Latent Chained Comments to Retweet Extraction on Twitter

Ryusei Takagi[1] and Yasunobu Sumikawa[1]

Depat. of Computer Science, Takushoku University, 815-1 Tatemachi, Hachioji-shi,
Tokyo, Japan
r88446@st.takushoku-u.ac.jp, ysumikaw@cs.takushoku-u.ac.jp

Abstract. Twitter, one of the social networking services, has a retweet function that displays tweets posted by other users on the retweeter's timeline. As this retweet function spreads information, past studies have been conducted to analyze posts that are spread by many people. However, some Twitter users use the retweet function not only to spread information, but also to discuss their opinions. In these discussions, they tend to do retweet first, followed by multiple tweets of their own opinions. We call the multiple tweets representing the opinions latent comments. In this study, we propose algorithms that collect tweets which talk about their previous retweets and are continuously posted. We have created a novel dataset including latent comments and evaluated our algorithm on the dataset. We have confirmed that the algorithm performs well on the dataset.

Keywords: Topic detection and tracking · Natural language processing · Machine learning · Twitter · commentary tweet

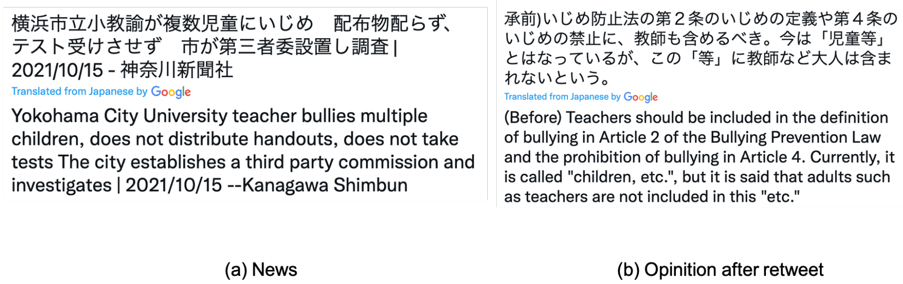
1 Introduction

Twitter provides the retweet function that allows Twitter users to share information posted by other users by showing the content on the users' timeline. As this function is effective in delivering information to many users, it is also used for corporate advertising. Indeed, the number of times a tweet has been retweeted is one of the important indicators of popularity [4, 8].

The number of retweets on Twitter is an important measure of popularity; however, as text is the basic way of communicating on Twitter, Twitter users sometimes do retweets to start a discussion on the retweeted contents. in this study, we define latent comments if the following two conditions are satisfied: 1) tweets classified as one of the three types: rule-, direct-, and indirect-based referring tweets related to the retweeted content, and 2) continuously chained on the user's timeline. Fig. 1 shows an example of latent comments. The tweet[1] shown in Fig. 1(a) is an original tweet whereas the tweet[2] in Fig. 1(b) asserts the user's opinion to the original tweet. Looking at Fig. 1(b), the user does the

[1] https://mobile.twitter.com/takeda828/status/1458052542269558786

[2] https://mobile.twitter.com/takeda828/status/1458053054398222341



横浜市立小教諭が複数児童にいじめ　配布物配らず、テスト受けさせず　市が第三者委設置し調査 | 2021/10/15 - 神奈川新聞社

Translated from Japanese by Google

Yokohama City University teacher bullies multiple children, does not distribute handouts, does not take tests The city establishes a third party commission and investigates | 2021/10/15 --Kanagawa Shimbun

承前)いじめ防止法の第２条のいじめの定義や第４条のいじめの禁止に、教師も含めるべき。今は「児童等」とはなっているが、この「等」に教師など大人は含まれないという。

Translated from Japanese by Google

(Before) Teachers should be included in the definition of bullying in Article 2 of the Bullying Prevention Law and the prohibition of bullying in Article 4. Currently, it is called "children, etc.", but it is said that adults such as teachers are not included in this "etc."

