Analysis of Reply-Tweets for Buzz Tweet Detection

Kazuyuki Matsumoto[1], **Yuta Hada**[2], **Minoru Yoshida**[1], **Kenji Kita**[1]
[1]Tokushima University, Graduate School of Technology, Industrial and Social Sciences
Minamijosanjima-cho 2-1, Tokushima, Japan
[2]Tech Information Corp., Technical support department,
Inubushihigashidani 6-23, Itano, Tokushiima, Japan
{matumoto;mino;kita}@is.tokushima-u.ac.jp

Abstract

In this study, we propose a method for predicting whether a tweet will create a buzz on the Internet by examining tweeted replies posted by others. We also investigate the distinguishing characteristics of replies to buzz tweets by analyzing feature amounts. Our proposed method first converts each reply tweet into a vector expression using a word distributed representation or some other vectorization method. We then apply a machine learning method for binary classification to determine whether the reply is to a buzz tweet or a non-buzz tweet. We classify the target tweet into "buzz" or "non-buzz" categories by comparing the total "buzz" and "non-buzz" scores produced by the classifier. The proposed method using StarSpace achieved 93.1% F1-score, while an approach that used number of retweets and number of favors ("likes") achieved 77.8% F1-score. We also found that there are a number of words that are characteristic of buzz tweet replies and a number of words that are characteristic of non-buzz tweet replies.

1 Introduction

In recent years, portable information and communication devices have become a common means by which individuals interact with one another via the Internet. Social networking sites such as Twitter, Facebook and Instagram are immensely popular. In Japan, the term "*Bazuru*" (or "buzz") is frequently used to describe a situation in which a topic expands dramatically in a short period of time, attracting the attention of many. The term typically refers to the rapid spread of a topic on the Internet through social media, etc. On Twitter, such a phenomenon is caused by followers or other Twitter users retweeting (RT), registering a "like" (or favor) or replying to a tweet.

In this paper, we describe a method for detecting a "buzz tweet" (i.e., a tweet that induces a "*Bazuru*" phenomenon) using reply tweets as features. This technique makes it possible to discover important tweets and topics from various viewpoints that were difficult to detect with conventional methods.

2 Related Works

Numerous studies have analyzed trend keywords on the web or social media (Cataldi et al., 2010),(Lau et al., 2012),(Cheong and Lee, 2009),(Yu et al., 2011),(Naaman et al., 2011),(Kaushik et al., 2015). Twitter, in particular, is one of the more popular targets of such studies, as it has superior real-time posting. Many of these studies have attempted to detect keywords or topics that sharply increase usage rates and analyze the stream of times considering the keywords as trend keywords. Their main focus has been on analyzing the relation between keywords in the posted tweets and trends in the real world and assessing the factor of the trend.

To clarify the mechanism by which a "buzz" is created, we believe that it is important to determine the distinctive features of buzz tweets by identifying the various types of buzz tweets that exist and analyzing the responses they produce. As a method to grasp the scale of the response, numerical information such as the number of retweets or favors

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 138-146, Hakodate, Japan, September 13-15, 2019

("likes"), as well as the number follows and various follower characteristics, can be useful. However, to fully assess the response to a target buzz tweet, it is necessary to analyze the contents of the replies to the posted tweet.

Counting the number of the retweets is not a particularly reliable indicator of "buzz," as retweeting is a mechanical and easy way to produce information diffusion. This renders simply counting the number of retweets of a given posted tweet a rather limited way to distinguish a true buzz phenomenon from an artificially created one. Various methods (Zaman et al., 2010),(Suh et al., 2010),(Morchid et al., 2014),(Firdaus et al., 2018) that predict the scale of information diffusion based on a change in relationships of users such as the number of follows, or followers, or retweets or favors have been proposed. However, none are capable of determining whether the information diffusion resulting from a particular tweet is due to a true popular phenomena.

There have also been studies that estimate the probability of retweeting by using the correspondence relation between the contents of a tweet and the interest of users (Imamori and Tajima, 2016). However, having interest is not equal to having a favorable opinion of a tweet, which makes it difficult to distinguish a buzz tweet from a "flame."

