Tweet Review Mining focusing on Celebrities by Machine Reading Comprehension based on BERT

Yuta Nozaki, Kotoe Sugawara, Yuki Zenimoto, Takehito Utsuro
Degree Programs in Systems and Information Engineering,
Graduate School of Science and Technology, University of Tsukuba,
1-1-1, Tennodai, Tsukuba, Ibaraki, 305-8573, Japan
{s2020776,s2120736,s2220753}_@_s.tsukuba.ac.jp, utsuro_@_.iit.tsukuba.ac.jp

Abstract

In this paper, we propose a method of mining tweets that represent reviews on celebrities by the machine reading comprehension model based on BERT. The purpose of this paper is to collect and aggregate reviews on tweets for celebrities in order to support the activities of fans of celebrities searching for information on celebrities' criticisms and impression trends on the Web. Specifically, we focus on the celebrities' names and adjective expressions that appear in tweets, and determine whether or not there is a relationship such as review and impression between the two words using the BERT machine reading comprehension framework. In the machine reading comprehension framework, the tweet is the context, the celebrity's name is the question, and an adjective that has a relationship such as review and impression with the celebrity is the answer. As the result of the evaluation experiment, the machine reading comprehension framework achieved fairly reliable performance.

1 Introduction

The purpose of this paper is to support the process by which TV drama viewers and celebrity fans search for information on critiques and interest trends in celebrities on the Web. In this paper, we propose a method for mining reviews of popular celebrities using tweets as the information source. Usually, when an event or an incident that is related to a popular celebrity occurs, a large number of tweets are posted, where, in those tweets, people express their own thoughts in the way they like. In Twitter, however, there exist no rule on the grammatical correctness of the posted tweets. Thus, this makes it unexpectedly difficult to correctly identify what people actually intend to express in their tweets, mainly due to the lack of grammatical correctness of tweet sentences (Sanguinetti et al., 2020). And, in the NLP community, in the case of other applications such as tweet sentiment analysis, it is quite common to avoid grammatically parsing tweet sentences, but to directly analyze sentiment of tweets (e.g, Nakov et al. (2016)).

Considering such a background, in this paper, we apply the BERT framework of machine reading comprehension (Figure 1) to tweet sentences, so as to detect the span of impression relations from text showing the review relations of celebrities. Specifically, first we collect tweets in which the name of the celebrity is included. We develop a dataset of tweets that are annotated with the span of adjectives indicating review relationships for the celebrity names co-occurring in the tweets. Next, in a machine reading comprehension framework, we train the BERT model (Devlin et al., 2019) to predict the spans that indicate reviews on previously unobserved celebrities.

Machine reading comprehension frameworks are often used in question answering systems. For example, Huang et al. (2020) proposed a machine reading comprehension model aimed at answering questions from Twitter, which is full of noisy, informal text. Also, Guo et al. (2021) proposed a model to detect spans mentioning health-related information from tweets. Xu et al. (2019) proposed

a Review Reading Comprehension task that predicts spans from review text that are answers to user questions.

The model proposed in this paper, which is based on a machine reading comprehension framework, uses the keyword Q of the celebrity name as the question, and the tweet in which the keyword Q and the adjective A co-occur as the context C. If the review of the celebrity name Q is expressed by an adjective in the tweet, the corresponding adjective A is the output. If the adjective indicates an opinion on the celebrity Q, the corresponding adjective A is also the output. And, if the adjective does not indicate an opinion, "Not Review" is the output.

Specifically, first we collect tweets in which the name of the celebrity is included. We develop a dataset of tweets that are annotated with the span of adjectives indicating review relationships with the celebrity names co-occurring in the tweets. Next, in a machine reading comprehension framework, we train the BERT(Devlin et al., 2019) model to predict the spans that indicate reviews of unknown celebrities.

In the evaluation, first we develop the dataset as follows. In the dataset, for each tweet that mentions a specific celebrity, we annotate whether the tweet expresses an opinion on the celebrity by means of an adjective. First, we collect tweets that contain the name of a specific celebrity for a certain period of time. The tweets that contain adjectives with a certain frequency are randomly selected to be included in the dataset. Then, to each candidate tweet, it is assigned whether there is a review relationship between the celebrity name and the adjective, and the review mining dataset of 1,500 tweet instances is developed.

