DEIM2022 B34-4

Early Detection of Fact Check-worthy Tweets by Using User Reactions

Yimou LIAO†, Masashi TOYODA††, and Naoki YOSHINAGA††

† The University of Tokyo, 7–3–1 Hongo, Bunkyo-Ku, Tokyo 113–8656, Japan
†† Institute of Industrial Science, the University of Tokyo, 4–6–1 Komaba, Meguro-Ku, Tokyo 153–8505, Japan
E-mail: †, ††{liao-y, toyoda, ynaga}@tkl.iis.u-tokyo.ac.jp

Abstract To reduce the spread of misinformation on Twitter, we should validate questionable facts in tweets as early as possible. Since manual fact-checking is costly and time-consuming, fact-checkers need to focus their work only on check-worthy tweets. The detection of fact check-worthy tweets has been studied as a shared-task proposed by CheckThat! Lab. In this task, the most retweeted tweets are selected, and they are ranked by their check-worthiness based only on their texts. However, user reactions, such as retweets and replies, are not used in the task, though they are useful and available at very early stage. We thus propose a method of detecting check-worthy tweets in a timely manner using early user reactions. We expand dataset provided by CheckThat! Lab with user reactions (retweets, quotes, and replies), and create linguistic, structural, temporal features for tweets based on these user reactions. We then use a combined neural network model that takes the outputs of the tweet-text-based baseline model, RoBERTa, in addition to user-reaction-based features for downstream classification and ranking layers. Our experimental results show that F1-score and MAP (Mean Average Precision) can be improved when using early user reactions. We also investigate the trade-off between accuracy and earliness of our method.

Key words SNS, Twitter, fact-checking, check-worthiness, user reaction

1 Introduction

The spread of misinformation on social networking services (SNSs) becomes a more serious problem as SNSs acquire a huge number of users. In fact, the amount of misinformation has been increasing explosively and rapidly [8]. Misinformation has many negative effects. It can mislead people, bring anxiety, and destroy public trust. Even worse, it might bring harm to our whole society in economics, politics, and public health. During the past year, COVID-19 vaccine misinformation circulated on social media had negative impact on people's vaccine beliefs and behaviors [7].

To fight against misinformation, fact-checking organizations including *FactCheck*,[1] *PolitiFact*,[2] *FullFact*,[3] have been launched. They are expected to validate questionable facts as early as possible. However, fact-checking is commonly a costly and time-consuming process, with multiple fact-checkers and several procedures needed before the final judgement is made. Fact-checking of social media postings like tweets is even more challenging, because of their explosive quantity, conversational nature, and lack of context [4]. To reduce the work of fact-checkers, it is important to determine whether a social media posting is worthy of fact-checking in the first place. Examples of a check-worthy tweet and a not check-worthy tweet is shown in Table 1.

check-worthy	A person who died last week in a Seattle hospital has since tested positive for coronavirus. The person lived in the same nursing home that has had numerous COVID19 cases and deaths. How many people were exposed due to lack of testing.
not check-worthy	I live in Seattle, I have all symptoms of COVID-19 and have a history of chronic bronchitis. Since I work in a physical therapy clinic with many 65+ patients and those with chronic illnesses, I decided to be responsible and go to get tested. This is how that went.

Table 1: Examples of check-worthy and not check-worthy tweets.

The detection of fact check-worthy tweets has been proposed as a shared-task by CheckThat! Lab [3]. They define the task setting as estimating fact check-worthiness of the most retweeted tweets about a certain topic (*e.g.*, COVID-19) during a certain period based on their texts. In this setting, tweets for estimation are assumed to have a number of user reactions, yet the role of these user reactions is ignored. Besides, the detection timing is also ignored, as most of the tweets in training data appear several weeks later than tweets in test data.

In this work, we propose a method of detecting fact check-worthy tweets in a timely manner by using early user reactions. In real-world scenario, fact-checkers have to validate questionable tweets as early as possible before they are retweeted by many users. Yet, tweets that

(1) : http://www.factcheck.org
(2) : http://www.politifact.com
(3) : http://fullfact.org

have not been retweeted have little social influence and tend to be not check-worthy. We thus come up with the idea of utilizing all the data available, including early user reactions. We focus on early user reactions because we assume that they can provide some important clues for the estimation of fact check-worthiness.

