpus of Stance Classification 2 is in Japanese, and the authors of this paper are not native Japanese speakers, it is not feasible to establish a stance detection model with rule-based methods that require a high level of proficiency in the Japanese language to analyze the linguistic rules. Therefore, this paper primarily adopts methods based on deep-learning neural networks to design the stance detection model.

3 Methodology

The motivations behind the design of the two proposed frameworks, i.e., the BERT–LSTM and GPT-based models, are explained in Section 1. Sections 3.1 and 3.2 elaborate on the details of these two methods, respectively.

3.1 BERT-LSTM Model

This model concatenates BERT and bidirectional LSTM (Bi-LSTM) [14] models. We believe that utterances can be transformed into semantic vectors: texts with the same stance should be closer in the vector space and those with opposing stances should be farther apart. Therefore, we can use a classifier to divide the semantic vectors of utterances into two groups: "賛成"(approval) and "反対"(disapproval). Based on this concept, this paper employed the BERT model to transform text into semantic vectors. However, as BERT can only process a limited number of words in a single text, we broke longer texts that exceeded BERT's limits into several segments, with each segment entered into BERT to generate its respective semantic vector and then integrated using Bi-LSTM.

Bidirectional Encoder Representations from Transformers (BERT) is a language model comprising a multilayer encoder based on the transformer [15] architecture. The BERT model is pre-trained with copious unlabeled text and employs attention mechanisms that consider relationships between all words in the entire text at once; this mechanism allows BERT to precisely transform text and words into vector coordinates in the language space. BERT's pre-training procedure employs two methods. The first approach is the Masked Language Model(MLM) method, which operates based on the principle of requiring the model to predict certain masked words during training. By predicting the masked words, the model can be adjusted to make more accurate predictions in the future, leading to more precise settings of hyperparameters affecting relationships between these words. The second pre-training method, known as the Next Sentence Prediction(NSP) method, requires the model to predict whether two input sentences are consecutive. The model is adjusted based on the prediction outcomes, which helps in setting hyperparameters affecting the semantic relationship between sentences more accurately.

The Long Short-Term Memory (LSTM) model is based on the recurrent neural network (RNN) architecture and addresses the vanishing and exploding gradient problems in RNNs. With the designed gate and cell memory mechanisms, LSTM can determine whether to retain previous computation results in the next step, thus allowing LSTM to effectively handle data with time sequence characteristics. As LSTM processes data sequentially, information processed earlier does not consider information later in the sequence during computation, which does not meet the requirement of this paper of considering the text as a whole when processing the information. Therefore, to fulfill the need for considering the context of the information, this paper employs Bi-LSTM, which processes the full sequence of text at once.

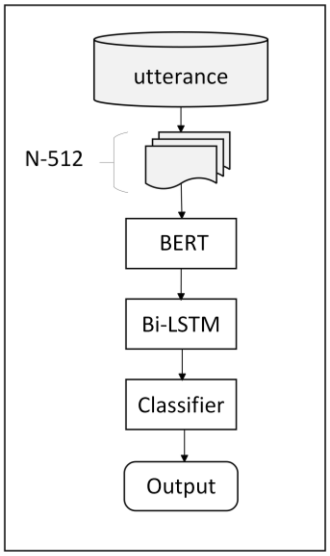


Figure 2: Architecture of BERT-LSTM.

Figure 2 illustrates the architecture of the BERT–LSTM model. Firstly, the model divided the utterances into several segments based on the maximum input length allowed by the BERT model. These segments were entered into BERT in a sequential order to obtain the semantic vectors for each segment. These semantic vectors were then entered sequentially into Bi-LSTM to generate the final vector output from the LSTM as the semantic vector of the utterance. This vector was then fed into the full connection classifier to determine the stance of the utterance.

3.2 GPT-based Model

This model concatenates summarization and QA models. As we considered the stance determination of utterances a QA problem, this paper attempted to employ the Curie QA model to assess the stance of utterance. Additionally, similar to BERT, the Curie model has limitations on the number of words it can process in a single text. As a result, this model divided longer texts over the QA model's limits into several segments; each segment underwent a generative summarization process through ChatGPT to shorten the text length in the segment without altering the meaning of the original content. Subsequently, the summarized texts of each segment were reassembled in sequential order to form a single summarized text that met Curie's length restrictions. This summarized text was then fed to Curie to determine the stance of the original utterance.