Using the developed dataset, we trained and evaluated the machine reading comprehension model (Figure 1) with BERT, where the results of comparison with the token classification model (Figure 2) with BERT as well as the classification model by SVM show that the machine reading comprehension model with BERT achieved the best performance. Especially, we further study to measure the performance of detecting spans that indicate the review relationship of previously unobserved celebrities. In this evaluation, it can be concluded that the machine reading comprehension model trained with tweets including previously observed celebrities' names is also effective in detecting the review relationship with previously unobserved celebrity names. This is quite contrastive in that the token classification model is not capable of detecting the review relationship with previously unobserved celebrity names.

We also investigate the effect of the number of adjectives co-occurring within a tweet on the performance of the machine reading comprehension model. From this analysis, it is shown that, in the case where the number of co-occurring adjectives is one, the performance of review relation detection is much higher than the case where the number of co-occurring adjectives is two or more. However, even in the case where the number of co-occurring adjectives is two or more, its precision is over 0.5, which is more or less satisfactory performance and this result confirms the effectiveness of the proposed approach.

2 Developing the Review Mining Dataset

2.1 Collecting Tweets

In this paper, we collected tweets on 5 celebrities[1] who are very popular in Japan. We collected tweets on each celebrity to develop the dataset of candidate tweets. First, we use the Twitter Search API[2] to collect tweets that contain each celebrity's name as the keyword, where the numbers of collected tweets are shown in Table 1[3]. Then, with the results of morphological analysis by JUMAN++[4] on the collected tweets, we transform each adjective into its representative notation (base form) and obtain the frequency statistics of the adjectives that co-occur with each celebrity name. Next, for each celebrity name, the most frequent 30 adjectives are used to collect candidate tweets for developing the dataset. Specifically, 10 tweets were randomly selected for each

[1]Satomi Ishihara, Haruma Miura, Masato Sakai, Yuko Takeuchi, and Ryoko Yonekura,

[2]https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets

[3]For roughly around the period from July to October, 2020, the numbers of collected tweets (not including retweets) are 555,701 for Satomi Ishihara, 130,734 for Haruma Miura, 61,108 for Masato Sakai, 33,7672 for Yuko Takeuchi, and 14,869 for Ryoko Yonekura.

[4]http://nlp.ist.i.kyoto-u.ac.jp/?JUMAN++

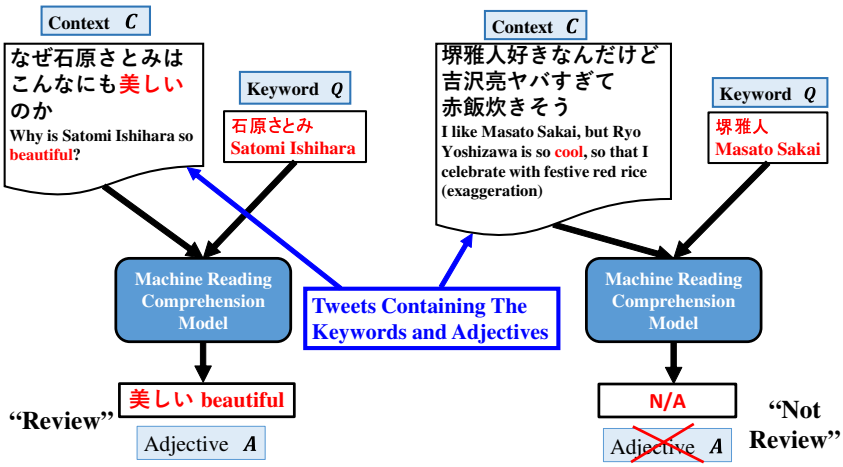


Figure 1: Framework of Machine Reading Comprehension Model for Mining Reviews on Celebrities represented by Adjectives

adjective. Then, 1,500 tweets in total are used as candidate tweets for developing the dataset.

