Information Engineering Express
International Institute of Applied Informatics
2022, Vol. 8, No. 1, IEE635

Recognizing a Participant Behavior in a Multi-party Conversation: Detection of a Participant That Returns to a Discussion That Is Already Over

Shunsuke Yonemitsu *, Kazutaka Shimada *

Abstract

In this paper, we discuss a task to recognize a participant's behavior in a multi-party conversation. We focus on a detection task of a participant that returns to a discussion that is already over; we call the participant "BDH (Beat a Dead Horse) participant." The target corpus contains 17 conversations about a topic with three participants, and one of three participants is the BDH participant. To detect the BDH participant, we apply machine learning methods. We compare three machine learning methods; naive bayes, decision tree, and support vector machines. In addition, we introduce a selection model based on the task setting. The experimental result shows the effectiveness of SVMs with our selection model.

Keywords: Discussion Support, Multi-Party Conversation, Role Prediction

1 Introduction

Discussions and meetings are frequently held to make a decision and to resolve various problems. Since the time for discussions is not unlimited, discussions should proceed smoothly. However, they often stagnate by various factors caused by participants and agendas. There are two approaches to a support system for smooth discussions. Some researchers have proposed supporting systems of the discussion during a meeting [1, 2]. One benefit of supporting discussions in real-time is that such systems can support consensus-building and immediately detect behaviors that negatively influence the discussion's progress. However, if participants depend on real-time support, it might be impossible to discuss smoothly without such supports. The other approach focuses on supports and analysis after meetings [3]. Although it is impossible to support discussions in real-time, it can help analyze their behaviors themselves objectively. It leads to upskilling about discussion for participants.

In this study, we focus on supports after discussion for improving participants' behaviors. Figure 1 shows an overview of our purpose. As the final goal, we generate some feedback to participants by using our system: e.g., the trigger utterance generation in the current meeting from the summary of the previous meeting, feedback generation from behavior analysis, and feedback generation from statistics of the previous meeting.

* Kyushu Institute of Technology, Fukuoka, Japan

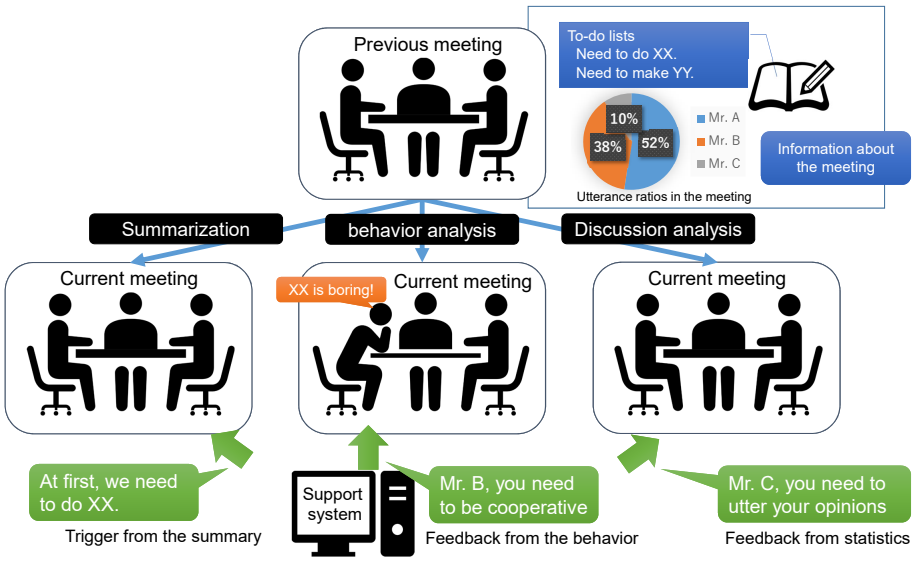

Figure 1: The goal of our system. Our motivation is to generate active and productive discussions with our system. In this paper, we focus on behavior analysis and stagnation situation in discussions.

In this paper, we discuss stagnation that appears in discussions. There are several definitions of the stagnation of discussions. Kodani et al. [4] have mentioned that utterances with disagreement lead to stagnation of discussions. Mizukami et al. [5] have regarded conditions that have not reached a solution due to out of ideas as stagnation of discussions. Yonemistsu and Shimada [6] have introduced a task to recognize a participant who returns to a done-deal in the previous discussion as a stagnation of the discussion. They call a participant in a conversation who returns to a done-deal "BDH (Beat a Dead Horse) participant." In this paper, we also deal with the BDH participant recognition task.