(a) News

(b) Opinition after retweet

Fig. 1. Example of a latent comment.

assertion by posting an original tweet instead of posting the tweet as a quote tweet. Several Twitter users post tweets representing their opinions as original tweets as well as the tweet referred to in the example. Therefore, it is necessary to design algorithms for detecting such opinion tweets.

In this study, we propose three algorithms for automatically collecting latent comments. The proposed algorithms are designed for the three types of latent comment: 1) rule-based latent comments contain commonly used notations such as “>RT” regardless of the topic, 2) direct-based latent comments contain the same words as the retweeted text, and 3) indirect-based latent comments do not contain the same words as the retweeted text.

To check the effectiveness of our algorithm, we collected 3,293 tweets posted by news agencies’ official accounts and manually collected 100 test data including latent comments of Twitter users who retweeted the tweets. Using these tweet data, we found that the proposed algorithm achieved a high accuracy of 85.9% as the F_1 score.

Definitions: We define a tweet set as a *latent comment* if the content of the tweets is related to the previous retweet and the tweets are continuously posted without including non-latent comments between them. Fig. 2 shows the concept of the latent comments. The solid and dotted lines indicate the connected tweets as latent comments and non-latent comments, respectively. There are two retweeted news tweets about an Olympic game and a trading event. Tweet1 follows the first retweet about the Olympic game and contains “Opening ceremony”; thus, we regard this tweet as a latent comment. As Tweet2 includes the keyword “Olympic” and follows Tweet1 and Retweet1, this tweet is also a latent comment for Retweet1. The text of the last tweet, Tweet3, is “I will get up early tomorrow”; it is unclear if the tweet points to the Olympic game. In addition, there is a retweet between Retweet1 and Tweet3. We do not consider Tweet3 as a latent comment for Retweet1.

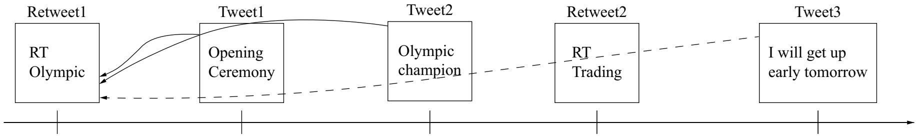


Fig. 2. Examples of chained latent comments

2 Related Work

2.1 Topic Detection and Tracking

Identifying latent comments is to extract tweet chains; thus, this study is one of the topic detection and tracking (TDT) studies. Looking at previous TDT studies, Radinsky and Davidovich proposed an algorithm that extracted news chains representing causal relationships from the descriptions of past events to predict possible future events [10]. Comparing with our study, the past study assumptions are based on long descriptions whereas our study assumptions are based on short descriptions. For TDT studies on Twitter, there are methods to detect and track hot events from online news streams [9], detect global and local hot events using local community detection mechanisms [11], and reporting real-life events to users in the human-readable form [7]. These past events use not only tweet texts but also context information such as trending and geography. By contrast, using this context information is difficult in our study; therefore, we use tweet text only.

2.2 Using Explicit Popularity on SNS

Twitter is sometimes used as a medium that reflects current trends or influences. Studies have been conducted over the past decade to predict which tweets will receive the most likes and retweets. [5] proposed a model that predicts the number of retweets within a given timeframe. Some recent studies proposed methods to discover influencers on Twitter in addition to the popularity of each tweet, and the number of retweets is used as one of the indicators of influencers [13, 12, 2]. In the aforementioned studies, retweets themselves are considered as an influence. While previous studies consider retweets as an indicator of influence, this study uses them as starting points for discussions and detects their tweet chains.