2.2 Distribution of Adjectives per Celebrity

For each of the three celebrity names "Satomi Ishihara", "Masato Sakai", and "Ryoko Yonekura", Figure 3 shows the statistics of the frequencies of the most frequent 30 adjectives. More or less general adjectives that are common among celebrity names such as "good" and "amazing" are observed in the highest ranks for "Satomi Ishihara" and "Masato Sakai". Other adjectives, however, are mostly specific to each celebrity name. Those celebrity specific ones sometimes represent events closely related to each celebrity such as the marriage of "Satomi Ishihara". Those celebrity specific ones include "envy", "congratulation", and "happy" for "Satomi Ishihara", "reliable", "painful", and "unique" for "Masato Sakai", "noisy", "close" and "annoying" for "Ryoko Yonekura"[5].

[5]In the evaluation, we apply the pre-trained BERT machine reading comprehension model, which helps to avoid overfit to adjectives seen in fine-tuning. Actually, we confirmed in the supplementary evaluation we omit due to the space restriction in this paper, that the model is capable of answering adjectives unseen in fine-tuning in the machine reading comprehension framework of this paper.

2.3 Criteria on Annotating Reviews on Celebrities represented by Adjectives

For the selected candidate tweets, we develop the review mining dataset by annotating whether or not the tweets show review relations between the celebrity name and adjectives. The results of annotating 1,500 candidate tweets based on the criteria can be divided into the following three classes.

- "Review" ··· For the keyword celebrity name, only one adjective in the tweet indicates the review relationship with the keyword celebrity name. Here, the adjective does not have to be the one specified at the time of tweet collection.
- "Not Review" ··· With the keyword celebrity name, none of the adjectives in the tweet show a review relationship[6].
- Others ··· With the keyword celebrity name, two or more adjectives in the tweet show a review relationship.

Among the 1,500 tweets[7], the tweets corresponding to others are excluded, and the same number of

[6]Examples of "Not Review" include tweets that indicate a review relationship with a celebrity other than the keyword celebrity or a TV drama in which the celebrity appears, and tweets that indicate a review relationship with a part of speech other than adjectives.

[7]As a result, as shown in Table 1, the number of tweets in "Review" totaled 636 of which 112 for Satomi Ishihara, 147 for Haruma Miura, 131 for Masato Sakai, 126 for Yuko Takeuchi, 120 for Ryoko Yonekura. The number of tweets in "Not Review" totaled 864 of which 188 for Satomi Ishihara, 153 for