The previous study [6] utilized only Support Vector Machines (SVMs) for the task. However, it is not clear whether SVMs are the best approach. In this paper, we compare three machine learning methods; not only SVMs but also naive bayes and decision tree. In addition, they did not employ the task setting for the recognition task. In their corpus, each discussion consists of utterances with three participants, and then one of three participants is the BDH participant. We introduce a selection model based on the discussion setting, namely just one BDH participant in one discussion. The contributions of this paper are as follows:

- Comparison of three machine learning methods to investigate the best model,
- Introduction of a selection model to the method to improve the accuracy.

2 Related Work

Benne et al. [7] have defined participants' roles as functional roles in group work, including discussions. The three main roles in the paper were task roles, social/maintenance roles, and dysfunctional roles. The three roles were subdivided into 26 roles. Li et al. [8] have

proposed a method for predicting a key speaker in meeting speech based on multiple features optimization. Hung et al. [9] have proposed a method for predicting the dominant person in a group meeting by using speaker diarization strategies. Sanchez-Cortes et al.[10] have proposed a method for identifying emergent leaders in small groups by using voice and video data. Rienks et al. [11] have reported a system that determines the dominance level of meeting participants to help each participant. Shiota et al. [12] have analyzed the characteristics of facilitators in multi-party conversations from a macro viewpoint. Muller et al.[13] have proposed a method for detecting leadership from non-verbal behaviors and evaluated it with a cross-dataset situation. Beyan et al.[14] have defined two types of leadership styles: autocratic leadership and democratic leadership. They reported that non-verbal features provided good prediction rates for some classification approaches and analyzed a correlation between the features and the results of social psychology questionnaire. These studies focused on positive aspects and influence in discussions. On the other hand, we focus on negative influence in discussions, such as dysfunctional roles defined by Benne et al.

Zhang et al. [15] have discussed the functional roles of the participants in group discussion and reported the results of the analysis of the relationship between communication skill impression and functional roles. In their work, they defined passive participants based on dysfunctional roles defined by Benne et al. The passive participants denote persons that rarely speak in the discussion. In this study, our target, namely BDH participants, is a person that leads to the stagnation of a discussion. BDH participants are non-cooperative and prolong the discussion. However, BDH participants is not always a passive participant because they probably utter negative opinions about a done-deal in the previous discussion from the definition of BDH. BDH participants are similar to "blocker" and "deserter" in dysfunctional roles by Benne et al. "blocker" is a role in disturbing the task progression and "deserter" is a role in leading to an off-topic conversation.

Cheng et al. [16] have analyzed user's behaviors that are banned from an online community and proposed a method for identifying antisocial users. Seah et al.[17] have proposed a method for detecting a troll that sows discord in an online forum. Cheng et al. [18] have discussed the cause of trolling behavior in discussion communities. They reported that negative moods and seeing troll posts increased the probability of trolling. These studies focused on offensive or aggressive persons. A participant that mentions an offensive comments often causes the stagnation of a discussion. However, BDH participants are not always offensive, although they tend to utter negative opinions about a done-deal. In other words, we need to consider many situations and behaviors.

3 Target Data

We use the BDH corpus by [6]. It is a text-based chat corpus. The corpus contains 17 Japanese conversations about two topics. The average number of utterances in one conversation is approximately 50 utterances. Table 1 shows an example. Timestamp in the table denotes the number of seconds from the starting time of the conversation. U-ID and P-ID in the table are the utterance ID (U1, ..., U_n) and participant ID (A, B, and C). In this example, assume that the participant B is the BDH participant, and the A and C are normal participants. The done-deals in the previous conversation are a coat, a tent, and a knife. The participant B utters a negative opinion about the coat in U19. On the other hand, the other participants are trying to keep the done-deal and the discussion, such as U20 and U26.

Table 1: An example of a conversation: the topic is “a list of necessary supplies on a desert island,” and the done-deals in the previous conversation are a coat, a tent, and a knife. The participant B is the BDH participant, namely our target in our recognition model. This table is cited from the previous study [6].