To propose the method, we investigate the following three research questions:

- **RQ1**: What are the differences between user reactions to check-worthy tweets and those to not check-worthy tweets?
- **RQ2**: How can we utilize these differences to improve the performance of existing check-worthiness estimation methods?
- **RQ3**: What is the relationship between the accuracy and earliness of our method?

To investigate **RQ1**, we expand existing dataset with retweets, quotes, and replies. We create linguistic features for quotes and replies, structural features and temporal features for retweets, quotes, and replies. We then analyze how these features differ between check-worthy tweets and not check-worthy tweets. To investigate **RQ2**, we combine these features with baseline Transformer-based models. We also compare the performances of different types of features by combining only part of them. To investigate **RQ3**, we control the number of user reactions as input and capture the changes of accuracy.

Our contributions are summarized as follows:

- We extend existing dataset with user reactions and firstly explore the relationship between user reactions and fact check-worthiness of tweets.
- We demonstrate the effectiveness of user reactions when combined with baseline models for the detection of check-worthy tweets.
- We simulate the real scenario for fact-checkers by evaluating the trade-off between earliness and accuracy of the detection method.

2 Related Work

In this section, we first introduce the shared task on check-worthiness estimation of tweets and the existing proposed methods. We will also point out some unnatural aspects of their task settings and how we revise the task settings to better reflect real-world scenario. We then introduce some existing work on analyzing user reactions for similar tasks of social media study and how we reconsider them to fit our task.

2.1 Check-worthiness Estimation in Tweets

Task of detecting check-worthy tweets was first proposed by CheckThat! Lab in the form of a contest [3]. They define the task as predicting which tweet from a stream of tweets on a topic should be prioritized for fact-checking. They provide both English and Arabic datasets, and all the tweets in English dataset are on the topic of COVID-19. Among the participants, Transformer-based models like BERT [6], RoBERTa [13], and COVID-Twitter-BERT [14] are used, along with other learning models like SVM(Support Vector Machine) [5] with TF-IDF features, and bidirectional LSTM(Long-Short Term Memory) [10] on top of GloVe [17] embeddings. The same contest is also organized in 2021 [15], where datasets of five different languages are provided. English dataset is enlarged with some lately posted tweets related to COVID-19. In this time, additional Transformer-based models like ALBERT [12], DistilBERT [18], and BERTweet [16] are used by some participants. Data augmentation through machine translation of other language data is also used. For evaluation of the contests, as organizers treat the task as a ranking problem where check-worthy tweets are expected to be ranked at the top, they use mean average precision (MAP) as the official evaluation measure, complement with other ranking evaluation metrics like reciprocal rank (RR), R-precision (R-P), and P@k for k $\in$ 1, 3, 5, 10, 20, 30. The evaluation results show that RoBERTa and BERTweet perform the best among the participants in 2020 and 2021, respectively. All the details can be found in the overviews of the shared tasks [3, 15].

Until now, all the work on check-worthiness estimation in tweets is based on this shared task and in the same task setting. However, we find the task setting a little unnatural. Tweets in their datasets are assumed to have a number of user reactions, in fact they are the most retweeted tweets about certain topics during certain periods. Even though they provide JSON file of tweets where metadata such as the total number of retweets are given, more detailed information such as creation time of these retweets happen is not given. In our work, we extend dataset by the missing information of user reactions and utilize it to facilitate the check-worthiness estimation in tweets.

2.2 Analysis of User Reactions on Social Media

User reactions (retweets, quotes, replies) in Twitter have been studied in some work to understand and detect misinformation. The structural features (*e.g.*, number of cascades, cascade size, cascade depth) of retweets are investigated to understand how true news and false news are spread by users in different ways [23]. Besides retweeting, users can also express their opinions in a more direct way by writing replies. Linguistic features of user replies are captured from different perspectives. Some work focus on extracting user sentiments (*e.g.*, positive, negative) and emotions (*e.g.*, joyful, sad) from their replying texts to infer whether they are replying to misinformation or not [26]. Some other recent work attempts to mine the stances (*e.g.*, agree, disagree) expressed by users towards a certain topic of misinformation on COVID-19 [25]. Temporal features (*e.g.*, time differences between two reactions) can be created in retweet networks or replying networks, and be used to analyze how users react to true news and fake news in different frequency [20].