Another study (Deusser et al., 2018) used the metadata of articles posted on Facebook as features to predict the popularity of an article. The authors of this study used binary classification to determine whether an article is popular by employing a machine learning method. Here, the contents of the article are not used as a feature, as the authors believed that the interaction between an article 's author and his/her friends (viewers) was more related to buzz. In the case of Facebook, the probability that an article will be read by complete strangers is lower than on Twitter. There are thus fewer uncertain elements and the amount of noise in feature values is thought to be smaller. For this reason, such a method may not be particularly effective for evaluating postings on Twitter.

3 Buzz Tweet Classification Method

3.1 Definition of problem

In this study, buzz tweets are tweet contents that have a large number of RTs, replies, likes, give strong impact to readers, and are sympathetic to many people, or that are accompanied by photos or videos. To obtain such tweets, in this paper we define buzz tweets as those listed on the websites (curation sites) that collects buzz tweets filtered by numerical indicators such as the number of RTs.

3.2 Target data

To detect a buzz tweet, it is necessary to collect buzz tweets as training data. For that purpose, a definition of "buzz tweet" is necessary. Because the definition of a buzz phenomenon is ambiguous and it is difficult to establish a clear basis for determining buzz, we collected buzz tweets from various buzz tweet roundup websites.

Many of the roundup sites determine the buzz/non-buzz status of a tweet by using the number of retweets, i.e., the tweet 's degree of diffusion. However, if buzz tweets are identified solely by counting the number of retweets, tweets by celebrities or flame tweets are more likely to be identified as buzz tweets.

In this study, we considered "flame" and "buzz" as different phenomena. Generally, "flame" indicates that a tweet has attracted negative attention, while "buzz" is associated with largely positive responses.

Tweets posted by famous persons such as entertainers, politicians, athletes and YouTubers tend to be diffused at a higher rate than those of general users, which means that their tweets often become buzz tweets. It is natural that the tweets of authors with more fans or friends will produce more retweets or favors, which increases the probability of buzz. When we surveyed the buzz tweet roundup websites, we found that there were several tweets posted by famous persons, but they were small in number. Therefore, we decided not to remove such tweets from the buzz tweets used in our study. Nevertheless, we recognize that diffusion in the case of tweets by famous persons reflects the attributes of the person rather than the contents of the tweet,

Type	# of replies	Avg. # of RTs	Avg. # of Likes
Buzz	13218	30253.6	80845.8
Non-Buzz	16012	2425.0	12925.0
Total	29230	20977.4	58205.5

Table 1: Number of replies and the average # of RTs/Likes.

which makes it difficult to establish them as representing a true buzz phenomenon.

Additionally, as pictures or videos can be posted on Twitter, tweets using such non-verbal media tend to attract more attention and, as a result, are more likely to be identified as buzz tweets.

The following are the websites from which we collected the buzz tweets used in the study. The period of collection was from December 2018 to August 2019:

Collection source of roundup websites:

- iitwi: `https://service.webgoto.net/iitwi/`
- Matome site: `https://matome.naver.jp/odai/2150908164548234501`

In this paper, we validate whether it is possible to classify buzz/non-buzz tweets by using reply tweets. Non-buzz tweets are posted by famous persons with a larger number of retweets or favors. The tweets of famous persons were selected from the tweets of users who ranked in the top 500 in number of followers during the period from 2016 through April 2019.

The number of unique user accounts was 150 for buzz tweets and 151 for non-buzz tweets. For each of the targeted buzz/non-buzz tweets, we collected replies. Table 1 shows the number of replies and the average number of Retweets/Likes for the buzz/non-buzz tweets used in the study.

Because Twitter 's API is unable to collect replies to specific tweets, we manually collected the replies that could be viewed. We defined seven categories, from A to G, for the buzz tweets based on their factors of buzz. A breakdown of the 150 buzz tweets is shown in Table 2.

As shown in the table, categories C (buzz tweets due to images or videos) and F (buzz tweets due to jokes or funny behavior/utterance) make up the vast majority of the buzz tweet factors. Table 3 shows the example of buzz tweet and its replies.