Timestamp	U-ID	P-ID	Utterance
…	…	…	…
288.8	U14	B	To be frank, I want to take food, not searching on the island. (食べ物は探すよりも素直に持っていきたいけどな)
290.1	U15	A	Do you want to bring emergency food? (非常食を持っていくってこと？)
295.1	U16	B	Yep. (そういうこと)
309.7	U17	A	It may be good. (割とありかも)
334.8	U18	A	Food can be solved with this one. (食料はこれ 1 つで解決できそう)
343.0	U19	B	I want water. The temperature seems quite hot and I want to stop taking the coat. I want water. (水も確保したいしさ、気温もかなりあついみたいだし コート持っていくのやめて水の確保がしたい)
370.6	U20	A	But if it rains it’s surprisingly cold. (でも雨が降ったりしたら意外と寒いと思うよ)
388.3	U21	A	I think it’s better to bring a coat. (コートはあった方がいいと思う)
395.2	U22	C	I want to use a coat instead of a blanket (コートは布団代わりに使いたい)
411.9	U23	B	It rains only three or four days a month and there is a tent, right? (一か月に 3,4 日しか降らないしテントもあるじゃん？)
425.7	U24	B	I think it’s riskier to go outside (むしろ外に出ていくほうがリスク高いと思うんだよね)
462.7	U25	B	I don’t even need a knife. (なんならナイフもいらないまである)
519.0	U26	A	This is a discussion of what to bring as an additional supply. Now, let’s consider not cutting other belongings. (そもそもあと 1 つ何持っていくかの議論だし、 現状は他の持ち物を削らない方向で考えよう)
533.7	U27	B	But is the coat completely useless in terms of this climate? (でもコートとか気候的に完全に無駄じゃん？)
540.3	U28	A	Anyway, let’s select the candidates that we should bring. (取り敢えず持っていきたいものの候補を絞ろう)
…	…	…	…

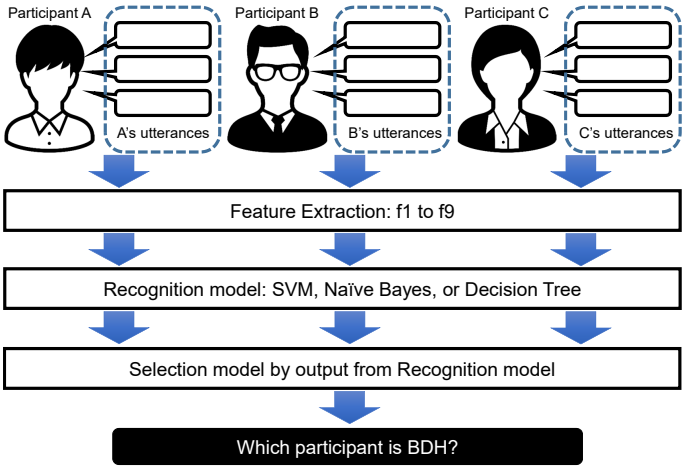

Figure 2: Overview of our recognition model about BDH participants. Our model classifies each participant into BDH or not by using features first. Then, it selects one BDH participant from them. One of our contributions is to introduce the selection model into the recognition model.

4 Method

In this section, we explain our recognition model. Figure 2 shows the outline of our method. The task is to recognize the BDH participant in each discussion. First, we explain features for machine learning methods. These features were proposed by the previous study [6]. Next, we explain three machine learning methods with our selection model through the reconfiguration of the task setting.

4.1 Features

We extract features for the model from each participant's utterances in the conversation. This process is the same as the related work by [6]. They proposed nine types of features as follows:

$f1$) **Ratio of own utterances**
They assumed that BDH participants tend to mention negative opinions frequently in the conversation. As a result, the ratio of utterances of BDH participants becomes larger than others.

$f2$) **Standard deviation of intervals between own utterances**
As mentioned in $f1$, BDH participants tend to mention negative opinions about done-deals because of the task setting. On the other hand, BDH participants do not contribute to the main subject of the conversation, e.g., selection of an additional supply in Table 1. Therefore, they assumed that the utterances of BDH occur sporadically. The standard deviation of intervals of utterances of each participant is computed from timestamps in the corpus.

$f3$) **Ratio of own consecutive utterances**
They defined consecutive utterances by one participant "block." For example, U23 to

U25 in Table 1 are a block by the participant B. As mentioned in $f2$, BDH participants tend to mention many utterances in a short time. As a result, their utterances tend to occur sequentially. On the other hand, normal participants tend to speak after listening attentively. Therefore, they assumed that the ratio of consecutive utterances of BDH differs from that of normal participants. The feature is computed as follows:

$$f3 = \frac{\text{\# of own blocks}}{\text{\# of own utterances}} \tag{1}$$

$f4$) **Ratio of own utterance blocks**
In a similar way to $f3$, the ratio of own blocks is computed as follows:

$$f4 = \frac{\text{\# of own blocks}}{\text{\# of blocks in the conversation}} \tag{2}$$

$f5$) **Standard deviation of frequencies of the most frequent five words**
Although normal participants utter their opinions about the main subject in the conversation, the BDH participant in the conversation speaks about done-dales that he/she wants to return to. In other words, the distribution of words between BDH and normal participants becomes different. This feature is based on the characteristic.