Likewise, we characterize user reaction by using linguistic features, structural features, and linguistic features. For linguistic features, besides sentiment and emotion, we also consider offensive, irony, and hate language, which are less taken into consideration in the context of misinformation detection or fact-checking. For structural features and linguistic features, previous related work tends to focus only on one certain type of user reactions [11, 22]. Although some work applies multiple types of user reactions, they create structural features and structural features for replies and retweets respectively [19, 20]. Since our goal is to detect check-worthy tweets as early as possible, we aim to use all the information in early stage. Therefore, we create structural and temporal features based on user reaction network containing all three types of user reactions.

3 Datasets on Check-worthiness of Tweets

In this section, we first describe an existing dataset on check-worthiness of English tweets, which is constructed by CheckThat! Lab. We then introduce how we expand it by user reactions including retweets, quotes, replies, and how we re-split the dataset in time order to simulate the real-world scenario of fact-checking.

3.1 Existing Dataset on Check-worthiness of English Tweets

In CheckThat! 2020 [3], organizers of the contest constructed a dataset for the task of check-worthiness estimation of English tweets. They collected tweets that match keywords and hashtags related to COVID-19 during March 2020, and selected the most retweeted tweets for manual annotation. For annotations, they considered several factors including tweet popularity in terms of retweets. They further ask annotators to answer the following five questions:

- Q1: Does the tweet contain a verifiable factual claim?
- Q2: To what extent does the tweet appear to contain false information?
- Q3: Will the tweet have an effect on or be of interest to the general public?
- Q4: To what extent is the tweet harmful to the society, person(s), company(s) or product(s)?
- Q5: Do you think that a professional fact-checker should verify the claim in the tweet?

A tweet is annotated as check-worthy if both Q1 and Q5 are answered with yes. Although the answers to Q2, Q3, and Q4 are not considered directly, they help annotators make a better decision for Q5. The annotations are performed by 2–5 annotators independently, and then consolidated after a discussion for the cases of disagreement. The details about the annotation instructions and setup can be found in [1]. The statistics about the data is shown in Table 2.

3.2 Augmenting the Dataset by Adding User Reactions to Target Tweets

Since the original dataset only contains tweet IDs, texts, URLs of target tweets (of which the check-worthiness is to be estimated), we augment it with user reactions on our own.[4] For each target tweet, we collect its retweets, quotes, and replies that appear within 24 hours. Since we aim to detect check-worthy tweets as early as possible, we assume that user reactions that appear later than 24 hours are too late for this task. For each user reaction, we record the following information:

- tweet id, reaction type (retweet, quote, reply)
- creation time, text (only for reply)
- tweet id of its source tweet (subject of this reaction)

Partition	Total	Check-worthy
Train	672	231
Dev	150	59
Test	140	60

Table 2: Statistics of English tweet dataset in CheckThat! 2020 [3].

3.3 Re-splitting the Dataset

Although the splitting method of original dataset is not mentioned in its description, we find that tweets in the test set appear earlier than some of the tweets in the training set. Therefore, we guess that they shuffle the dataset before splitting. Although most work split dataset in this way, it has a shortcoming for this task.

In the real-world scenario, when a fact check-worthy tweet emerges, we only have previous tweets to train the detector. In terms of this, we think it is better to split dataset chronologically. Chronologically data splitting is also used by [26] in their work of fake news detection for the similar reason. By doing this, we can avoid the unnatural situation of using future information to predict past information.

Specifically, we make sure the following in our dataset:

- Creation time of tweets in training set is earlier than that of tweets in development set
- Creation time of tweets in development set is earlier than that of tweets in test set
- Creation time of user reactions in training set and development set is earlier than that of user reactions in test set

Besides, even though the description of the original dataset states that the tweets in their dataset are during March 2020, we find 6 tweets out of this range. One of them is created in 2013, and the other five are created in January 2020 or February 2020. For the difficulties in collecting user reactions for them, we do not include the 6 outliers in our dataset. The statistics of our user reaction augmented and chronologically split dataset is shown in Table 3.