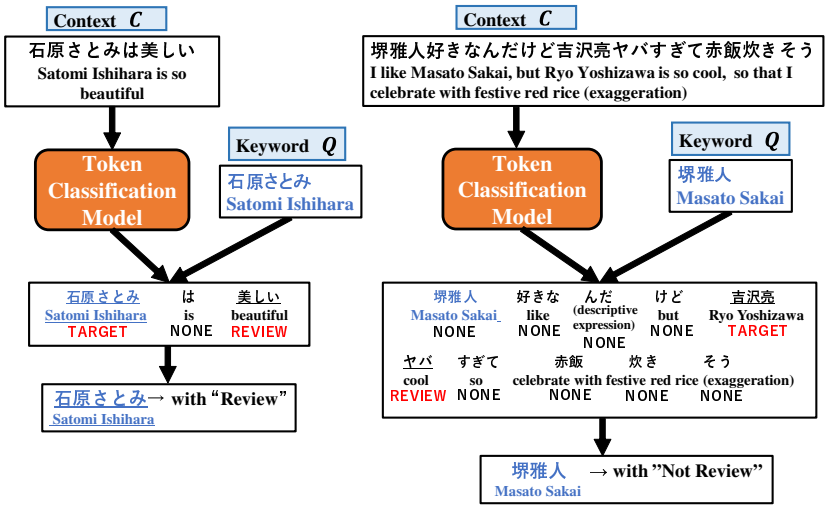


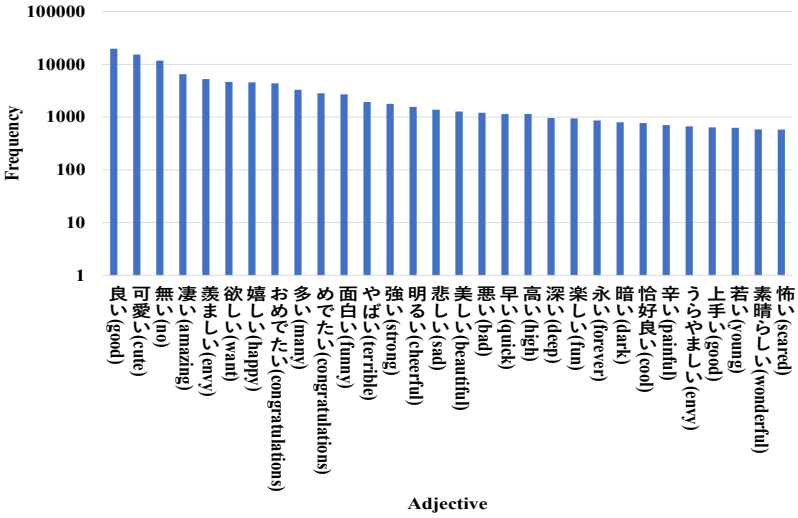
Figure 2: Framework of Token Classification Model for Mining Reviews on Celebrities represented by Adjectives

Table 1: Numbers of Collected Tweets and those for Evaluation

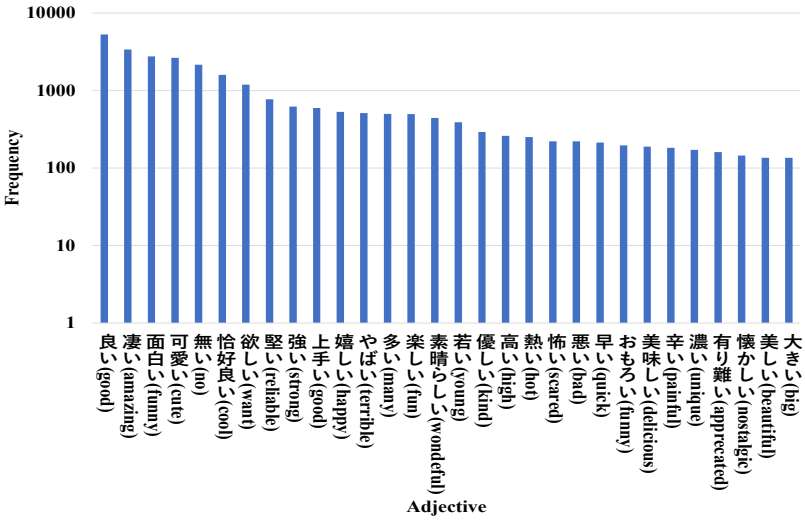
celebrity name	number of tweets	Number of tweets excluding retweets	collection period	Number of training and evaluation tweets		
				Review for 1 adjective only.	not Review	total
Satomi Ishihara	1,677,692	555,701	2020/8/2 ∼ 2020/10/24	112	188	300
Haruma Miura	557,796	130,734	2020/7/22 ∼ 2020/8/15	147	153	300
Masato Sakai	161,404	61,108	2020/8/2 ∼ 2020/10/24	131	169	300
Yuko Takeuchi	868,979	337,672	2020/9/27 ∼ 2020/10/24	126	174	300
Ryoko Yonekura	37,445	14,869	2020/8/2 ∼ 2020/10/24	120	180	300
total	-	-	-	636	864	1,500

Table 2: Number of Adjectives Co-occurring in a Tweet (%)