First, the top-five frequent words in the conversation are extracted. Next, the method counts the number of times that the words appeared in the utterances of each participant. Then, the standard deviation is computed as the feature.

$f6$) **Number of words that are used in the most frequent five words**
In a similar way to $f5$, for the top-five words in the conversation, the method counts the number of words that each participant used in the conversation.

$f7$) **Number of vocabulary words in own utterances**
As mentioned in $f5$, the distribution of words between BDH and normal participants is probably different. Since the purpose of the BDH participants is to return to a done-deal in the previous conversation, the words that he/she uses in the conversation focus on the done-deal. Therefore, they assumed that the size of the vocabulary of BDH becomes smaller than that of normal participants. The method counts the number of words that each participant used in the conversation as the feature.

$f8$) **Frequencies of n-grams in own utterances**
Surface expressions are one of the essential features for text classification tasks. It is natural to introduce n-gram features to the task. Therefore, three types of n-grams in the following conditions are extracted.

* $f8$ uni-gram: we extract nouns, verbs, adjectives, and adverbs in conversations.
* $f8$ bi-gram: top-five bi-grams about the BDH participants
* $f8$ tri-gram: top-ten tri-grams about the BDH participants

$f9$) **Ratio of negative utterances**
BDH participants tend to use negative expressions due to their purpose. Therefore, they assumed that the distributions of negative expressions between BDH and normal participants become different. The numbers of negative utterances for each participant in the conversation are computed as the feature. In addition, the ratio of the

negative utterances is computed as the feature. For the scoring process, the word polarity dictionary reported by [19] is used. Each word in the dictionary possesses a polarity score within [-1, +1]. If a word is strongly negative, the score is close to -1. If a word is strongly positive, the score is close to +1. The method computes the polarity score of each utterance on the basis of the summation of the word scores in the utterance.

$$PNscore(u_i) = \sum_{w_i \in u_i} w_i \tag{3}$$

where u_i is the target utterance, and w_i denotes the polarity score of a word in [19]. Then, the value of each utterance is determined on the basis of $PNscore(u_i)$ as follows:

$$NegU(u_i) = \begin{cases} 1 & (PNscore(u_i) \leqq 0) \\ 0 & (Otherwise) \end{cases} \tag{4}$$

Finally, the final score, namely the ratio of negative sentences, is computed.

$$Final(p_j) = \frac{\sum_{u_i \in p_j} NegU(u_i)}{n_{p_j}} \tag{5}$$

where p_i denotes a participant and n_{p_j} is the number of utterances of p_j.

4.2 Proposed method

Yonemitsu and Shimada [6] reported the accuracy of SVMs and the effectiveness of each feature. Although SVMs are a robust classifier, it is not clear yet that SVMs are the best classifier for the task. In this paper, we compare three machine learning models; SVMs, Naive Bayes, and Decision Tree. We implement these models by scikit-learn[1]. The parameters of these models are default parameters (the kernel of SVMs is "linear.")

In this section, we introduce a selection model as the post-processing of the machine learning models. The BDH corpus that we handle in this paper consists of utterances with two normal participants and one BDH participant. In other words, the task is to recognize one BDH participant in each discussion. However, the previous model by [6] just handled that it is a classification task; each participant is BDH or normal.

We reconfigure this task; which participant is the most likely BDH participant in each discussion. For this purpose, we regard the output value of each machine learning model as a confidence measure. We compare the output values of each participant and then select the participant with the highest value as the BDH participant in the discussion. For example, assume that the values of participants A, B, and C from SVMs are 0.7, 0.8, and -0.3, respectively. In this situation, the model of the previous work selects the participants A and B as the BDH participant because the two values are positive. On the other hand, our model selects only the participant B by the selection model[2].

5 Experiment

We evaluated our method explained in Section 4.2 with the BDH corpus by [6]. The corpus contains 17 conversations with three participants (one BDH participant in them). The results

[1] https://scikit-learn.org/stable/index.html

[2] In fact, we obtain the output values by the "predict_proba" option in scikit-learn.