(4)：Twitter data is provided by NTT Data.

	Check-worthiness	#target tweets	#user reactions
Train	check-worthy	229	1,264,486
	not check-worhty	389	2,782,114
	Total	618	4,046,600
Dev	check-worthy	57	294,141
	not check-worhty	93	618,073
	Total	150	912,214
Test	check-worthy	62	1,091,010
	not check-worhty	126	3,230,990
	Total	188	4,322,000
Total	check-worthy	348	2,649,637
	not check-worhty	608	6,631,177
	Total	956	9,280,814

Table 3: Statistics of our user reaction augmented and chronologically split dataset.

4 Proposed Method

In this section, we propose a method of detecting check-worthy tweets using early user reactions. We first introduce how we create the 50 linguistic features, 48 structural features, and 24 temporal features for a tweet by using its early retweets, quotes, and replies. We then introduce how we use a combined neural network model to combine our created numerical features with the outputs of baseline Transformer-based model using only the texts of target tweets.

4.1 Characterizing User Reactions

We characterize user reactions for analyzing and detecting check-worthy tweets. We refer to the tweet whose check-worthiness is to be determined as a target tweet. The target tweet is assumed to be posted by influential accounts with a large number of followers. These followers, and other users exposed to the target tweet, can share the target tweet to their own followers by retweeting. This is a great way to pass along news and interesting discoveries on Twitter.[5] When comments are added before retweet, they become quotes. In this case, followers can not only see the target tweet retweeted, but also the attached comments. A reply is a response to another user's tweet.[6] Users express their opinions to the target tweet by replies. They can also reply to each other to form a conversation thread. To characterize user reactions, we extract linguistic features from quotes and replies, structural features and temporal features from retweets, quotes, and replies.

a) Linguistic Features

We create linguistic features based on the following perspectives and the types of features are shown in Table 4.

Sentiment occurs in social media with a user's information sharing behavior [21]. We attempt to explore whether users express different sentiments to check-worthy tweets, which contain a claim [15] informing users of some key information.

(5) : https://help.twitter.com/en/using-twitter/how-to-retweet

(6) : https://help.twitter.com/en/using-twitter/mentions-and-replies

Sentimental features	Emotional features	Other features
(1) negative	(4) anger	(8) irony
(2) neutral	(5) joy	(9) offensive
(3) positive	(6) optimism	(10) hate
	(7) sadness	

Table 4: Linguistic features extracted from different perspectives.

Target tweet might inspire certain emotions of users [23], and then users express their emotions in the texts by quoting or replying. Since check-worthy tweets are assumed to have an effect on or be of interest to the general public [15], it is likely that emotions expressed by users are different from those of not check-worthy tweets, we thus take the factor of emotion into consideration.

Irony language, offensive language and hate speech are used by some users to express their individual biases against a person or a group of persons [24]. One of the characteristics of check-worthy tweets is that they do harm to the society, a person, a company, or a product [15]. For check-worthy tweets doing harm to certain group of persons, or countries (*e.g.*, the origin of COVID-19), individual biases might be triggered and expressed by users.

These features can all be calculated by using Twitter-RoBERTa-base model, which is trained on more than 58 million tweets, and then finetuned for specific tasks with the TweetEval benchmark [2]. We calculate feature values of these 10 types (negative, neutral, positive, anger, joy, sadness, irony, offensive, hate) for each target tweet, quote, and reply. We then create features for each target tweet. For a target tweet having n linguistic reactions (quote or reply), the ith ($i = 1, 2, ..., n$) linguistic reaction has feature values of 10 types. Feature value of the jth ($j = 1, 2, ..., 10$) type is denoted as s_{ij}. To aggregate linguistic reactions' feature values of the jth ($j = 1, 2, ..., 10$) type for a target tweet, we calculate their mean values, maximum values, minimum values, and median values by using

$$s_j^{mean} = mean(s_{1j}, s_{2j}, ..., s_{nj}) \quad (1)$$

$$s_j^{max} = max(s_{1j}, s_{2j}, ..., s_{nj}) \quad (2)$$

$$s_j^{min} = min(s_{1j}, s_{2j}, ..., s_{nj}) \quad (3)$$

$$s_j^{median} = median(s_{1j}, s_{2j}, ..., s_{nj}) \quad (4)$$

We also use the target tweet's feature value of the jth type itself as a linguistic feature for target tweet, which can be denoted as s_j^{self}. Then, for the jth feature type, we have 5 feature values for a target tweet. Since we have 10 feature types, we obtain totally 50 linguistic features for a target tweet.