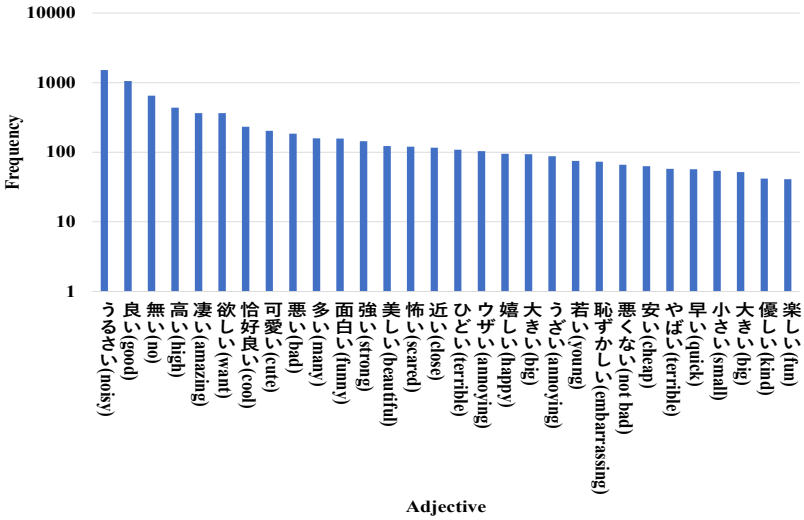
Review or Not Review	Number of co-occurring adjectives in a tweet		total
	1	≥ 2	
Review for 1 adjective only.	335 (22.3)	300 (20.0)	635 (42.3)
Not Review	487 (32.5)	378 (25.2)	865 (57.7)
total	822 (54.8)	678 (45.2)	1,500 (100)



(a) Satomi Ishihara



(b) Masato Sakai



(c) Ryoko Yonekura

Figure 3: Statistics of the Frequencies of Adjectives Co-occurring with Celebrity Names (30 Types of Adjectives with the Highest Frequencies)

tweets are additionally collected from the tweets in which the pair of celebrity name c and adjective a co-occur.

3 Review Mining Models

3.1 Machine Reading Comprehension Model

For modeling with the machine reading comprehension model, we use the question-answering framework of Figure 1. In this framework, tweets which contain a celebrity name are used as the context C, the celebrity name is used as the keyword Q, and the adjective A is the output if the adjective indicates a review relation to the celebrity name Q in the tweet. If all the adjectives in the tweet do not indicate a review relationship with the celebrity name Q, nothing is output. With this framework, it is judged whether or not there exist any review relation with the celebrity name Q in the tweet.

Specifically, in the case of the tweet in the "Review" class, we store the example as a tuple of the context C, the question as the keyword Q for the celebrity name, and the token position L_A of the answer adjective A indicating the review relationship with Q.

Context C: the tweet in which the keyword Q is co-occurring with the adjective A, which has the review relationship with Q.

Question Q: the keyword Q as the celebrity name

Token position L_A of answer A : the token position of the adjective A, which has the review relation with Q.

In the case of the tweet in the "Not Review" class, on the other hand, we store the example as a tuple of the context C for the machine reading comprehension model (Devlin et al., 2019) training, the question Q as the keyword for the celebrity name, and answer A' = N/A.

Context C: the tweet where the keyword Q and all adjectives in the tweet do not have review relationship.

Question Q: the keyword Q as the celebrity name

Answer A': N/A

Haruma Miura, 169 for Masato Sakai, 174 for Yuko Takeuchi, 180 for Ryoko Yonekura. The number of tweets excluded as others totaled 169 of which 20 for Satomi Ishihara, 64 for Haruma Miura, for 21 Masato Sakai, for 44 Yuko Takeuchi, for 20 Ryoko Yonekura.

3.2 Token Classification Model

Weinzierl and Harabagiu (2020) divided tweets into tokens and predicted what events each token was related to using a multi-class classification model. In this paper, we also divide tweets into tokens and build a token classification model that predicts whether each token indicates the name of a celebrity, a review of that celebrity, or something else.

For modeling with the token classification model, we use the named entity recognition framework of BERT (Devlin et al., 2019). The token classification model in this paper is shown in Figure 2. In the training data of the token classification model, the sequence of all morphemes in a tweet which contains a celebrity name is denoted as L_1,...,L_n, and each morpheme L_i ($i = 1, \ldots, n$) is regarded as a token, and is annotated with one of the 3 classes "REVIEW", "TARGET" and "NONE". The "REVIEW" class is a class that indicates an opinion on the celebrity, the "TARGET" class is a class that indicates the name of the celebrity, to whom the opinion of the "REVIEW" class is directed, and the "NONE" class is a class that indicates other morphemes.