Table 2: Experimental result of the three models without the selection model.

Model	Class	Precision	Recall	F-score	Average F
SVM	BDH	0.667	0.353	0.462	0.639
	Normal	0.738	0.912	0.816	
Naive Bayes	BDH	0.250	0.176	0.207	0.446
	Normal	0.641	0.735	0.685	
Decision Tree	BDH	0.562	0.529	0.545	**0.664**
	Normal	0.771	0.794	0.783	

Table 3: Comparison with the selection model. The result of the Decision Tree method at the top row is the best result in Table 2, namely the best method without the selection.

Classifier	Selection	Class	Precision	Recall	F-Score	Average F
Decision Tree	NO	BDH	0.562	0.529	0.545	0.664
		Normal	0.771	0.794	0.783	
SVM	YES	BDH	0.588	0.588	0.588	**0.691**
		Normal	0.794	0.794	0.794	
Naive Bayes	YES	BDH	0.294	0.294	0.294	0.471
		Normal	0.647	0.647	0.647	
Decision Tree	YES	BDH	0.529	0.529	0.529	0.647
		Normal	0.765	0.765	0.765	

in this experiment are based on conversation-level cross-validation for 17 conversations; one conversation as the test data and the others as the training data, and iteration of the 17 combinations.

First, we compared three machine learning methods without the selection model. Table 2 shows the Precision, Recall, F-score, and Averaged F-score of each method without the selection model. The bold score denotes the best average F-score. As the result mentioned, the best method was the Decision Tree method. The method outperformed the method reported in [6], namely SVMs without the selection.

Next, we evaluated three methods with the selection model. In this situation, the best method was SVMs with the selection model. The method outperformed the best F-score without the selection (0.691 vs. 0.664). This result shows the effectiveness of our selection model. Although the F-score of normal participants was relatively high, that of BDH participants was not always sufficient. The reason is that the numbers of BDH and normal participants were not balanced; BDH : normal = 1 : 2.

Recently, the state-of-the-art methods in various classification tasks are usually based on deep learning approaches. Thus, we also applied Long Short-Term Memory (LSTM) [20] that is an artificial recurrent neural network (RNN) architecture, to the task. We used the output of BERT [21] that is a pre-training model, as the input of LSTM. In other words, we input each utterance into BERT, and then we obtained the embeddings from BERT. We handled the [CLS] token of the 11th layer on BERT as the input embeddings of LSTM, and then the method learned the model in the same manner as mentioned above, namely conversation-level cross-validation for 17 conversations. However, the F-score was extremely low (less than the chance rate). The reason is that the size of the BDH corpus is not suitable for deep learning approaches, namely a small dataset.

6 Conclusions

In this paper, we focused on the stagnation of discussions for developing a discussion support system. The task is to recognize a participant who returns to a done-deal in a multi-party conversation. We called such a participant "BDH (Beat a Dead Horse) participants."

Although the previous work reported the accuracy of an SVM-based method, it was not clear whether the method is the best way for the task. Therefore, we compared three methods; SVMs, Naive Bayes, and Decision Tree. As a result, the method based on Decision Tree outperformed the SVM method on the average F-score (0.664 vs. 0.639).

We also introduced a selection model to the method. By using the selection model, we obtained a higher F-score as compared with the Decision Tree method without the selection model (0.691 vs. 0.664). In addition, we applied an LSTM-BERT method to the task. However, the accuracy was lower than SVMs with our selection model due to the small dataset. To apply deep learning methods appropriately, scaling up the corpus is one of the most important future work. These results show the effectiveness of our proposed method in the BDH corpus.

The final goal of our study is to detect several problems in conversations by our support system and then offer some feedback to participants for productive discussions/conversations. For the purpose, we analyzed the behaviors of BDH participants. However, the behaviors depended on each participant; e.g., a BDH participant generated negative responses and a BDH participant craved neat explanations for the done-deal. We need to investigate the behaviors of each BDH participant more deeply. We are also investigating the behaviors of facilitators in other papers [12, 22]. Multiple analysis of them is important future work.

In this paper, we handled text-based conversations. However, for the behavior recognition, non-linguistic information is also important. Shiota et al. [23] have reported the importance of multimodal information in a conversation analysis task. The construction of a multimodal corpus for the BDH participant recognition task is an interesting research issue. In addition, the BDH behavior also appears in various situations. Recognizing such behavior is also useful for other tasks, such as discussion quality assessment in multi-party conversation [24].