3 Data Collection

3.1 Tweet collection

We collected all tweets by the Twitter official API. Twitter provides three kinds of tweets: tweets, retweets and quote tweets. If a Twitter user posts an original

Table 1. Statistics of collected tweet dataset

Number of news tweet	3,293
Number of news articles	3,293
Number of retweet IDs	2,267,322
Period of timestamps	24 Oct. 2021 – 2 Mar. 2022
Number of news categories	51

text, it is defined as a tweet. If a tweet is re-posted by a user using Twitter's official retweet function, it is called a retweet. If a tweet is re-posted with new content, it is called a quote tweet; thus, a quote tweet can be called as a re-tweet with a comment. In this study, we treat all quote tweets as original tweets as they include additional information/text.

We collected tweets from official accounts of news agencies to create our dataset. The reason of this selection is twofold. The first one is that the accounts have numerous numbers of tweets. The second one is that the number of people viewing those tweets is also large. These reasons are necessary to perform a detailed analysis including what kind of tweets are latent referring tweets as discussed in Section 5.

We collected news tweets only from Japanese news agencies' accounts and performed manual dataset creation so that we can properly analyze whether Twitter users' tweets are references to news. The news agencies we collected were the Yomiuri Shimbun, Asahi Shimbun, NHK, Mainichi Shimbun, and Yahoo News[3]. For these accounts, we used user_timeline[4] an official Twitter API to retrieve tweets from the news agencies' accounts.

3.2 Latent comment candidate collection

In this study, we used two Twitter official APIs, retweeters[5] and user_timeline API, to collect the Twitter users who retweeted each news tweet from the news agency tweets and their latent comments.

The detailed collection procedure is as follows. First, to collect the tweets that retweeted the news agencies' tweets, we used retweeters to obtain the user IDs of the users who retweeted them. For these user IDs, we retrieved past tweets using the user_timeline API and collected 10 tweets each before and after the retweet of the news tweet. Tab. 1 shows the statistics of the collected tweets.

[3] The names of the news agency accounts we collected are Yomiuri_Online, asahi, nhk_news, mainichi, YahooNewsTopics.

[4] `https://developer.twitter.com/en/docs/twitter-api/v1/tweets/timelines/api-reference/get-statuses-user_timeline`

[5] `https://developer.twitter.com/en/docs/twitter-api/v1/tweets/post-and-engage/api-reference/get-statuses-retweeters-ids`

4 Algorithm

Our algorithm determines whether the given tweet is a latent comment to the given retweet or not. If the algorithm decides that the tweet is a latent comment, the algorithm returns true; otherwise, it returns false. If there are other tweets between the two given IDs, the algorithm traces them one by one to determine if each tweet is a comment on the given retweet. For this determination, we propose three types of latent comments: rule-, direct-, and indirect-based comments. The details of identifying chained comment tweets and the 3 types of tweets and the algorithms for detecting them are in the remainder of this section.

4.1 Chained comment tweet identification

This algorithm determines whether a given tweet is a comment on another given retweet. As a single tweet can only contain 280 characters, Twitter users sometimes post their opinions in multiple tweets. To track such tweets, we retrieve other tweets between a given tweet ID and retweet ID. As the purpose of this retrieving is to reveal posts that divide one's opinion into multiple posts, it is assumed that there are no other retweets between them. We apply the proposed algorithm recursively to the obtained tweets to analyze whether they are latent comments or not. If we find other retweets posted at this time, then the recursive analysis is terminated and false is returned. We also terminate the recursive analysis if false is obtained by applying the proposed algorithm described in the following sections. If the recursive search visits the retweet ID and determines that it is a latent comment, the algorithm returns true. Once this recursive search returns false, our algorithm considers the given tweet as a non-latent comment.

4.2 Rules in latent comments

On Twitter, there are widely used keywords such as "> RT" when expressing own opinion after a retweet. We consider rule-based latent comments as tweets including these keywords in their texts. The algorithm to find this type of tweet is based on the presence of such keywords in the text.