Curie and ChatGPT are constructed primarily based on the technology of GPT-3. GPT [16] was introduced to address the problem of the vast amounts of hard-to-obtain labeled text required for supervised learning. GPT involves pre-training a base model with a large amount of unlabeled text and then fine-tuning the base model on a small set of labeled text specific to the task. Like BERT, GPT is a transformer-based model. However, GPT uses the decoder part of the transformer architecture, whereas BERT focuses on the encoder part. Figure 3 illustrates the architecture of GPT.

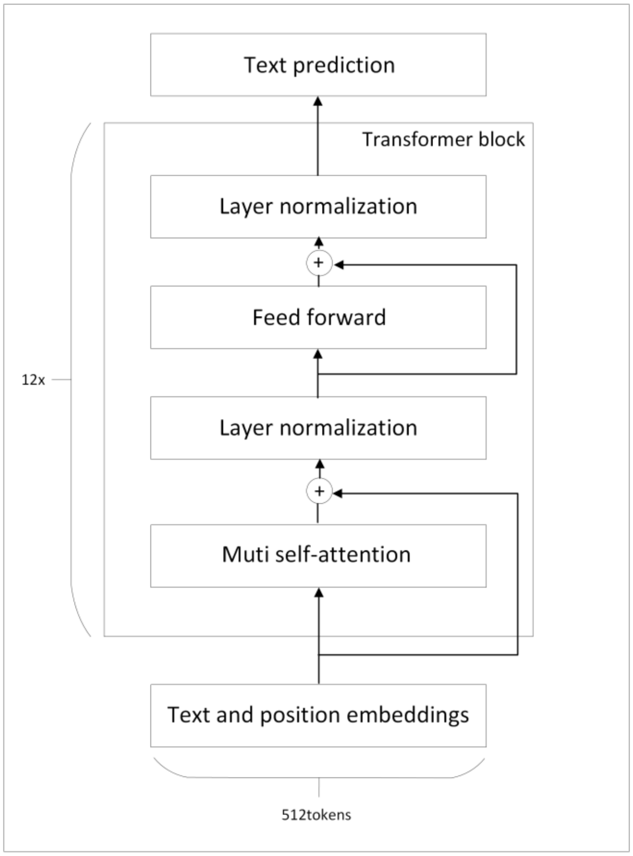


Figure 3: Architecture of GPT.

GPT-2 [17] is an improved version of GPT. GPT-2 involves moving the layer normalization module to the input of feed-forward and multi-self-attention networks; this layer normalization module is added after the last transformer block. While GPT-3 and GPT-2 possess the same architecture, GPT-3 alternates between dense and locally banded sparse attention models in transformer blocks. In the traditional transformer, during self-attention calculations, every two vectors need to compute attention with each other. Meanwhile, when alternating between dense and locally banded sparse attention models, the dense attention model calculates vectors at fixed intervals, whereas the locally banded sparse attention calculates a few vectors adjacent to the current position. This modification reduces the computational complexity of the transformer, thus enabling the efficient handling of longer texts.

In the second framework proposed in this paper, we employed ChatGPT as the summarization model. ChatGPT is a LLM built from GPT-3.5, which is GPT-3 fine-tuned with reinforcement learning from human feedback (RLHF) model training. RLHF, used by ChatGPT, consists of three steps. The first step is to record various human responses and use them as training data for a GPT-3.5 model in a supervised learning approach. In the second step, a set of questions is answered using various machine models, with the quality of the answers ranked manually; the sorted data are then used to train a reward model. The third step uses proximal policy optimization to fine-tune the hyperparameters of the model trained in the first step. Responses from the fine-tuned model are then scored by the reward model from the second step. The final scores are used to further refine the model; this process is iterated multiple times.

In the second framework proposed in this paper, the QA model employs Curie, a GPT-3 sub-model chosen from among the various GPT-3 sub-models provided by OpenAI owing to its faster computation speed and better performance in sentiment analysis tasks compared to other sub-models.