As shown in Figure 2, when testing, given the tweet which is denoted as the context C (containing the query keyword Q of the celebrity name) and the query keyword Q of the celebrity name, the token classification model predicts the class of each morpheme L_i ($i = 1, \ldots, n$) into 3 classes "REVIEW", "TARGET" and "NONE". Then, if the morpheme L_j ($1 \leq j \leq n$) of the query keyword Q of the celebrity name is predicted as "TARGET", and there exists at least one morpheme L_k ($1 \leq k \leq n$) that is predicted as "REVIEW", then, the overall output as "with REVIEW" is predicted.

4 Evaluation

4.1 The Procedure

In this paper, we used PyTorch implementation of BERT (Devlin et al., 2019) for both the machine reading comprehension model and the token classification model. For BERT, we used the NICT BERT Japanese Pre-trained model[8], which was pre-trained using the entire Japanese Wikipedia except for the

[8] https://alaginrc.nict.go.jp/nict-bert/index.html

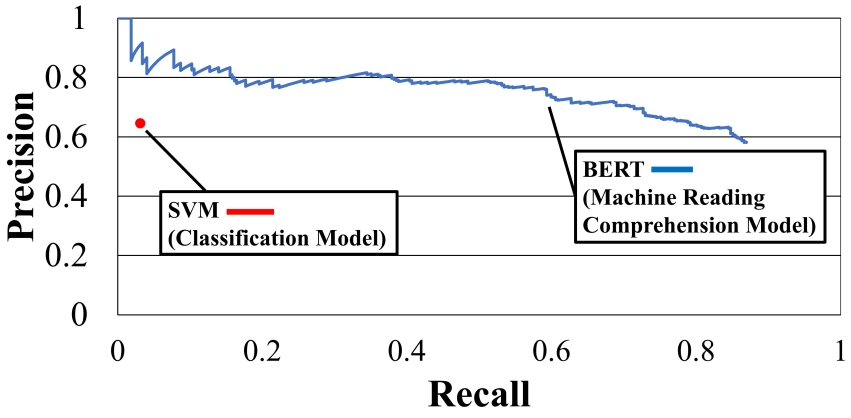


Figure 4: Evaluation Results where the Model Output is "Review" (5-fold Celebrities Cross-Validation)

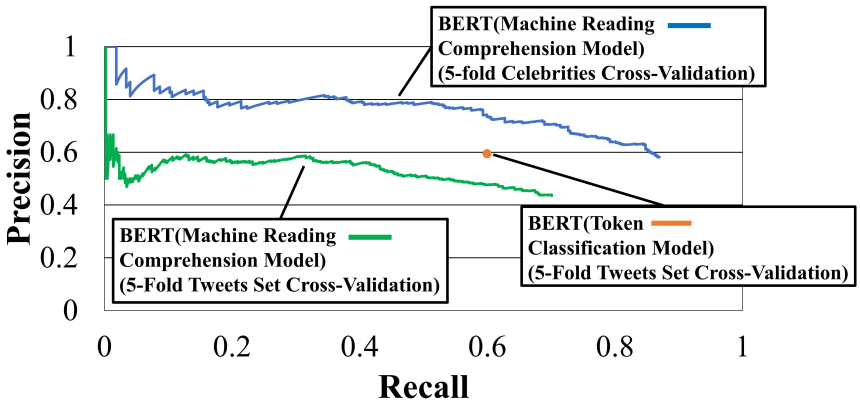


Figure 5: Comparative Evaluation Results of 5-fold Celebrities Cross-Validation and 5-fold Tweets Set Cross-Validation (where the Model Output is "Review")