4.3 Direct-based latent comments

We call tweets direct-based latent comments when the tweets explicitly use words contained in the retweeted content. To collect these tweets, we use the Jaccard coefficient that calculates the similarity between two texts based on the number of words they share. If the score is over a given threshold, we consider the tweets as latent comments. In this study, we set 0.2 as the threshold. The formal definition is as follows:

$$Jaccard(A, B) = \frac{\mid T_A \cap T_B \mid}{\mid T_A \cup T_B \mid} \tag{1}$$

where: $\mid \cdot \mid$ is the size of the set and T_A and T_B are the tokens included in tweet A and B, respectively. The higher the score of the measurements, the more correlated they are.

4.4 Indirect-based latent comments

We call tweets indirect-based latent comments if the tweets use indirect-referring words to refer to the retweets instead of using words contained in the retweeted content. To collect these tweets, we cannot use direct keywords as a filter; thus we perform latent semantic analysis such as LDA [1], LSA [3], and Doc2Vec[6]. We first create feature vectors with the latent semantic analysis models equipped with TF-IDF model. TF-IDF indicates the importance of a word to a document in the dataset. This score is a multiplication of the term frequency and inverse document frequency. Term frequency refers to how frequently each term (word) occurs in each document, whereas the inverse document frequency represents how rarely each term occurs in all documents. The formal definition is as follows.

$$TFIDF(w, d, \mathcal{D}) = tf_{w,d} * \frac{\mid \mathcal{D} \mid}{\mid \{d' \in \mathcal{D} \mid w \in d'\} \mid} \tag{2}$$

where $tf_{w,d}$ is the number of times a word w occurs in a document d and $\mid \bullet \mid$ is the size of $\bullet$. The second term of this equation gives the number of all labeled data divided with labeled data including w.

We then compute the cosine similarity between the feature vectors of retweets and news agencies' tweets. We regard the tweet as latent comment if the score of cosine similarity is over 0.5 in this study.

4.5 Overview of identifying latent referring tweet algorithm

At the beginning, we input two tweet IDs: an original tweet ID and a retweet ID. We then collect other tweets between the given two tweet IDs described in Sec. 4.1. Thereafter, we apply morphological analysis, lemmatization, stop word removal and other fundamental natural language processing techniques to create tokens for each tweet. We then sequentially apply the procedures described in Sec. 4.2, 4.3, and 4.4 to analyze if each tweet is a comment on a given retweet. After applying the three processes, we return the result of applying logical OR to all the results obtained.

5 Experimental Evaluations

5.1 Experimental setting

Dataset. We have created a dataset including latent comments for this study because there was no ground truth dataset for latent comment extraction task. We first collected 3,293 news tweets as described in Section 3.1 from Oct. 24, 2021 to Mar. 2, 2022. We then randomly selected 1,000 news tweets and collected tweets posted after retweeting the news tweets. We manually checked whether the tweets are latent comments for retweets. If the tweets were determined as latent comments for the retweets, we combined them as pairs in our dataset. This was performed by two workers, including Ph.D. holder working on machine

Table 2. Rules used in latent comments. We add English words to Japanese to make it easier to understand though we used only Japanese in the experimental evaluation.

> RT	→ RT	(RT)	RT >	RT	RT:
↓	承前 (continue)	（RT 参照） (RT reference)			

Table 3. Micro-average precision, recall, and F-scores of latent comment or not classification by each algorithm.

	P	R	F_1
Rule	52.0%	52.0%	52.0%
Direct reference	86.0%	86.0%	85.9%
Indirect reference	81.0%	81.0%	81.0%
All	*86.0%*	*86.0%*	*85.9%*

learning research. Then, the collected pairs were verified on agreement of the two workers, and the verified pairs were kept for the experiment. As results, the dataset contained 50 latent comments. We also added 50 non-latent comments into the dataset. After removing the test data, we trained LDA on the collected tweet data to use the model as an indirect-based latent comment identifier.

Except for the correct answer data above, the authors extracted rules from 5,000 randomly selected tweets that could be used in a rule-based algorithm. Tab. 2 lists all the keywords obtained from this manual process. These keywords are used in this study.