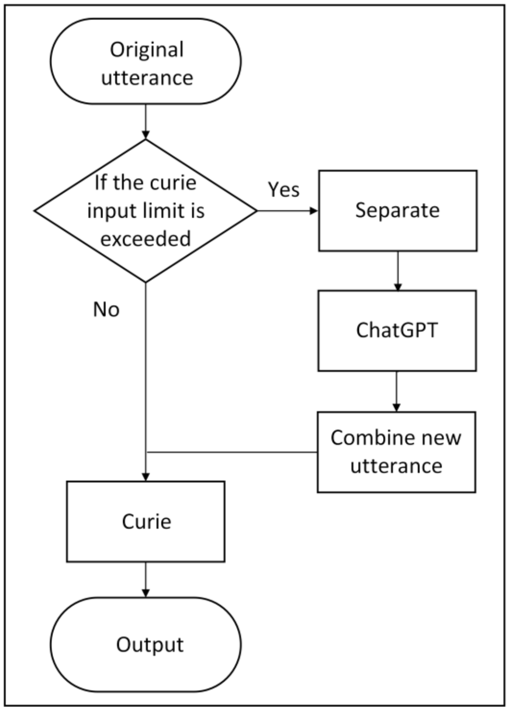


Figure 4: Architecture of GPT-based Model.

Figure 4 illustrates the architecture of the proposed GPT-based model. First, this model determined whether the length of the utterance exceeded Curie's maximum input length. If the text was within the word limit, it was directly fed to Curie for stance detection; if the text exceeded the input limit, this model segregated it into several segments. This process involved identifying sentences with [STANCE] markers and considering sentences between two markers as a segment. Moreover, text from the beginning to the first marker and text from the last marker to the end were considered individual segments.

This model added prompts, which this paper designed, to each segment and then fed it to ChatGPT for summarization. To avoid extremely short summaries, which could result in significant semantic differences between the summarized segments, we used the following prompt:

"この記事を絶対に n 字以内で日本語でまとめる。"
(Summarize this text in Japanese with an absolute maximum of n characters.)

Where n was calculated using the following formula:

$$n = floor((t - a)/b),$$

where t denotes the total number of words in the utterance, a denotes the total number of words in sentences containing the [STANCE] marker within the utterance, and b denotes the number of segments.

Finally, the summaries of each segment and the sentences containing the [STANCE] marker were reorganized in sequence to create the summary text of the original utterance. The summary text was subsequently used to fine-tune Curie. Since the input of Curie requires a prompt in addition to the summary text, the prompt this paper created for Curie is as follows:

"記事中の[STANCE]は[賛成]とすべきか[反対]とすべきか?"
(Should the [STANCE] in the text be [approved] or [disapproved]?)

The two prompts mentioned above are selected after testing many different prompts. In Section 4, we provide further examples of how to select the prompts of ChatGPT and Curie from various prompts.

4 EXPERIMENTS

This paper used the dataset provided by Stance Classification 2 for training and testing. The dataset included 8,534 utterances in the training set and 2,240 utterances in the testing set, with each utterance manually labeled with a stance (approved or disapproved). Owing to an error in an item of the training dataset and the fact that this paper extracted the first 1,000 utterances out of 8,533 as the validation dataset, the actual number of utterances used for training was 7,533; apart from the aforementioned dataset, no additional datasets were used in this study. The utterances in the testing set did not contain manually labeled stances. The results of the proposed models on the testing set were uploaded on the NTCIR-17 website, where the website automatically calculates the accuracy of the proposed model.

In the GPT-based model, two crucial parameters, i.e., feeding prompts to ChatGPT and Curie, were considered. This paper experimented with the following three prompts for ChatGPT:

1."日本語で n ワード以内で記事に要約する";
(Summarize the text content in Japanese with no more than n words.)

2."必ず日本語で n 字以内に要約してください";
(Summarize within n characters using Japanese.)

3."この記事を絶対に n 字以内で日本語でまとめる";
(Summarize this text in Japanese with an absolute maximum of n characters.)
This paper tried the following three prompts for Curie:

1."[法案]に関するあなたの立場は? [法案] に関連する質問のみに回答してください";
(What is the stance of this utterance? Only provide answers related to this utterance.)