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 20K12110.

References

[1] F. Nihei, Y. Hayashi, and Y. Nakano. Detecting discussion state shifts in group discussions. The 28th Annual Conference the Japanese Society for Artificial Intelligence, 2014.

[2] R. Kirikihira and K. Shimada. Discussion map with an assistant function for decision-making: A tool for supporting consensus-building. In *International Conference on Collaboration Technologies*, pages 3–18, 2018.

[3] K. Okamoto, S. Matsubara, and K. Nagao. Automatic evaluation of statements in meetings based on acoustic and linguistic features. In *The 78th National Convention of IPSJ*, number 1, pages 521–522, 2016.

[4] T. Kodani, K. Keki, T. Matsui, and T. Okamoto. Development of discussion supporting system based on the "value of favorable words' influence". In *Transactions of the Japanese Society for Artificial Intelligence*, volume 19, pages 95–104, 2004.

[5] E. Mizukami, L. Liu, and M. Ikuyo. An analysis of participants' behavior in the transition from a stagnant phase of discussion in group discussions. *SIG-SLUD*, B5(03):50–55, 2018.

[6] S. Yonemitsu and K. Shimada. Don't beat a dead horse: Recognizing a person who returns to a done-deal in a multi-party conversation. In *Proceedings of the 8th International Conference on Smart Computing and Artificial Intelligence*, 2020.

[7] KD. Benne and P. Sheats. Functional roles of group members. In *Journal of social issues*, volume 4, pages 41–49, 1948.

[8] W. Li, Y. Li, and Q. He. Estimating key speaker in meeting speech based on multiple features optimization. In *International Journal of Signal Processing, Image Processing and Pattern Recognition*, volume 8, pages 31–40, 2015.

[9] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez. Estimating the dominant person in multi-party conversations using speaker diarization strategies. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2197–2200, 2008.

[10] D. Sanchez-Cortes, O. Aran, MS. Mast, and D. Gatica-Perez. A nonverbal behavior approach to identify emergent leaders in small groups. In *IEEE Transactions on Multimedia*, volume 14, pages 816–832, 2011.

[11] R. Rienks and D. Heylen. Dominance detection in meetings using easily obtainable features. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 76–86, 2005.

[12] T. Shiota, T. Yamamura, and K. Shimada. Analysis of facilitators' behaviors in multiparty conversations for constructing a digital facilitator system. In *International Conference on Collaboration Technologies*, pages 145–158, 2018.

[13] PM Muller and A Bulling. Emergent leadership detection across datasets. In *2019 International Conference on Multimodal Interaction*, pages 274–278, 2019.

[14] C Beyan, F Capozzi, C Becchio, and V Murino. Prediction of the leadership style of an emergent leader using audio and visual nonverbal features. volume 20, pages 441–456. IEEE, 2017.

[15] Q. Zhang, H. Hung, S. Kimura, S. Okada, N. Ohata, and K. Kuwahara. Analysis on the participants' functional roles and their transitions in group discussion. In *The transaction of Human Interface Society*, volume 20, pages 31–44, 2018.

[16] J Cheng, C Danescu-Niculescu-Mizil, and J Leskovec. Antisocial behavior in online discussion communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, 2015.

[17] CW Seah, HL Chieu, KMA Chai, LN Teow, and LW Yeong. Troll detection by domain-adapting sentiment analysis. In *2015 18th International Conference on Information Fusion (Fusion)*, pages 792–799, 2015.

[18] J Cheng, M Bernstein, C Danescu-Niculescu-Mizil, and J Leskovec. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1217–1230, 2017.

[19] H. Takamura, T. Inui, and M. Okumura. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL2005)*, pages 133–140, 2005.

[20] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[21] J. Devlin, M-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[22] T. Sembokuya and K. Shimada. Timing prediction of facilitating utterance in multi-party conversation. In *Computational Linguistics, Communications in Computer and Information Science*, volume 1215, pages 267–279, 2020.

[23] T. Shiota, K. Honda, K. Shimada, and T. Saitoh. Leader identification using multimodal information in multi-party conversations. In *Proceedings of the International Conference on Asian Language Processing (IALP)*, pages 7–12, 2020.

[24] T. Shiota and K. Shimada. The discussion corpus toward argumentation quality assessment in multi-party conversation. In *Proceedings of the 9th International Conference on Learning Technologies and Learning Environments*, 2020.