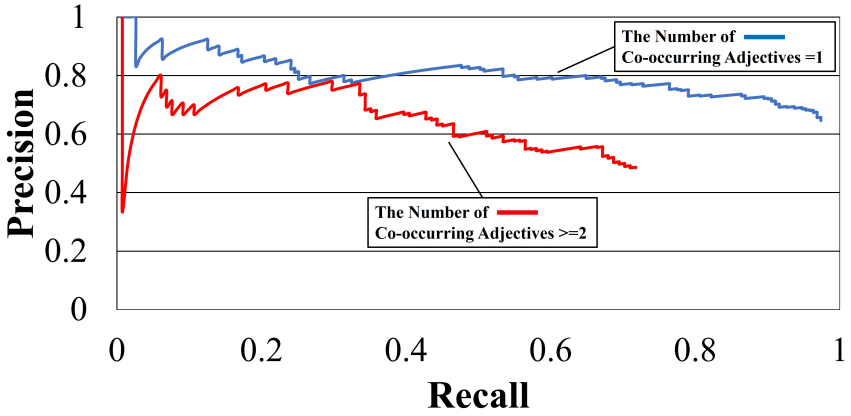


Figure 6: Evaluation Results per the Numbers of Co-occurring Adjectives in a Tweet (where the Model Output is "Review" and the Model is BERT (Machine Reading Comprehension Model))

title. The context C is segmented into morpheme sequences by JUMAN++. Then, based on the BERT specification, the WordPiece module[9] was applied to segment the context C further into subword units with a vocabulary of 32,000. For fine-tuning of the machine reading comprehension and the token classification models, we used the modules[10] of Huggingface[11]. For the evaluation, we conducted two types of cross-validation, namely, 5-fold celebrities cross-validation and 5-fold tweets set cross-validation. In 5-fold celebrities cross-validation, we trained on the data of 4 of the 5 celebrities and evaluated on the data of the remaining one celebrity. The 5-fold tweets set cross-validation is the usual 5-fold cross-validation for the whole set of collected tweets. The whole set of collected tweets was randomly divided into 5 subsets regardless of the name of the celebrity, and the training was conducted on 4 of the 5 subsets and the evaluation was conducted on the remaining one subset.

As a baseline, we modeled the text classification model[12] with SVM. The implementation was done using `scikit-learn`. As the feature scaling of the dataset, each tweet was vectorized using tf-idf to train a linear SVM classification model[13], and 5 celebrities cross-validation was performed.

4.2 Evaluation Result

With the "Review" class as the positive examples, Figure 4 plots the recall-precision curve of the machine reading comprehension model as "BERT (Machine Reading Comprehension Model)", where the lower bound p_0 of the output probability of BERT softmax function is changed in descending order[14]. Since the output probability of BERT softmax function is reliable as the confidence of BERT softmax function, the precision decreases as the recall increases. Figure 4 also plots precision and recall for the baseline as "SVM (Classification Model)"[15]. The figure shows the results of the 5-fold celebrities cross-validation. Here, in the 5-fold celebrities cross-validation, the token classification model cannot predict the query celebrity name that is not observed in training as "TARGET". Thus, we omit the plot for the token classification model. It is obviously shown that the machine reading comprehension model outperforms the baseline.

For the machine reading comprehension model, Figure 5 compares the 5-fold celebrities cross-validation with 5-fold tweets set cross-validation. Figure 5 also plots precision and recall for the token classification model as "BERT (Token Classification Model)". For the the machine reading comprehension model, the 5-fold celebrities cross-validation outperforms the 5-fold tweets set cross-validation, suggesting that the machine reading comprehension model is a model that does not over-fit the features of each celebrity. From this result, it can be concluded that the machine reading comprehension model trained with tweets including previously observed celebrities' names is also effective in detecting the review relationship with previously unobserved celebrity names. For the case of 5-fold tweets set cross-validation, the precision of the token classification model is relatively low compared with the machine reading comprehension model in the high confidence range of prediction.