b) Structural Features

User reactions have network structures, where the root is the target tweet. It can be retweeted, quoted, or replied. Then, these quotes and replies can also be retweeted, quoted, or replied. This is typically how information propagates on social media. In this process, user reaction networks can be denoted as graphs, where a node represents

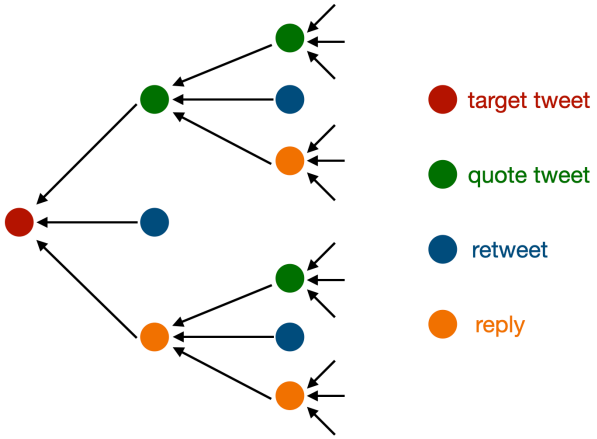

Figure 1: An example of user reaction network for a target tweet.

	Reaction to target tweet	Reaction to quote	Reaction to reply
Retweet	retweet of target tweet	retweet of quote	retweet of reply
Quote	quote of target tweet	quote of quote	quote of reply
Reply	reply to target tweet	reply to quote	reply to reply

Table 5: Nine types of link between user reactions.

a user reaction. An example of user reaction networks is shown in Figure 1.

There are four types of node in the networks, which are target tweet, retweet, quote, and reply. We assume that (i) target tweet is not a reaction of another user reaction, and (ii) a retweet can not be retweeted. Other than these two exceptions, the node of user reactions can form nine types of link, which are shown in Table 5. Since we assume that the information hidden behind these links might give a clue, we create features of link types node depth. Different from previous work that constructs a network for each reaction type [20], our proposed user reaction networks include all types of user reaction.

To characterize link types of reaction networks, for each target tweet, we use the quantity and proportion of each link type as its features. Since there are 9 types of link, we obtain 18 features for each target tweet. We also calculate:

- the quantity and proportion of retweet, quote, reply
- the quantity and proportion of reaction to target tweet, reaction to quote, and reaction to reply

We thus have additional 12 features and totally 30 features for each target tweet based on the link types of its reaction network.

To characterize node depth of reaction networks, for each target tweet, we calculate mean depth and maximum depth of the end node for different types of links. Notice that depth of the end node for retweet of target tweet, quote of target tweet, reply to target tweet always equal to 1. We calculate mean depth and maximum depth of the end node for the rest types of link. We then have 18 features for each target tweet on the node depth of its reaction network. Combined with the 30 features based on link types, we totally obtain 48 features as structural features for a target tweet.

c) Temporal Features

For some tweets, users react to them as soon as they are released; while for others, users react later. Likewise, some tweets are reacted more frequently and others more scarcely. Here, we design temporal features of user reactions by using elapsed time and adjacent time, denoted as T_{elapsed} and T_{adjacent}, respectively. For each user reaction, we use t_{reaction} to denote the creation time of itself, and t_{root}, t_{parent} to denote the creation time of its root, and parent, respectively, in the user reaction networks. We can then calculate T_{elapsed} and T_{adjacent} by using

$$T_{\text{elapsed}} = t_{\text{reaction}} - t_{\text{root}} \tag{5}$$

$$T_{\text{adjacent}} = t_{\text{reaction}} - t_{\text{parent}} \tag{6}$$

Then, for each target tweet, we capture the distribution of T_{elapsed} and T_{adjacent} of its retweets, quotes, and replies, respectively, by using mean value, maximum value, minimum value, and median value as features. For each one of the 3reaction types, we obtain 4 features for T_{elapsed} and 4 features for T_{adjacent}. We totally obtain 24 features as temporal features for each target tweet.