2."記事中の[STANCE]は[賛成][反対]のどちらで書くべきでしょうか?[法案]に関する回答のみ回答してください";
(Is the [STANCE] in this utterance [approved] or [disapproved]? Only provide answers related to this utterance.)

3.”記事中の[STANCE]は[賛成]とすべきか[反対]とすべきか?";
(Is the [STANCE] in this utterance [approved] or [disapproved]?)

We randomly selected 100 utterances from the train dataset that need to be summarized by ChatGPT as the validation data. Then, we use the 9 combinations of the prompts mentioned above as the inputs of the proposed model to calculate the prediction accuracy of the model on the validation data. Table 1 lists the predicting accurate rates of the proposed model under various combinations. Table 1 shows that the prediction accuracy of the proposed model is higher when Curie uses prompt 3 as input. Moreover, the highest accuracy is obtained when the proposed model uses ChatGPT's prompt 3 and Curie's prompt 3. Therefore, the two prompts 3 (i.e. the prompts mentioned in Subsection 3.2) are adopted by the proposed model.

Table 1: the Prediction Accuracy of the Proposed Model under Various Combinations of Prompts

Selected Prompts		ChatGPT		
		Prompt1	Prompt2	Prompt3
Curie	Prompt1	0.73	0.84	0.82
	Prompt2	0.81	0.75	0.74
	Prompt3	0.91	0.87	0.92

Table 2 lists the stance determination results of the two proposed models on the testing set. As the pilot experiments suggested no significant differences in accuracy among different prompts, GPT-based model only employed the prompts mentioned in Section 3.2. Table 2 indicates that the GPT-based model outperformed the BERT-BiLSTM model in terms of accuracy.

Table 2: Accuracy of the Proposed Model on the Testing Set

Model	Accuracy
BERT-BiLSTM	0.916
GPT-based	0.932

According to the results in Table 3, we believe that the text summarized by ChatGPT may not fully capture the semantics in the original content; moreover, the semantics discarded in the summarization process could be crucial for Curie's stance determination. Based on this assumption, this paper analyzed the number of misjudgments for texts summarized and not summarized by ChatGPT, respectively. Table 3 shows the results of analysis mentioned above. Table 3 shows that the model error rate was 0.045 in utterances without summaries and 0.114 in utterances with summaries (a 2.53 times higher error rate). This significant difference demonstrates that the summarization step decreases the accuracy of stance determination.

Table 3: Error Rates of the GPT-based Model for Two Types of Utterances

	Total	True	False	Error Rate
Utterances	2240	2089	151	0.067
Utterances do not require summarization	1533	1463	70	0.045
Utterances require summarization	707	626	81	0.114

Figure 5 is an example that illustrates the model's incorrect prediction for an utterance that has been summarized by the GPT. This example is the ChatGPT summaries of the utterance with id 113247-00023 in the training dataset. In contrast to the original utterance, ChatGPT cannot retain the critical content in the utterance that can be used to predict the stance of the utterance, and even experts cannot also identify the stance of the utterance from the summary. This may be the reason why the prediction accuracy of the proposed model decreases for summarized utterances.

16番、杉本です。日本共産党議員団を代表し、議案第11号平成22年度三芳町一般会計予算案に対し、[STANCE]討論を行います。- 日本の経済状況は深刻で、デフレ状態が続いている。
- 平成22年度の三芳町予算は前年比でマイナス成長となり、地方債の増加が主な要因となっている。
- 特徴的な事業として、子ども手当の増加や地方財政の改革が挙げられる。
- 予算編成により、削減や廃止が行われ、正規職員の削減には懸念がある。
- 日本共産党議員団は、計画的な行財政運営や公共施設の管理委託、待機児童対策などの提案・要望を行っている。以上、町長並びに当局の一層の取り組みを要請をして、[STANCE]討論といたします。

Figure 5: An Example of the Abstract of an Utterance summarized by the ChatGPT.

5 CONCLUSIONS

This paper demonstrated our approach to tasks related to stance classification. We designed two frameworks for determining politicians' stances using utterances, with the GPT-based model achieving an accuracy of 0.932. The findings indicate that text processed through summarization still cannot fully replace the original text; moreover, summarization affects the accuracy of the GPT-based model. For text that expresses meaning subtly or indirectly, language models have difficulty comprehending the true intent of the text, leading to the false determination of opposite stances.