Table 2 shows the statistics of the number of adjectives co-occurring in a tweet. This table shows that there exist certain percentages of the cases where the number of co-occurring adjectives is two or more, while one of those co-occurring adjectives has the review relationship with a celebrity name. Figure 6 further compares the recall-precision curves of the following two cases: (i) where the number of co-occurring adjectives is one, and (ii) where the number of co-occurring adjectives is two or more. This result shows that, in the case where the number of co-occurring adjectives is one, the performance of review relation detection is much higher than the case where the number of co-occurring ad-

[9] `tokenization.py`

[10] `run_squad.py` and `run_ner.py` were used respectively where the number of epochs as 2, the batch size as 8, and the learning rate as 0.00003.

[11] transformers-2.2.1

[12] The model takes "the celebrity name token sequence + [SEP] token + the tweet token sequence" as the input and classifies the input into 2 classes "Review" or "Not Review".

[13] C=1.0

[14] In this case, the recall does not reach 1 because there exist cases where the answer span predicted by the model does not match the reference answer. Thus, we only show the recall-precision curves, whereas we do not show ROC-curves in this paper.

[15] Although it is also possible to plot the recall-precision curve for SVM classification model, we omit it since the recall of SVM classification model is too low.

jectives is two or more. However, even in the case where the number of co-occurring adjectives is two or more, its precision is over 0.5, which is more or less satisfactory performance and this result confirms the effectiveness of the proposed approach.

5 Related Work

Compared with our analysis of Twitter mentions of celebrities, as a related work, Wiegmann et al. (2019) studied author profiling of celebrities in Twitter. Wiegmann et al. (2019) collected Wikipedia entries and Twitter feeds for 71,706 celebrities and developed a corpus containing an average of 29,968 words and a maximum of 239 personal traits per celebrity. They developed a model to predict gender and occupation from tweets using deep learning methods.

Span prediction has been also studied in previous tasks. For example, Alhuzali and Ananiadou (2021) proposed a model that casts the emotion classification task as span prediction. In the context of named entity recognition, Fu et al. (2021) studied the strengths and weaknesses of the span prediction model and compared it with the sequence labeling framework. In this paper, we employed the machine reading comprehension model based on span prediction to extract adjectives that indicate a review relationship with a celebrity name.

6 Conclusion

In this paper, we focus on the celebrities' names and adjective expressions that appear in tweets, and determine whether or not there is a relationship such as review and impression between the two words using the BERT machine reading comprehension framework. We also compared 5 celebrities cross-validation with 5-fold cross-validation, which showed that the machine reading comprehension model is robust enough not to overfit the features of each celebrity.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 21H00901.

References

Hassan Alhuzali and Sophia Ananiadou. 2021. SpanEmo: Casting multi-label emotion classification as span-prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1573–1584, Online, April. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Jinlan Fu, Xuanjing Huang, and Pengfei Liu. 2021. SpanNER: Named entity re-/recognition as span prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7183–7195, Online, August. Association for Computational Linguistics.

Yuting Guo, Yao Ge, Mohammed Ali Al-Garadi, and Abeed Sarker. 2021. Pre-trained transformer-based classification and span detection models for social media health applications. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 52–57, Mexico City, Mexico, June. Association for Computational Linguistics.

Rongtao Huang, Bowei Zou, Yu Hong, Wei Zhang, AiTi Aw, and Guodong Zhou. 2020. NUT-RC: Noisy user-generated text-oriented reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2687–2698, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, California, June. Association for Computational Linguistics.

Manuela Sanguinetti, Cristina Bosco, Lauren Cassidy, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2020. Treebanking user-generated content: A proposal for a unified representation in Universal Dependencies. In *Proceedings of the 12th Language Resources and Evaluation Conference*,

pages 5240–5250, Marseille, France, May. European Language Resources Association.

Maxwell Weinzierl and Sanda Harabagiu. 2020. HLTRI at W-NUT 2020 shared task-3: COVID-19 event extraction from Twitter using multi-task hopfield pooling. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 530–538, Online, November. Association for Computational Linguistics.

Matti Wiegmann, Benno Stein, and Martin Potthast. 2019. Celebrity profiling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2611–2618, Florence, Italy, July. Association for Computational Linguistics.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2324–2335. Association for Computational Linguistics.