Evaluation Criteria. We evaluated the proposed algorithm as a binary classification problem to determine whether the result is a latent comment or not. Therefore, we evaluated the algorithm with micro-average precision (P), recall (R), and F1-score (F_1), which are widely used to evaluate classification. To better understand the accuracy of the proposed algorithms, we evaluated the prediction results of each algorithm as a multi-class classification problem. As the number of test data is different for each label, we used the macro-averages precision (maP), recall (maR), and F1-score (maF).

5.2 Results

Q. How accurate is the proposed algorithm in detecting whether a comment is latent or not?
A. The proposed algorithm achieved 86% as P, R, and F_1 scores.

Tab. 3 shows P, R, and F_1 scores of four cases where each of the three algorithms proposed in this study is applied and where all results, represented as *All*, are combined by OR operator. We can see that using all the three algorithms and Direct reference are the best among the four cases. As all the P, R, and F_1

Table 4. Macro-average precision, recall, and F-scores of multiclasss classification by each algorithm.

	maP	*maR*	*maF*
Rule	37.7%	49.5%	41.7%
Direct reference	45.0%	46.3%	45.4%
Indirect reference	43.1%	27.7%	26.0%
All	*95.2%*	*73.6%*	*75.3%*

scores achieved approximately 86%, we can conclude that our algorithm works well.

Q. How accurate is each algorithm for identifying each latent comment type?
A. The algorithm of finding each latent comment type achieved a score of approximately 26% to 45% as *maF* scores.
A. The algorithm combining the three algorithms achieved 75% as a *maF* score.

Tab. 4 shows *maP*, *maR*, and *maF* scores for identifying each type of the latent comment by each of the three algorithms and *All* proposed. This result also indicates that using all the three algorithms is the best among the four cases. Especially, our algorithm achieved 75% as *maF* score.

To analyze the features of the proposed algorithm in detail, Fig. 3 shows the results of our algorithm in correctly predicting whether a tweet is a latent comment or not. The results show that our algorithm correctly predicted all non-latent comments as non-latent; however, it mispredicted some latent comments as non-latent. To analyze what kind of data was mispredicted, the confusion matrix summarizing the prediction results of each algorithm is shown in Fig. 4. We can see that all rule-based latent comments were correctly predicted. By contrast, many indirect-based latent comments were mispredicted as not latent comments. Our dataset includes only 10% indirect-based latent commenting tweets; thus, this misprediction had a small impact on the overall evaluation results. However, for improving accuracy, more sophisticated algorithms can be used to properly discover indirect-based latent comments rather than the LSA used in this study.

5.3 Limitations

Although we focused on Japanese news tweets in this study, several Twitter users post latent comments without limiting to news tweets. To analyze the exact overall trend of Twitter users, we need to collect not only news tweets but also all tweets and their retweets. If we can solve the problem of limited use of Twitter API and collecting all tweets or Twitter providing the data, the above bias can be removed with the proposed algorithm. However, currently we cannot solve the issue of data collection; thus we intend to study this issue as a future work.

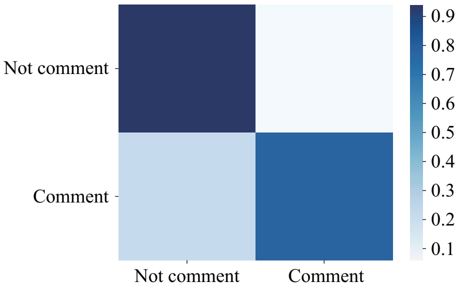


Fig. 3. Confusion matrix for binary classification on identifying latent comment or not.

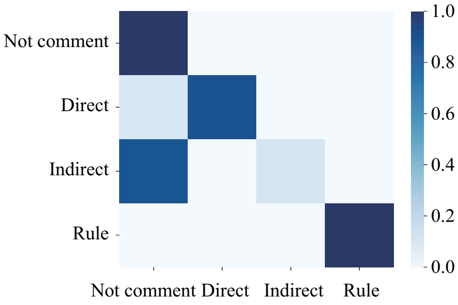


Fig. 4. Confusion matrix for prediction results of each algorithm.