4.2 Combining User Reaction Features with Target Tweet Text Features

We use Multimodal-Toolkit [9] to incorporate numerical data of user reaction features with text data of target tweet text for check-worthiness detection. It uses transformers as the base model for text features and adds a combining module that takes the outputs of the transformer in addition to numerical features to produce rich multimodal features for downstream classification layers. Text data and numerical data will be combined in the following combining methods:

- Concatenate transformer output, numerical features all at once before final classifier layer.
- MLP on numerical features then concatenated with transformer output before final classifier layer.
- Gated summation of transformer outputs and numerical features before final classifier layer.
- Weighted sum of transformer outputs and numerical features for each feature dimension before final classifier layer.

5 Experiments

In this section, we evaluate our proposed method by using metrics of both classification and ranking. We compare the performances in different timing by controlling the number of user reactions used. We also compare the performances of the different 3 feature types. We then analyze the distributions of some features to investigate how

	f1	prec.	recall	map	mrr	rp	p1	p3	p5	p10	p20	p30
baseline	0.634	0.663	0.608	0.724	0.883	0.769	0.800	0.833	0.860	0.860	0.830	0.783
10u	0.647	0.666	**0.631**	0.733	**1.000**	0.774	**1.000**	0.867	0.880	0.870	0.830	0.783
30u	**0.652**	0.676	**0.631**	0.738	0.950	0.778	0.900	0.867	**0.920**	**0.890**	0.825	**0.787**
50u	0.651	**0.682**	0.624	**0.740**	0.950	**0.780**	0.900	0.900	0.840	0.870	**0.835**	0.780
90u	0.646	0.676	0.619	0.734	0.900	0.776	0.800	0.900	0.880	0.850	0.815	0.773
160u	0.645	0.679	0.614	0.738	0.950	0.777	0.900	0.867	**0.920**	**0.890**	0.825	**0.787**
200u	0.642	0.680	0.608	0.738	0.950	0.776	0.900	**0.967**	0.900	0.860	0.820	**0.787**

Table 6: Results of proposed method using all types of features in different stages.

	f1	prec.	recall	map	mrr	rp	p@1	p@3	p@5	p@10	p@20	p@30
baseline	0.634	0.663	0.608	0.724	0.883	0.769	0.800	0.833	0.860	0.860	0.830	0.783
linguistic features	0.635	0.663	0.613	0.727	0.950	0.769	0.900	**0.933**	0.880	0.880	**0.840**	**0.800**
structural features	0.644	0.664	0.626	0.725	0.800	0.772	0.600	0.800	0.860	0.860	0.815	0.790
temporal features	0.646	**0.682**	0.616	0.735	**1.000**	**0.778**	**1.000**	**0.933**	0.880	0.830	0.830	0.777
all features	**0.652**	0.676	**0.631**	**0.738**	0.950	**0.778**	0.900	0.867	**0.920**	**0.890**	0.825	0.787

Table 7: Results of proposed method using certain feature type of the earliest 30 user reactions.

the user reactions to check-worthy tweets are different from those to not check-worthy tweets.

5.1 Evaluation

We evaluate the performance of our proposed method when using all created features. We combine the user reaction features with the output of Transformer-based model RoBERTa. After validation by using development dataset, we select the combining method of MLP on numerical features then concatenated with transformer output before final classifier layer, and set the hyper parameter of transfer model as following. We set batch size to 32, learning rate to 2.5e-5, adam epsilon to 1e-6, number of epochs to 10, and keep the other parameters as default except seed. For each set, we run 10 times with seed=0,1,...,9, and calculate the mean value for each evaluation metrics. We use f1-score, precision, recall, map(mean average precision), mrr(mean reciprocal rank), rp(r-precision), p@k(k=1,3,5,10,20,30) as evaluation metrics with following considerations:

- Precision measures whether it can reduce the useless work to check the not check-worthy tweets.
- Recall measures whether it can reduce the ignored tweets which might bring negative effects.
- Metrics for ranking measures whether it can assist fact-checkers to set order of priority for their work.

a) Performance in Different Timing

Since time is very important in the real word scenario for fact-checkers, we explore the relationship between accuracy and earliness of our proposed method. We investigate the performance of using different number of user reactions in early stages. We set the number of earliest user reactions n as $n = 10, 30, 50, 90, 160, 200$. We use the data available in these different stages for training and prediction, and the results are shown in Table 6.

We find that in the stage of 10u, our proposed method perform the best in p@1, mrr; in the stage of 30u, it performs the best in f1-score, recall, p@5, p@10, p@30; in the stage of 50u, it performs the best in precision, map, and rp, p@20; in the stage of 160u, it performs the test in p@5, p@10, and p@30; in the stage of 200u, it performs the best in p@3 and p@30. Overall, even using only the earliest 10 user reactions can improve the performance for both classification and ranking. In particular, in the stage of 30u and 50u, our proposed method perform the best. This is probably because user reaction between check-worthy tweets and not check-worthy tweets differs more in the early stage, thus giving more important clues for estimation of check-worthiness.

b) Performance of Different Feature Types

We have created three types of features including linguistic features, structural features, temporal features, we assume different types of features might improve the performance in different ways. We then investigate how proposed methods perform when using only certain types of features. Results of proposed method using certain feature type of the earliest 30 user reactions is shown in Table 7.

Overall, when using all the three types of features, it perform better for both classification and ranking, having the highest f1-score, recall, map, rp, p@5, p@10. Besides, linguistic features perform the best in p@3, p@20, p@30; temporal features perform the best in precision, mrr, rp, p@1, and p@3. Although the performance of structural features is not very good, it also outperforms the baseline in most all the metrics.

5.2 Analysis

From the results of evaluation, the earliest 30 user reactions are valid to improve the accuracy of both classification and ranking, we then analyze the differences between the earliest 30 user reactions to check-worthy tweets and not check-worthy tweets based on our

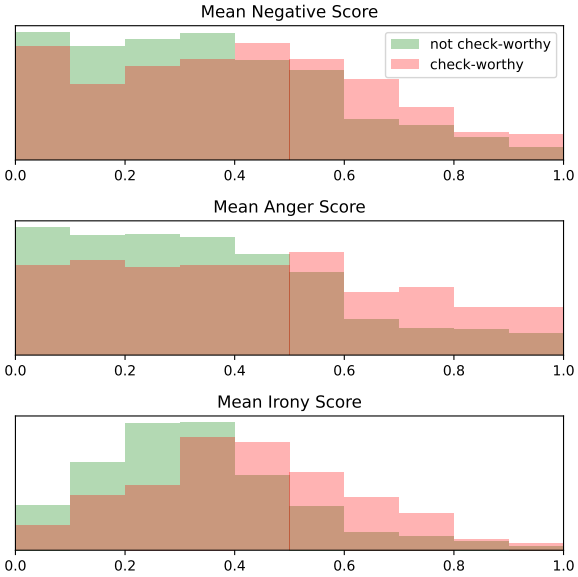

Figure 2: Distributions of linguistic features calculated by using the earliest 30 user reactions.

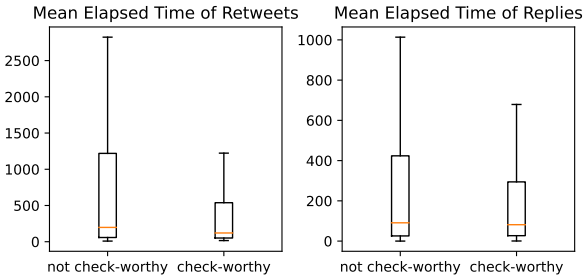

Figure 3: Distributions of temporal features calculated by using the earliest 30 user reactions.

created features.

The distributions of part of the linguistic features calculated by the earliest 30 user reactions to check-worthy tweets and not check-worthy tweets are shown in Figure 2. We find that users express more negative sentiment, emotion of anger, and use more irony language in their quotes or replies to check-worthy tweets.

We also compare how users react to check-worthy tweets and not check-worthy tweets in different earliness. We compare T_{elapsed} of retweets and replies for check-worthy tweets and not check-worthy tweets. From the results shown in Figure 3, we find that retweets and replies of check-worthy tweets appear earlier than those of not check-worthy tweets.

6 Conclusion

In this work, we extend existing dataset with user reactions and explore the relationship between user reaction and fact check-worthiness. We create 122 features of user reactions from the perspectives of language, structure and time. To simulate the real-world scenario for fact-checkers, We set 9 different stages where certain amount of user reactions can be used. We then combine these features with baseline Transformer-based model RoBERTa, and evaluate their performances.

Though feature analysis, we find that users express more irony language, more angry emotion, and more negative sentiments to check-worthy tweets than not check-worthy tweets. They also express less joyful emotion and less neutral sentiment to check-worthy tweets than not check-worthy tweets. Besides, check-worthy tweets and the reactions of them are retweeted more quickly and frequently than not check-worthy tweets and the reactions of them.

By evaluating our proposed methods, we find that accuracy for both classification and ranking can be improved by using user reactions. Besides, we find that using limited amount of user reactions perform better than using larger amount of user reactions for our proposed method.

In our future work, we will characterize user reactions in more detailed ways. Since we figure out that user reaction in different timing have different importance for check-worthiness detection, we will take consideration the timing for each user reaction rather than aggregating all the user reaction for a target tweet by typical statistical value. We will also characterize the text information of user reactions by using other model and extract more information from them. Then, we will set the timing more earlier and examine the least amount of user reactions needed for improving the accuracy of check-worthiness estimation.

Acknowledgement

This work was supported by JST CREST Grant Number JPMJCR19A4, and JSPS KAKENHI Grant Number JP21H03445, Japan.

References

[1] F Alam, S Shaar, A Nikolov, H Mubarak, GDS Martino, A Abdelali, F Dalvi, N Durrani, H Sajjad, K Darwish, et al. Fighting the covid-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. 2020.

[2] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1644–1650, 2020.

[3] Alberto Barrón-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, Fatima Haouari, Nikolay Babulkov, Bayan Hamdan, Alex Nikolov, et al. Overview of checkthat! 2020: Automatic identification and verification of claims in social media. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 215–236. Springer, 2020.

[4] Sylvie Cazalens, Philippe Lamarre, Julien Leblay, Ioana Manolescu, and Xavier Tannier. A content management perspective on fact-checking. In *Companion Proceedings of the The Web Conference 2018*, pages 565–574, 2018.

[5] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Jieyu Ding Featherstone and Jingwen Zhang. Feeling angry: the effects of vaccine misinformation and refutational messages on negative emotions and vaccination attitude. *Journal of Health Communication*, 25(9):692–702, 2020.

[8] Miriam Fernandez and Harith Alani. Online misinformation: Challenges and future directions. In *Companion Proceedings of the The Web Conference 2018*, pages 595–602, 2018.

[9] Ken Gu and Akshay Budhkar. A package for learning on tabular and text data with transformers. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 69–73, 2021.

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[11] Yonghun Jang, Chang-Hyeon Park, and Yeong-Seok Seo. Fake news analysis modeling using quote retweet. *Electronics*, 8(12):1377, 2019.

[12] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[14] Martin Müller, Marcel Salathé, and Per E Kummervold. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*, 2020.

[15] Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Watheq Mansour, et al. Overview of the clef–2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 264–291. Springer, 2021.

[16] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*, 2020.

[17] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[18] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[19] K Sharma, S Seo, C Meng, S Rambhatla, and Y Liu. Covid-19 on social media: Analyzing misinformation in twitter conversations. 2020.

[20] Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 626–637, 2020.

[21] Stefan Stieglitz and Linh Dang-Xuan. Emotions and information diffusion in social media—sentiment of microblogs and sharing behavior. *Journal of management information systems*, 29(4):217–248, 2013.

[22] Tetsuro Takahashi and Nobuyuki Igata. Rumor detection on twitter. In *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems*, pages 452–457. IEEE, 2012.

[23] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[24] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.

[25] Maxwell Weinzierl, Suellen Hopfer, and Sanda M Harabagiu. Misinformation adoption or rejection in the era of covid-19. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 787–795, 2021.

[26] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. Mining dual emotion for fake news detection. In *Proceedings of the Web Conference 2021*, pages 3465–3476, 2021.