Category	Example (Factor)	Count
A	knowledge	13
B	surprise news	3
C	image, video	61
D	celebrity news, information	7
E	moral, social remarks	11
F	joke, funny behavior/utterance, etc.	51
G	common thing	4

Table 2: Number of tweets for each buzz category.

3.3 Flow of the buzz tweet classification

Our proposed method constructs a binary classifier (buzz/non-buzz) that uses the reply texts posted to buzz/non-buzz tweets as features and uses the buzz/non-buzz tweet to which a reply text is posted as a label.

To judge whether an unknown tweet is a buzz or non-buzz tweet, a buzz/non-buzz classification score is produced by the classifier for each posted reply to the tweet. These scores are then aggregated to produce a total classification score for each of the two classifications (buzz/non-buzz). The larger of the two scores determines whether the unknown tweet should be judged a buzz tweet or a non-buzz tweet.

Eq.1 and 2 show the buzz score calculation and label judgement criteria. $Prob_{i,x}$ indicates the probability of label x estimated by the classifier for reply i posted to tweet t; $label_t$ shows the result of the label judgment for tweet t, determined by comparing the magnitude of the two total label scores.

$$Score_x = \sum_{i=1}^{N} Prob_{i,x} \tag{1}$$

$$label_t = \begin{cases} buzz & (Score_{buzz} > Score_{nonbuzz}) \\ nonbuzz & (Score_{buzz} \leq Score_{nonbuzz}) \end{cases} \tag{2}$$

The flow of the proposed method is shown in Fig.1.

3.4 Conversion of reply tweet into vector

It is not easy to identify features from the reply tweets posted to a buzz tweet without formatting. Therefore, we embedded each reply into the feature

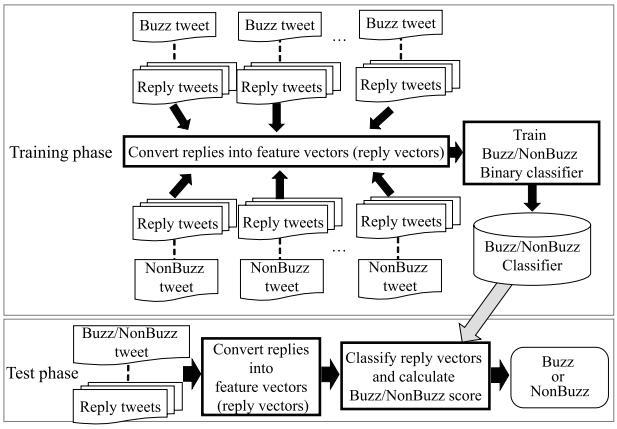

Figure 1: Flow of the proposed method.

space. Recently, techniques such as word2vec that express words or sentences with fixed length vectors are being used not only in studies but also in various actual services.

Buzz tweet (# of RT: 13034, # of Favorite: 23504)
If the insistence that "an artist is arrested and all the work of that person can not be used" is strictly used, perhaps the most significant impact on Japan at the time of becoming a suspect is probably "Illust-ya's creator".
Example of replies
Ex.1) Conspiracy of people who are trying to capture the share of "Illust-ya" starts to move. I understand. **Ex.2)** I think Asei Kobayashi is quite good (lol). ← I checked again by chance.

Table 3: Example of buzz tweet and its replies.

In this study, we employ a method to convert reply data into vectors by using unsupervised pre-training based on a large-sized corpus. By pre-training the reply vectors, we are able to create a buzz/non-buzz judgement model that is robust to unknown words based on a small-sized corpus.

As a baseline, we apply the dimensional compression method that applies TF-IDF-based keyword weighting to bag of words vectors. This method is often used for document retrieval or document classification.

In this study, we used the following vectorizing methods and machine learning models:

- Averaged word vector (AWV)
- CNN, bi-LSTM, bi-GRU
- character-AutoEncoder-Decoder trained by CNN, LSTM, and GRU
- StarSpace
- BERT (Bidirectional Encoder Representations from Transformers)
- Baseline: bag of words vector (tfidf-weight)

The following subsections explain each method in sequence.

3.4.1 Averaged word vector(AWV)

This method employs the averaged vector(AWV) of a word distributed representation trained by the fastText (Joulin et al., 2016) algorithm using a Japanese tokenized corpus. Because the fastText algorithm can consider the sub word information of words, this method is more robust to unknown words than word2vec. The buzz/non-buzz binary classifier was created by training feed forward neural networks (FFNN) using averaged vectors as features.

In our experiment, we use 300 dimension distributed representations that were trained based on Japanese Wikipedia articles.

3.4.2 CNN, bi-LSTM, bi-GRU

Here we created a buzz/non-buzz binary classifier by training Convolutional Neural Networks (CNN), Bidirectional Long-short Term Memory (bi-LSTM), Bidirectional Gated Recurrent Unit (bi-GRU) using pre-trained word distributed representations as features, which is the same as the features used in the averaged word vector method.

Because the length of the various reply texts differs, we applied padding to the reply data as preprocessing. From the average number of words, we set the maximum word number as 30.

We also created a classifier by neural network using a simple attention mechanism (Luong et al., 2015). The structure of the self-attention bi-LSTM network is shown in Fig.2.

3.4.3 Character-AutoEncoder-Decoder trained by CNN, LSTM, and GRU

Because the reply texts include several character strings that are difficult to divide morphologically, we applied training per character. We first created

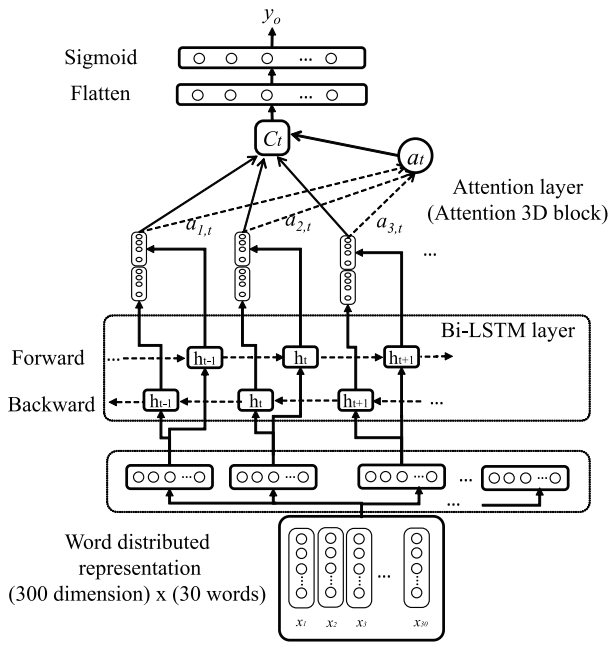

Figure 2: Bi-LSTM attention network.

	Layer type	In/out	Tensor shape
Encoder	Input	in:	(None, 140, 500)
		out:	(None, 140, 500)
	Conv1D	in:	(None, 140, 500)
		out:	(None, 140, 256)
	MaxPooling1D	in:	(None, 140, 256)
		out:	(None, 70, 256)
	Conv1D	in:	(None, 70, 256)
		out:	(None, 70, 128)
	MaxPooling1D	in:	(None, 70, 128)
		out:	(None, 35, 128)
	Conv1D	in:	(None, 35, 128)
		out:	(None, 35, 64)
Decoder	Conv1D	in:	(None, 35, 64)
		out:	(None, 35, 64)
	UpSampling1D	in:	(None, 35, 64)
		out:	(None, 70, 64)
	Conv1D	in:	(None, 70, 64)
		out:	(None, 70, 128)
	UpSampling1D	in:	(None, 70, 128)
		out:	(None, 140, 128)
	Conv1D	in:	(None, 140, 128)
		out:	(None, 140, 256)
	Conv1D	in:	(None, 140, 256)
		out:	(None, 140, 500)

Figure 3: CNN auto encoder-decoder note: ('None' is the batch dimension).

character-based one-hot-vectors (maximum character length: 140), then trained an encoder-decoder that reproduced the original texts by using CNN, LSTM and GRU. By using the output of the encoder of the trained model, we converted the reply texts into fixed length vectors.

The buzz/non-buzz binary classifier was created by training the FFNN using the obtained vector as input. Fig. 3 shows the character-based AutoEncoder-Decoder by CNN. The AutoEncoder-Decoder based on LSTM and GRU, respectively, consists of four layers for the encoders and three layers for the decoders. There were 15 training epochs for CNN, 50 for LSTM, and 8 for GRU.

3.4.4 StarSpace

The "StarSpace" (Wu et al., 2018) algorithm converts text into distributed representations. Because StarSpace can learn effective distributed representations for the text classification task, we were able to create a model to classify replies accurately without pre-learning.

We trained a one-to-one classifier (which estimates one label for the one inputted text) by StarSpace, and used it to classify replies as buzz/non-buzz. We set the parameters n of the word n-gram at 3 and the number of word distributed representation dimensions at 100. We used Sentencepiece (Sentencepiece,) as a tokenizer with vocabulary size is 10000.

3.4.5 BERT(Bidirectional Encoder Representation from Transformers)

BERT, developed by Google (Devlin et al., 2018), is a model that can produce versatile distributed representations. To apply the BERT model to Japanese reply texts, we generated 768 dimensional distributed representation vectors by using the pre-trained BERT model trained with Japanese Wikipedia articles (BERT,). Using the vectors as feature vectors, we created a buzz/non-buzz classifier using a perceptron without hidden layers.

3.4.6 Baseline: bag of words vector(TF-IDF)

As feature words, we selected words with high importance values (applying a threshold) based on TF-IDF values, set vector dimensions, respectively, for the important words, and created vectors with TF-IDF values as dimension values. We constructed a buzz/non-buzz classifier by FFNN using the TF-IDF vectors as features. In this paper, we removed

words with TF-IDF values under the threshold and decided on 197 dimensions.

4 Experiment

4.1 Preliminary experiment

As features other than the reply tweets to the buzz/non-buzz tweets, numerical measures such as the number of users following the poster, the poster ' s number of followers, and the number of retweets, favors, etc., as well as features obtained from the tweets themselves or from user profiles, were available. We conducted a preliminary experiment to evaluate variations of a classifying method based on combining these features. A support vector machine (SVM) was used to train the classifier. We divided the data by 10 and conducted a cross validation to evaluate the performance of the method. We used Recall, Precision and F1-score as evaluation scores (see Eq.3, 4, 5). TP_x means true positive of label x, FP_x means false positive of label x, and FN_x means false negative of label x.

$$R(Recall)_x = \frac{TP_x}{TP_x + FN_x} \times 100 \quad (3)$$

$$P(Precision)_x = \frac{TP_x}{TP_x + FP_x} \times 100 \quad (4)$$

$$F1(F1 - Score)_x = 2 \times \frac{R_x \times P_x}{R_x + P_x} \quad (5)$$

Table 4 shows the features used in the experiment. Results from the preliminary experiment are shown in Table 5. As indicated, the M2 feature combination produced the highest F1-score level (77.8%) for buzz tweets. The lowest F1-Score for buzz tweets of combinations M6 shows that the number of Follows/Followers and total number of favors ("likes") were more important than the features obtained from the tweets themselves (feature ID: 14). With the exception of combination M7, the various feature combinations showed higher F1-Score in the non-buzz tweet classification than in the buzz tweet classification.

These results suggest that famous persons who posted target non-buzz tweets might have distinctive characteristics with respect to the number of retweets or favors.

Because the buzz tweet classification F1-Scores were all under 80% in the preliminary experiment, we concluded that features from the buzz tweets themselves and user account information were not suitable for identifying buzz tweets.

ID	Feature type
1	# of replies
2	# of Retweets
3	# of Favors
4	# of Follows by the account
5	# of Followers of the account
6	Total # of Favors
7	Total # of List
8	Total # of Moment
9	Total # of tweets
10	Elapsed days from the date when the account registered
11	Whether the account is locked
12	Whether an image is attached
13	The # of characters of the tweet
14	The averaged word vector of the tweet (300 dimension)
15	The averaged word vector of the profile text (300 dimension)

Table 4: Feature type.

Method	Buzz			Non-Buzz		
(Feature IDs)	R	P	F1	R	P	F1
M1 (1-13)	84.8	71.1	77.4	75.4	87.4	81.0
M2 (1-12)	84.9	71.8	**77.8**	**75.9**	87.4	**81.2**
M3 (1-3, 12)	62.0	53.7	57.6	59.6	67.5	63.4
M4 (1-3)	**89.7**	17.4	29.2	54.6	**98.0**	70.1
M5 (4-11)	85.0	68.5	75.8	73.9	88.1	80.4
M6 (13, 14)	55.3	17.4	26.5	51.4	86.1	64.4
M7 (15)	50.0	**84.8**	62.9	46.3	13.4	20.8

Table 5: Result of preliminary experiment.

4.2 Evaluation experiment

Table 6 shows the experimental results when features of the replies were used. We conducted our cross validation test by using the same data divided into 10 groupings as in the preliminary experiment. As indicated, 94.7% classification precision for buzz tweets was achieved when BERT was used. In the case of AWV, TF-IDF, bi-LSTM+Attention, and LSTM-AE, the classification precisions for buzz tweets were over 85%.

We think the reason of the highest F1-score of StarSpace is mainly due to supervised learning of word embedding. On the other hand, BERT and the

other methods used unsupervised pre-trained word embedding.

The baseline method using TF-IDF produced 85.0% precision for buzz tweets, which was not particularly low. In fact, the classification recall for buzz tweets was over 94%.

Method	Buzz			Non-Buzz		
	R	P	F1	R	P	F1
AWV	89.3	86.5	87.9	86.1	89.0	87.5
TF-IDF	94.7	85.0	89.6	83.4	94.0	88.4
CNN	93.3	83.8	88.3	82.1	92.5	87.0
bi-LSTM	95.3	84.1	89.4	82.1	94.7	87.9
bi-GRU	**96.7**	83.3	89.5	80.8	**96.1**	87.8
bi-LSTM + Attention	95.3	86.1	90.5	84.8	94.8	89.5
CNN-AE	80.0	82.2	81.1	82.8	80.6	81.7
LSTM-AE	88.0	86.3	87.1	86.1	87.8	87.0
GRU-AE	90.0	84.9	87.4	84.1	89.4	86.7
StarSpace	94.7	91.6	**93.1**	91.4	94.5	**92.9**
BERT	88.2	**94.7**	91.3	**94.3**	87.4	90.7

Table 6: R, P, F1 of each method.

5 Analysis and Discussion

One of the primary aims of our study was to analyze the features of buzz tweets. Accordingly, in this section, from the training results of the classifier, we present our analysis of the distinguishing characteristics of replies to buzz tweets and replies to non-buzz tweets. We first randomly extracted 10000 reply vectors based on BERT. We compressed the vectors into two dimensions using the t-SNE algorithm (Maaten and Hinton, 2008) and plotted them in two-dimensional space. As can be seen in Fig.4, the replies to buzz tweets and non-buzz tweets are not clearly divided into buzz and non-buzz groupings.

To analyze this plot, the two-dimensional reply vectors compressed by t-SNE were clustered into eight clusters by k-means algorithm. Among these clusters, there were two clusters (cluster1, cluster4: these clusters are "magenta" and "lime" in Fig.5) where the numbers of non-buzz replies were twice as large as the numbers of buzz replies. As for these two clusters, we investigated the frequently appearing expressions in non-buzz replies by calculating the frequency of word appearance. The feature expressions frequently appeared in the non-buzz replies are shown in Table 7.

This result showed that among the non-buzz

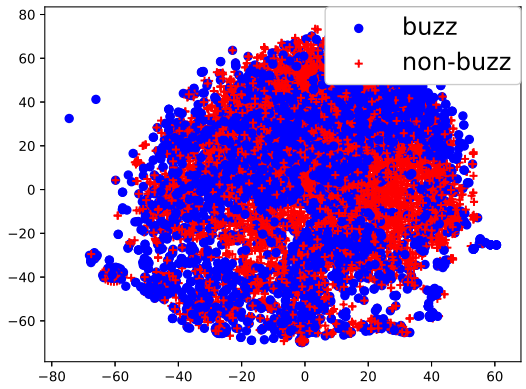

Figure 4: t-SNE plotting of BERT reply vectors.

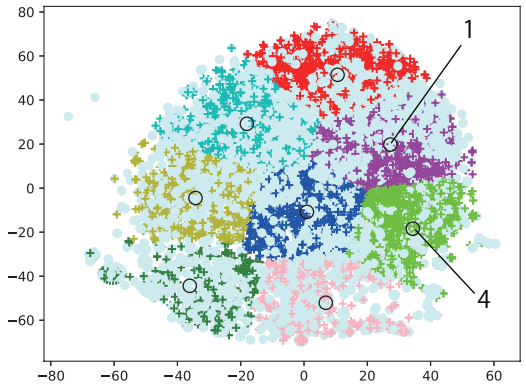

Figure 5: 8-clusters by k-means.

replies, there were comments on the events such as the concert or on their performance on TV programs, expression of thanks, support messages from the fans to their admiring famous persons. Next, we analyzed the differences between buzz and non-buzz tweets according to the word distributed representations obtained by training the reply texts, using features in the replies to buzz and non-buzz tweets based on the distributed representations trained by StarSpace.

StarSpace classifies texts by calculating the inner product of the label distributed representations and the vector summation of the distributed representations of the words in the texts. Therefore, we believed that the feature words for buzz/non-buzz could be obtained by calculating the similarity be-

Cluster	Frequently expressions in non-buzz replies
1	*Arigato* (Thank you) *Omedeto* (Congratulations) *ouen* (support)
4	*Tanoshimidesu* (I 'm looking forward it) *Yoroshiku* (Glad to see you) *Saikoudeshita* (It was the best)

Table 7: Frequently expressions in non-buzz replies.

tween the word distributed representations and the label distributed representations of buzz/non-buzz.

A partial list of the feature words is provided in Table 8. The numerical values indicate the cosine similarity with the given label. The table shows that in the replies posted to buzz tweets, distinctive expressions such as "*buzztteru*," "RT," and "FF" (an abbreviation of Follow/Follower) appeared, all of which are related to buzz phenomena on Twitter.

In contrast, in the replies posted to non-buzz tweets, there were many proper nouns (names/nicknames of famous persons, affiliations, etc.), as well as emojis (emoticons) or greeting expressions. This is thought to be because the fans (primarily, followers) of famous persons often post relatively polite replies that include greeting expressions or emotional expressions with many emojis.

Because many of the buzz tweet authors are not famous persons, they typically have a relatively small number of followers. Therefore, the attributes of the reply users are not limited to followers and cover a wider range than the reply users of famous figures. This may be one important reason why replies would be effective features for buzz/non-buzz classification.

On the other hand, proper nouns did not appear with exceptional frequency in the replies to buzz tweets. However, a large number of admiration expressions; such as "*warota*"(means have laughed), "genius" were found among the expressions that appeared in the buzz tweet replies.

6 Conclusions

In this paper, we proposed a method to classify buzz tweets by using reply features, which contain much

Buzz			
RT (Retweet)	0.90	***buzztteru***[1](now buzzing)	0.81
warota[1] (laughed)	0.89	***maji***[1](really)	0.81
kusa[1] (laugh)	0.87	*Garigarigarikuson*[2]	0.81
dekusa[1] (laugh)	0.84	Snape teacher	0.79
FF (Follow-Follower)	0.83	Excel	0.78
operation	0.83	station	0.77
Wakuwakusan[2]	0.83	*waratta* (laughed)	0.74
foreign citizen	0.82	*tensai* (genius)	0.73
Mac	0.82	(; ∀;)[3]	0.69
Non-Buzz			
Yoshimoto	0.94	politician	0.87
election	0.93	**emoji** (cherry blossom)	0.86
I'm looking forward to it	0.92	It was the best	0.86
pleasure	0.90	I support you	0.86
Sugichan[2]	0.89	***Murosan***[2]	0.85
Korea	0.87	participation	0.82
TV	0.87	***KeisukeHonda***[2]	0.85
jerky	0.87	**emoji** (dazzle)	0.83

Table 8: Example of terms that are important in each category.

richer information than numerical information such as the number of replies or favors. The proposed method converts the replies posted to buzz tweets into feature vectors and constructs a buzz tweet classification model by training the vectors with a machine learning method.

Based on results from an evaluation experiment and treating tweets by famous persons as non-buzz tweets, the method using StarSpace with SentencePiece tokenizer classified buzz tweets with 93.1 F1-score. This score is significantly higher than that of other methods that use such features as the number of retweets, number of follows, etc., all of which achieved less than 80 F1-score.

In the future, we plan to analyze whether the level of accuracy would change if we consider the posting times of the replies. According to our analysis of the differences in feature words between buzz and non-buzz tweet replies, we perceived a certain bias in the appearance of expressions related to differences in various attributes of the replying users.

As part of our extended investigation, we intend to determine whether the proposed method using reply features is capable of distinguishing buzz tweets

[1]Japanese slang
[2]Name of famous person
[3]Japanese emoticon

from flame tweets through additional experiments. We also plan to include additional features such as whether the replying user has a relationship to the author of the original tweet (e.g., is a follow or follower of the author), as this would seem to be an important feature for classification.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP18K11549.

References

Mario Cataldi, Luigi D. Caro, and Claudio Schifanella. 2010. Emerging topic detection on Twitter based on temporal and social terms evaluation *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, Article 4.

JeyHan Lau, Nigel Collier, and Timothy Baldwin. 2012. On-line Trend Analysis with Topic Models: #twitter trends detection topic model online, *Proceedings of COLING 2012: Technical Papers* 15191534.

Marc Cheong and Vincent Lee. 2009. Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base, *Proceedings of the 2nd ACM workshop on Social web search and mining*, 1-8.

Louis Yu, Sitaram Asur and Bernardo A. Huberman. 2011. What Trends in Chinese Social Media, *Proceedings of The 5th SNA-KDD Workshop '11 (SNA-KDD '11).*

Mor Naaman, Hila Becker and Luis Gravano. 2011. Hip and Trendy: Characterizing Emerging Trends on Twitter, *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, 62(5):902918.

R. Kaushik, S. Apoorva Chandra, Dilip Mallya, J. N. V. K. Chaitanya and S. Sowmya Kamath. 2015. Sociopedia: An Interactive System for Event Detection and Trend Analysis for Twitter Data, *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics*, 63-70.

Tauhid R. Zaman, Ralf Herbrich, Jurgen V. Gael and David Stern. 2010. Predicting Information Spreading in Twitter, *Computational Social Science and the Wisdom of Crowds Workshop (colocated with NIPS 2010).*

Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. 2010. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network, *2010 IEEE Second International Conference on Social Computing*.

Mohamed Morchid, Georges Linares, and Richard Dufour. 2014. Characterizing and Predicting Bursty Events: The Buzz Case Study on Twitter, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 27662771.

Syeda N. Firdaus, Chen Ding and Alireza Sadeghian. 2018. Retweet: A popular information diffusion mechanism A survey paper, *Online Social Networks and Media*, 6(2018), 26-40.

Daichi Imamori and Keishi Tajima. 2016. Predicting Popularity of Twitter Accounts through the Discovery of Link-Propagating Early Adopters, *Proceedings of Conference of Information and Knowledge Management (CKIM2016)*.

Clemens Deusser, Nora Jansen, Jan Reubold, Benjamin Schiller, Oliver Hinz and Thorsten Strufe. 2018. Buzz in Social Media: Detection of Short-lived Viral Phenomena, *WWW '18 Companion Proceedings of the The Web Conference 2018*, 1443-1449.

Armand Joulin, Edouard Grave, Piotr Bojanowski and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Volume 2, Short Papers, 427431.

Thang Luong, Hieu Pham and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP2015)*, 1412-1421.

Ledell Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, and Jason Weston. 2018. StarSpace: Embed All The Things!, *Proceedings of the thirty-second AAAI conference on artificial intelligence (AAAI-18)* , 5569-5577.

Sentencepiece: https://github.com/google/sentencepiece

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805.

BERT Japanese Pretrained Model: http://nlp.ist.i.kyoto-u.ac.jp/index.php?BERT

Laurens V. D. Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE, *Journal of Machine Learning Research*, 9(2008), 2579-2605.