6 Conclusions

In this study, we have proposed algorithms to detect latent comments for the retweets. These latent comments are tweets in which Twitter users discuss their opinions after retweeting others' tweets. The algorithms are designed for detecting three types of latent comments: rule-, direct-, and indirect-based latent comments. To evaluate the effectiveness of our algorithm, we created a test dataset and tested as classification problems. The results showed that we confirmed that our algorithm obtained about 86% micro-averaged and 75% macro-averaged F1 scores.

There are two possible directions for future research. The first one is to estimate authentic popularity measures on Twitter. There have been studies using Twitter popularity; however, most of them simply use the number of retweets. However, as the definition proposed in this study, Twitter users sometimes retweet to start discussions or express their opinions. We will combine the algorithm proposed in this study with sentiment analysis to identify whether the latent comments for each retweet are positive or negative. The second one is that we will apply our algorithm to other language tweets as this paper focused only on Japanese tweets.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (Mar 2003)
2. De Salve, A., Mori, P., Guidi, B., Ricci, L., Pietro, R.D.: Predicting influential users in online social network groups. ACM Trans. Knowl. Discov. Data **15**(3) (Apr 2021)
3. Deerwester, S., T. Dumais, S., W. Furnas, G., Thomas K., L., Harshman, R.: Indexing by latent semantic analysis. J. Amer. Soc. Inform. Sci. **41**(6), 391–407 (1990)

4. Hong, L., Dan, O., Davison, B.D.: Predicting popular messages in twitter. In: Proceedings of the 20th International Conference Companion on World Wide Web. pp. 57–58. WWW '11, Association for Computing Machinery, New York, NY, USA (2011)
5. Kupavskii, A., Ostroumova, L., Umnov, A., Usachev, S., Serdyukov, P., Gusev, G., Kustarev, A.: Prediction of retweet cascade size over time. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. pp. 2335–2338. CIKM '12, Association for Computing Machinery, New York, NY, USA (2012)
6. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. vol. 32, pp. 1188–1196. ICML'14, Bejing, China (22–24 Jun 2014)
7. Lei, Z., Wu, L.d., Zhang, Y., Liu, Y.c.: A system for detecting and tracking internet news event. pp. 754–764. PCM' 05, Springer-Verlag, Berlin, Heidelberg (2005)
8. Naveed, N., Gottron, T., Kunegis, J., Alhadi, A.C.: Bad news travel fast: A content-based analysis of interestingness on twitter. In: Proceedings of the 3rd International Web Science Conference. WebSci '11, Association for Computing Machinery, New York, NY, USA (2011)
9. Qi, Y., Zhou, L., Si, H., Wan, J., Jin, T.: An approach to news event detection and tracking based on stream of online news. vol. 2, pp. 193–196 (2017)
10. Radinsky, K., Davidovich, S.: Learning to predict from textual data. J. Artif. Int. Res. **45**(1), 641–684 (Sep 2012)
11. Tan, Z., Zhang, P., Tan, J., Guo, L.: A multi-layer event detection algorithm for detecting global and local hot events in social networks. Procedia Computer Science **29**, 2080–2089 (2014), 2014 International Conference on Computational Science
12. Zhang, Z., Zhao, W., Yang, J., Paris, C., Nepal, S.: Learning influence probabilities and modelling influence diffusion in twitter. In: Companion Proceedings of The 2019 World Wide Web Conference. pp. 1087–1094. WWW '19, Association for Computing Machinery, New York, NY, USA (2019)
13. Zheng, C., Zhang, Q., Young, S., Wang, W.: On-demand influencer discovery on social media. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. pp. 2337–2340. CIKM '20, Association for Computing Machinery, New York, NY, USA (2020)