Therefore, several works for enhancing the performance of the proposed frameworks could be developed in the future. First, we aim to experiment with other QA models as the judgment model. As Curie is not an open-source model, our training frequency and method were considerably limited. With other new open-source QA models based on different LLMs available, we can explore the performance of judgment models based on alternative QA models. Second, ChatGPT—adopted for generating summaries in this study—was not fine-tuned concurrently in the training process of the judgment model. However, the error analysis revealed a significantly higher error rate in the summarized text compared to that in the un-summarized text. Hence, fine-tuning may potentially improve the performance of the summary generation model and consequently enhance the accuracy of the stance determination model. Third, some ways of expressing opinions in utterances are less common. For instance, because of the serious problems that would arise if the bill was passed, the politician was disapproved of the bill. While these expressions are not common, they are not rare and still constitute a certain proportion of the dataset. Therefore, enabling the model to recognize such utterances may help improve the judgment accuracy of the model.

Acknowledgment

This study was partially supported by the Ministry of Science and Technology, under the grant MOST 110-2511-H-992-003-MY3. It was also financially supported by the "Institute for Research Excellence in Learning Sciences" of National Taiwan Normal University (NTNU) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

REFERENCES

[1] Ogawa, Y., Kimura, Y., Shibuki, H., Ototake, H., Uchida, Y., Takamaru, K., Kadowaki, K., Akiba, T., Sasaki, M., Kobayashi, A., Yoshioka, M., Mori, T., Araki, K., Sekine, S., & Mitamura, T. 2023. Overview of the NTCIR-17 QA Lab-PoliInfo-4 Task. In *Proceedings of the 17th NTCIR Conference*. https://doi.org/10.20736/0002001326

[2] Kimura, Y., Shibuki, H., Ototake, H., Uchida, Y., Takamaru, K.,Ishioroshi, M., Mitamura, T., Yoshioka, M., Akiba, T., Ogawa, Y.,Sasaki, M., Yokote, K., Mori, T., Araki, K., Sekine, S., & Kando, N. 2020. Overview of the NTCIR-15 QA Lab PoliInfo Task. In *Proceedings of The 15th NTCIR Conference*,101-112.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K . 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 1, 4171-4186.

[4] Hochreiter, S., & Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8), 1735-1780.

[5] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G.,Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I ., & Amodei, D. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.

[6] OpenAI. 2023. *Models-OpenAI API*. https://platform.openai.com/docs/models/overview

[7] OpenAI. 2023. *Introducing ChatGPT*. https://openai.com/blog/chatgpt

[8] Kato, R., & Sasaki, M. 2020. Ibrk at the NTCIR-15 QA Lab-PoliInfo-2. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologie*s,172-174.

[9] Sasazawa, Y., & Okazaki, N. 2020. WER99 at the NTCIR-15 QA Lab-PoliInfo-2 Classification Task. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*, 124-129.

[10] Atsumi, K., & Kano, Y. 2020. knlab Team: NTCIR-15 QA Lab-PoliInfo-2 Stance Classification Task. In *NTCIR 15 Conference: Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*, 113-115.

[11] Maćkiewicz, A., & Ratajczak, W. 1993. Principal components analysis (PCA). *Computers & Geosciences*, 19(3), 303-342.

[12] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. 2017. LightGBM: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 3149-3157.

[13] Nekomoto, T., Shirato, D., Ohsugi, R., Akiba, T., & Masuyama, S. 2020. akbl at the NTCIR-15 QA Lab-PoliInfo-2 Task. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologiess*, 153-159.

[14] Cornegruta, S., Bakewell, R., Withey, S., & Montana, G. 2016. Modelling radiological language with bidirectional long short-term memory networks. *arXiv preprint arXiv:1609.08409.*

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30, 5998-6008.

[16] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. 2018. Improving language understanding with unsupervised learning. https://openai.com/research/language-unsupervised. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[17] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. 2019. Better language models and their implications. https://openai.com/research/better-language-models https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf