# Assessment of Tweet Credibility with LDA Features

Jun Ito, Hiroyuki Toda, Yoshimasa Koike
NTT Service Evolution Laboratories,
NTT Corporation
Yokosuka, Kanagawa, Japan
{ito.jun,toda.hiroyuki,koike.y}@lab.ntt.co.jp

Jing Song, Satoshi Oyama
Graduate School of Information Science and
Technology, Hokkaido University
Sapporo, Hokkaido, Japan
songjing@complex.ist.hokudai.ac.jp
oyama@ist.hokudai.ac.jp

## ABSTRACT

With the fast development of Social Networking Services (SNS) such as Twitter, which enable users to exchange short messages online, people can get information not only from the traditional news media but also from the masses of SNS users. However, SNS users sometimes propagate spurious or misleading information, so an effective way to automatically assess the credibility of information is required. In this paper, we propose methods to assess information credibility on Twitter, methods that utilize the "tweet topic" and "user topic" features derived from the Latent Dirichlet Allocation (LDA) model. We collected two thousand tweets labeled by seven annotators each, and designed effective features for our classifier on the basis of data analysis results. An experiment we conducted showed a 3% improvement in Area Under Curve (AUC) scores compared with existing methods, leading us to conclude that using topical features is an effective way to assess tweet credibility.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

## Keywords

Information Credibility, Social Media, Twitter, Topic Model

## 1. INTRODUCTION

Twitter[1], a microblogging service, has emerged as a new medium that enables people to find out things that are happening when they are happening. Twitter propagates information much faster than traditional news media like newspapers and television [7, 8]. Twitter users can post and exchange 140-character-long messages known as tweets, and this system limitation facilitates real-time propagation of information to a large group of users. Over 284 million monthly active users generate about 500 million tweets per day[2], and this huge activity supports the information distribution environment on Twitter.

We can get information from Twitter quickly and easily, but sometimes we get spurious or misleading information. According to the research of Mendoza et al. [10], baseless rumors caused insecurity and chaos during the Chilean earthquake of 2010. Other researchers also carried on research to detect the credibility of tweets propagated on Twitter in an emergency situation [13, 14]. Thus, a major research topic is to evaluate information credibility on SNS for solving social problems such as hoax spreading.

Our main contributions in tackling the problem of assessing the credibility of trendy tweets are as follows.

**(i)** We showed basic analysis results on how people judge the credibility of a tweet from 2,000 trendy tweets posted in Japan in April 2014.

**(ii)** We proposed methods to assess information credibility of a tweet by using two new features, the "tweet topic" and "user topic" features derived from the Latent Dirichlet Allocation (LDA) model. We also conducted experiments to verify their effectiveness.

**(iii)** We built two hypotheses based on a user's "expertness" and "bias" and designed four methods to extract additional features. We conducted experiments to reveal which hypothesis is correct and which method works.

## 2. RELATED WORK

Many researchers have an interest in what credibility is and how people judge it [5, 11, 12]. Fogg and Tseng discussed the credibility of computers in 1999 [5]. They described credibility as a perceived quality composed of multiple dimensions and advocated that there are four types of credibility: presumed, reputed, surface, and experienced. Morris et al. [11] focused on how people evaluate the credibility of tweets. They conducted various kinds of experiments and showed that user names and user images affect people's judgment. O'Donovan et al. [12] analyzed the distribution of the salient features on Twitter that can be used to find interesting, newsworthy, and credible information. Their results show that the best indicators of credibility include URLs, mentions, retweets, and tweet length and that salient features occur more prominently in data describing emergency and unrest situations. These studies provided us directions about how to choose features of tweets in pursuing our goal of evaluating the credibility of tweets automatically.

[1] `https://twitter.com`

[2] `https://about.twitter.com/company`


*WWW 2015 Companion,* May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742569.

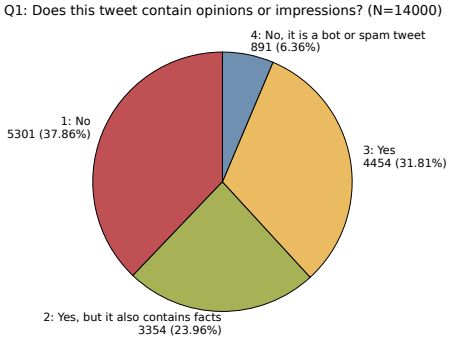


**Figure 1: Answer results for Q1.**

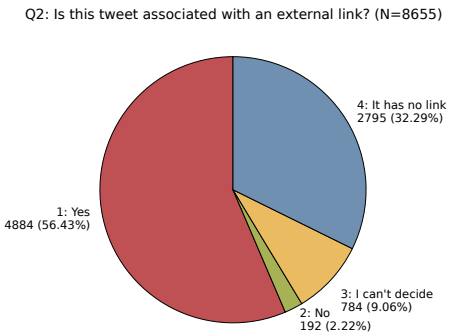


**Figure 2: Answer results for Q2.**

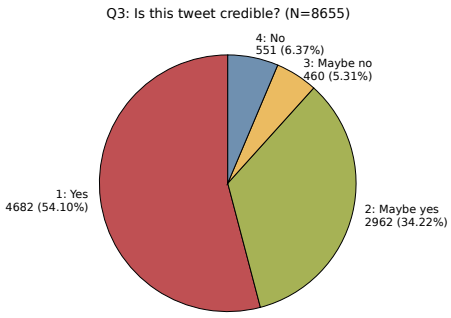


**Figure 3: Answer results for Q3.**

As we have in our work, a number of researchers have addressed the problem of how to assess tweet credibility. Castillo et al. [3] utilized four types of features (message-based, user-based, topic-based, and propagation-based) to make a classifier for evaluating the credibility of tweets. In their research, they focused on the level of credibility of every trend on Twitter. In contrast, we focused on how to access the credibility of every tweet instead of every trend. Gupta et al. [6] proposed a credibility analysis approach enhanced with event graph-based optimization. Their hypothesis is that tweets written about the same event should have similar credibility scores. Unlike us, they focused on the similar credibility scores of events and did not consider the users' topic distribution.

# 3. DATA COLLECTION AND ANALYSIS

We explain how we collected our data in Sec. 3.1 and results we got in analyzing it in Secs. 3.2–3.6.

## 3.1 Data Collection

**Collecting Tweets:** We accessed Twitter's trends/place API[3] every five minutes to get trendy words in Japan during April 2014. After that, we checked whether the trendy words also appeared in Google News[4] titles at that time. Words that did were removed from the trendy words list. In this way, we were able to get the trendy words appearing in relatively more news items. Then, the first author extricated the ten trends shown in Table 1 by referring to the remaining trendy words. We randomly collected 200 tweets with trendy words for each trend from preliminarily collected tweets by using Twitter's statuses/sample API[5]. One hundred tweets had unduplicated URLs, and the remaining 100 did not have URLs or duplicated text. In the end we collected 10 trends, with 200 tweets for each trend.

**Annotating Credibility:** We requested the annotators to label the credibility of every tweet collected. We employed 14 annotators who were widely distributed by age and sex and who were all used to Twitter. In the process of evaluating tweet credibility, we asked seven randomly assigned annotators to answer the four questions for each tweet. The annotators were allowed to see the tweet's text, posted time, user name, and webpages (if URLs were in the tweet). In

[3]`https://dev.twitter.com/rest/reference/get/trends/place`

[4]`https://news.google.com/news`

[5]`https://dev.twitter.com/streaming/reference/get/statuses/sample`

**Table 1: Details of Twitter trends used in our data.**

| # | Trend |
|---|---|
| 0 | Magnitude 8.2 earthquakes strikes off Chilean coast. |
| 1 | Tomioka Silk Mill to become a World Heritage Site. |
| 2 | The magazine "Koakuma Ageha" ceases publication. |
| 3 | Main actor chosen for "Attack on Titan" live-action movie. |
| 4 | Sinking of the MV Sewol. |
| 5 | Club NOON cleared of violating anti-dancing law. |
| 6 | Japan to bend overtime rules for white-collar workers. |
| 7 | Dr. Obokata says she created STAP cells "over 200 times." |
| 8 | The 2nd Escort Ship's Curry Grand Prix in Yokosuka. |
| 9 | President Obama dines at Sukiyabashi Jiro in Tokyo. |

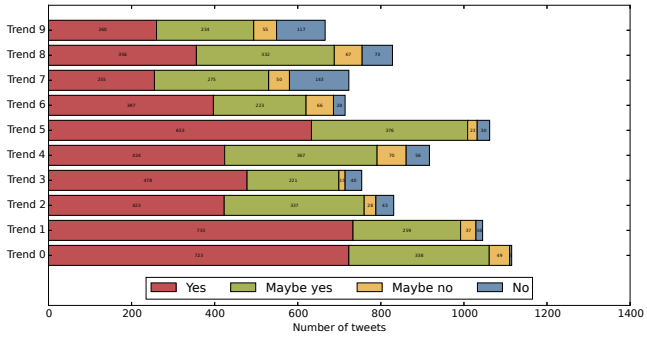


**Figure 4: Answer results for Q3 for each trend.**

Q1, we only asked whether the tweet contained opinions or impressions because by our definition credibility cannot be evaluated for subjective expressions. For tweets containing objective expressions they answered the next three questions, otherwise they quitted to answer. We asked about URLs in the tweet in Q2, credibility of the tweet in Q3, and the reasons why the annotator thought the tweet was or was not credible in Q4-1/Q4-2. In the end we got up to 14,000 labeled tweets for each question.

## 3.2 Answer Results for Q1 and Analysis

Question Q1 was *"Does this tweet contain opinions or impressions?"* The purpose of Q1 was to omit subjective tweets whose text contained only opinions or impressions, because by our definition credibility cannot be evaluated for subjective expressions. In Figure 1, we can see that 6.36% of the tweets were judged to be bots or spams, and 31.81% contained only self-opinions. Therefore, we were unable to evaluate the credibility of up to 38.17% of the trendy tweets.

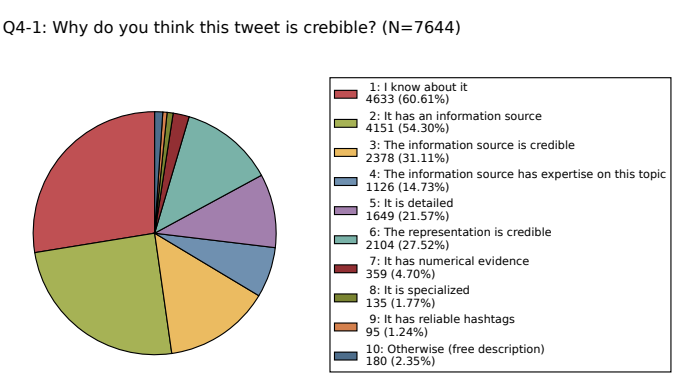


**Figure 5: Answer results for Q4-1.**

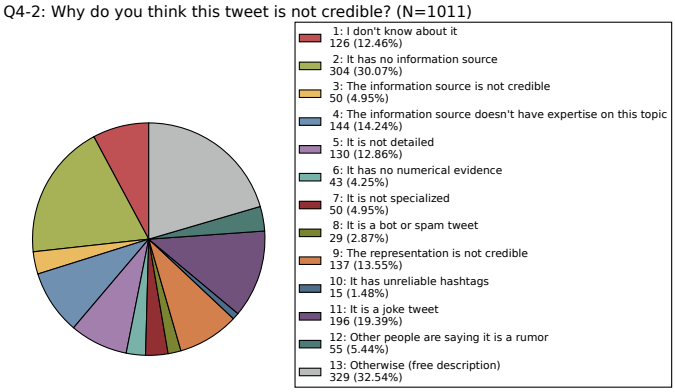


**Figure 6: Answer results for Q4-2.**

A total 8,655 tweets relevant to answer number one and two became targets for Q2.

## 3.3 Answer Results for Q2 and Analysis

Question Q2 was *"Is this tweet associated with an external link?"* The purpose of Q2 was to check whether URLs had relevance to the tweet, because spam tweets (especially fishing tweets) often appear in trendy tweets. The answer results (Figure 2) showed that most of the contents of tweets with URLs were related to the external links to which the URLs referred. Only 2.22% of the tweets were irrelevant ones, such as spams. The appearance of a few expired links made annotators select *"I can't decide"* as their answer; these links appeared because the annotation tasks were carried out four months after the tweets had been collected, and some linked webpages were deleted in the interim.

## 3.4 Answer Results for Q3 and Analysis

Question Q3 was *"Is this tweet credible?"* In Figure 3, we can see that most of the tweets (88.32%) were judged to be credible or relatively credible. There were fewer non-credible tweets than we had expected because subjective tweets were eliminated in Q1, and objective tweets tend to be mass media information. We checked the answer distribution for each trend in Q3. Figure 4 shows that the distribution differed from trend to trend and that relatively serious news topics such as trends 0, 1, and 5 tended to have more credible tweets. Conversely, innocuous topics such as trends 7, 8, and 9 tended to have more non-credible tweets, since innocuous topics get more joke tweets than serious topics.

## 3.5 Answer Results for Q4-1 and Analysis

Question Q4-1 was *"Why do you think this tweet is credible?"* Only annotators who answered Q3 with *"Yes"* or *"Maybe yes"* answered this question. We prepared nine answer choices to lessen the annotators' "thinking load", and added an *"Otherwise"* choice to allow them to answer the question freely. The annotators were required to select at least one of the 10 choices. In Figure 5, we can see that most people referred to their basic knowledge or the presence of an information source when they decided a tweet was credible. Furthermore, from the *"Otherwise"* answers we found that some annotators considered the reliability of the tweet's writer as a reason, for instance, if the writer was a journalist or a person who was right there when the incident in question happened, then the tweet seemed more credible.

## 3.6 Answer Results for Q4-2 and Analysis

Question Q4-2 was *"Why do you think this tweet is* ***not*** *credible?"* Only annotators who answered Q3 with *"No"* or *"Maybe no"* answered this question. We prepared 12 answer choices, and added an *"Otherwise"* choice to allow the annotators to answer the question freely. The annotators were required to select at least one of the 13 choices. In Figure 6, we can see that the presence of an information source was again an important factor in judging tweet credibility, but the annotators seemed to rely less on their basic knowledge. Interestingly, a key factor was whether the tweet seemed a joke. There were more *"Otherwise"* answers than Q4-1, and most annotators pointed out that a tweet from an unfamiliar writer did not seem to be credible.

## 3.7 Analysis Summary and Feature Design

The results obtained for Q4-1 and Q4-2 made it clear that the presence of an information source is the most important factor in a person's deciding that information has credibility, and the *"Otherwise"* answers told us the writer's reliability is also important. Furthermore, the level of tweet credibility may differ from topic to topic. The tweets written about serious topics such as earthquakes are more likely to be credible than tweets written about frivolous or innocuous topics such as gossip items.

We designed features for our classifier, which evaluates the information credibility of a tweet, on the basis of our analysis results and existing research work [3, 6]. We first defined the baseline features shown in Table 2. These features have been reported as being effective in assessing information credibility, and they cover most of the question choices in Q4-1 and Q4-2. However, they do not take into account two things we found, i.e., that the type of trendy topic and the writer's reliability are significant.

# 4. PROPOSED METHODS

We propose new methods to automatically assess tweet credibility by using two features, "tweet topic" and "user topic", in Sec. 4.1. We also present additional features based on a user's "expertness" and "bias" that are expected to enhance assessment accuracy in Sec. 4.2.

## 4.1 Assessment with Tweet and User Topics

The LDA model [1] is a well-known generative model for clustering words into topics and documents into mixtures of

**Table 2: Features used in the baseline.**

| Feature | Description |
|---|---|
| LENGTH_CHARS | Length of the tweet in characters. |
| LENGTH_WORDS | ... in number of words. |
| CONTAINS_? | Whether the tweet contains '?'. |
| CONTAINS_! | ... '!'. |
| CONTAINS_MULTI_?! | ... multiple '?' or '!'. |
| NUMBER_OF_URLS | Number of URLs in the tweet. |
| CONTAINS_URL | Whether the tweet contains a URL. |
| CONTAINS_MEDIA | ... a media URL. |
| CONTAINS_# | ... a hashtag. |
| CONTAINS_$ | ... a symbol. |
| CONTAINS_@ | ... a mention. |
| IS_RETWEET | Whether the tweet is a retweet. |
| REGISTRATION_AGE | Date the user is registered. |
| STATUSES_COUNT | Total number of tweets. |
| FOLLOWERS_COUNT | Number of followers. |
| FRIENDS_COUNT | ... friends. |
| LISTED_COUNT | ... lists. |
| IS_VERIFIED | Is the user verified. |
| LENGTH_BIO | Length of bio. |
| HAS_PROFILE_URL | Is URL contained in bio. |
| HAS_LOCATION | Is location contained in bio. |
| DEFAULT_PROFILE | Is bio default. |
| DEFAULT_PROF_IMG | Is the image in bio default. |
| USE_BG_IMG | Is background image used. |
| CONTRIB_ENABLED | Whether contributors can be used. |
| GEO_ENABLED | Whether geo can be used. |

topics. We collected past tweets users had written before April 2014 by using Twitter's statuses/user_timeline API[6] and used the concatenation of the tweets as a document in LDA. Because one document corresponds to one user, the topic of the document equals the topic of the user. We define "tweet topic" $P_t$ and "user topic" $P_u$ by utilizing the document-topic probability $\theta_{dt}$ and the topic-word probability $\phi_{tw}$ generated from LDA:

$$P_t(W) = \frac{\sum_{w \in V,W} \phi_{tw}}{\sum_t \sum_{w \in V,W} \phi_{tw}}, \quad (1)$$

$$P_u(d_u) = \theta_{d_u t}. \quad (2)$$

A word list $W$ (which is not a set) in a target tweet for evaluating credibility is used to calculate $P_t$ (Eq. 1). A word $w$ should appear both in $W$ and the word set $V$ used in LDA. $P_t$ is normalized by dividing it by the summation of each topic probability. $P_u$ equals $\theta_{dt}$, and we can get a user topic probability by referring to the row at the user's document index in the probability matrix of $\theta_{dt}$ (Eq. 2). Note that only nouns with appearance frequency over ten are used as $V$ to enhance the clustering accuracy of LDA.

We add "tweet topic" and "user topic" to the baseline features shown in Table 2 and use a machine learning method to train a classifier. On the basis of previous research work and our preliminary experiments, we choose Random Forests [2] as our classifier.

## 4.2 Additional Features: Expertness and Bias

We propose two additional features, which we refer to as "expertness" and "bias". They are based on the two hypotheses below.

**Hypothesis 1 (expertness):** *If a Twitter user often writes tweets about some specified topics, the user must know much about those topics, and the tweets the user has written about those topics should have relatively higher credibility.*

**Hypothesis 2 (bias):** *If the topic distribution of a Twitter user diverges much from the average topic distribution of all the users, he/she might be a bot or a very biased user, and the tweets written by the user should have lower credibility.*

For comparison with "user topic", "expertness" uses the "tweet topic" probability distribution, and "bias" uses the averaged "user topic" probability distribution of all users. Now we let $P$ be "user topic" probability distribution and $Q$ be the probability distribution of "tweet topic" or averaged "user topic". The size of both probability distributions $P$ and $Q$ is $K$. We calculate the distance between $P$ and $Q$ by using four types of equations as follows.

**Jensen-Shannon Divergence (JSD) [9]:** This is the information divergence between two probability distributions.

$$\mathrm{JSD}(P||Q) = \frac{1}{2}\mathrm{KLD}(P||M) + \frac{1}{2}\mathrm{KLD}(Q||M), \quad (3)$$

$$M = \frac{1}{2}(P+Q),\ \mathrm{KLD}(A||B) = \sum_i A(i) \ln \frac{A(i)}{B(i)}.$$

**TOP1:** This is a binary value whether or not the indices of maximum probability in the two probability distributions are the same.

$$\mathrm{TOP1}(P,Q) = \begin{cases} 1 & (\text{if } \operatorname{argmax} P == \operatorname{argmax} Q) \\ 0 & (\text{otherwise}) \end{cases}. \quad (4)$$

**Root Mean Squared Error (RMSE):** This is the square root of the mean of the square of all of the error.

$$\mathrm{RMSE}(P,Q) = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(P_i - Q_i)^2}. \quad (5)$$

**Squared Error (SE):** This is the square of all of the error.

$$\mathrm{SE}(P,Q) = \sum_{i=1}^{K}(P_i - Q_i)^2. \quad (6)$$

Equations 3, 4, and 5 return a binary value and Eq. 6 returns a vector with size $K$, which equals the size of topics in LDA. These values are added as new features to the existing features proposed in Sec. 4.1.

# 5. EXPERIMENTS

**Data:** We used the same labeled 2,000 tweets reported in Sec. 3.1. The tweets labeled "Yes" or "Maybe yes" by at least four of seven annotators were defined as positive class (credible), otherwise negative class. The reason we did not use only tweets labeled "No" or "Maybe no" as negative class is that these tweets are rare (see Figure 4); using them would make the data imbalanced. The details of our data are shown in Table 3. The past tweets for applying the LDA are the same as those described in Sec. 4.1.

**Tools:** We employed GibbsLDA++[7] for generating topics and the *RandomForestClassifier* in the scikit-learn[8] package for building the classifier. We set *n_estimators* to be 100 in *RandomForestClassifier*, otherwise we used default parameters. For segmenting a tweet into words, we used MeCab[9] with the IPA dictionary and our customized dictionary.

[6] `https://dev.twitter.com/rest/reference/get/statuses/user_timeline`

[7] `http://gibbslda.sourceforge.net`

[8] `http://scikit-learn.org`

[9] `http://code.google.com/p/mecab`

**Table 3: Number of positive and negative tweets in each trend.**

| No. | Positive | Negative | No. | Positive | Negative |
|---|---|---|---|---|---|
| 0 | 155 | 45 | 5 | 150 | 50 |
| 1 | 151 | 49 | 6 | 99 | 101 |
| 2 | 117 | 83 | 7 | 82 | 118 |
| 3 | 102 | 98 | 8 | 116 | 84 |
| 4 | 124 | 76 | 9 | 87 | 113 |

**Table 4: Performance of four sets of features. Bolded score means over the baseline and the * and ** are significance level of 5% and 1%, respectively.**

| $K$ | baseline | w/ tweet | w/ user | w/ tweet&user |
|---|---|---|---|---|
| 2 | 0.7843 | **0.7873** | **0.7905** | **0.7860** |
| 4 | 0.7843 | 0.7798 | **0.7927** | **0.7917** |
| 8 | 0.7843 | **0.8006**$^{*}$ | **0.7931** | **0.8035**$^{**}$ |
| 16 | 0.7843 | **0.7919** | 0.7825 | **0.7987** |
| 32 | 0.7843 | **0.7919** | 0.7824 | **0.8044**$^{*}$ |
| 64 | 0.7843 | 0.7820 | 0.7768 | **0.7967** |
| 128 | 0.7843 | 0.7734 | 0.7786 | **0.7912** |

**Evaluation:** We based the experiments on 10-fold cross validation and measured the Area Under Curve (AUC) for whole prediction outputs. The AUC equals the area under the Receiver Operating Characteristic (ROC) curve and takes a value from 0 to 1, with 1 being best and 0 being worst. The closer the plot of the ROC curve approaches the upper left corner, the better. We also used the Delong test [4], which is a nonparametric approach to test the significant difference between two ROC curves. We evaluated the difference from the baseline.

## 5.1 Effectiveness of Tweet and User Topics

We evaluated four different sets of features: the baseline, the w/ tweet ("tweet topic"), the w/ user ("user topic"), and the w/ tweet&user (both of the two topics). We varied the number of topics $K$ from 2 to 128 in a geometric sequence.

Figure 7 and Table 4 shows that the w/ tweet&user always gave the best performance. Compared with the baseline, it works the best when $K$ is 32 at the significance level of 5% (p-value of 0.011). Furthermore, the w/ tweet and the w/ user also outperform the baseline for some $K$ values, especially for 8. This value may be suitable for clustering our data since it is neither too big nor too small.

We conclude that both "tweet topic" and "user topic" are useful to evaluate the credibility of a tweet, when the topics are clustered by appropriate size. Additionally, the performance increases when the both topics are used at the same time. In the following subsections, we considerate the reason why these topics work well.

### 5.1.1 Why w/ tweet works

By checking the true positive (TP) number and the true negative (TN) number between the baseline and the w/ tweet, we found that TP increased from 887 to 919 but TN decreased from 564 to 561. This increase of TP overcomes the decrease of TN, therefore the w/ tweet outperforms the baseline. Especially, when we focused on trend number 0, 1, and 5, we found that their TP increased 26, 13, 8 respectively, which means that our classifier learned that the tweets of these three trends have higher credibility. In fact, these trends have many positive tweets than the other trends (see Figure 4 and Table 3). The "tweet topic" works because that the possibility of a tweet to be credible varies in different trends, e.g. earthquakes or gossips.

### 5.1.2 Why w/ user works

By checking the TP and TN between the baseline and the w/ user, we found that TP changed from 887 to 882, and TN changed from 564 to 562 after adding the "user topic". In spite of the decrease of the numbers in TP and TN, why did the AUC score increase 0.0088 points at $K$ is 8?

To reveal the reason, we plotted the ROC curve in Figure 8. We can see that when the false positive rate is between 0.0 and 0.2, the w/ user is closer to the upper left area than the baseline, which means that the w/ user works better around the range. Hence, we checked the location where the w/ user intersects with the baseline. Sequencing the probability of a tweet to be credible in descending order, we find that the cross point lies in the $849_{th}$ area from the top. Before the $849_{th}$ area, there are more TPs in the w/ user than in the baseline, which makes the AUC score of the w/ user higher. Specifically, the value that the classifier of the w/ user outputs is likely to be TP when the classifier assesses with high confidence.

The next question is what kinds of tweets made TP increase by adding the "user topic". We found that the tweets in trend 5 made TP increase in the top 849 areas. Figure 9 shows the relationship between the trend topic and the user topic. The lightness of every lattice means the possibility that a user topic is related to a trend. A brighter lattice indicates a higher probability. For example, the lattice at [topic 3, trend 8] is bright, because the users who like games such as KanColle[10] got passionate about trend 8. We know that tweets in trend numbers 0, 1, and 5 are more likely to be credible according to our data analysis (see Sec. 3.4), and these trends have the highest probability in topic 6. Therefore, after adding the "user topic", the classifier learned that if a user has high probability in topic 6, his/her tweet is more likely to be credible. Here, topic 6 was a daily life topic that included words such as "work", "photos", and "today".

### 5.1.3 Why w/ tweet&user works

By checking the TP and TN values between the baseline and the w/ tweet&user, we found that adding the "tweet topic" and "user topic" simultaneously helped TP to increase from 887 to 892 and TN to increase from 564 to 579. Since both TP and TN increased, the AUC score got larger. We believe the reason that adding both the "tweet topic" and "user topic" features works is because doing so gives the classifier more information about the relationship between the tweet and the user, which helps the classifier make better decisions.

## 5.2 Effectiveness of Expertness and Bias

We evaluated two additional features we call the user's "expertness" and "bias" by adding them to the w/ tweet&user each in turn. We tried this for four methods (JSD, TOP1, RMSE, and SE) while changing $K$ in the same way as related in Sec. 5.1. Out of the 28 combinations (four methods and seven $K$s), the "bias" worked better than the "expertness" 20 times (see Table 5 and Table 6). This indicates that the second hypothesis given in Sec. 4.2 might be more convincing than the first one. The best method, which showed

[10] `http://www.dmm.com/netgame/feature/kancolle.html`

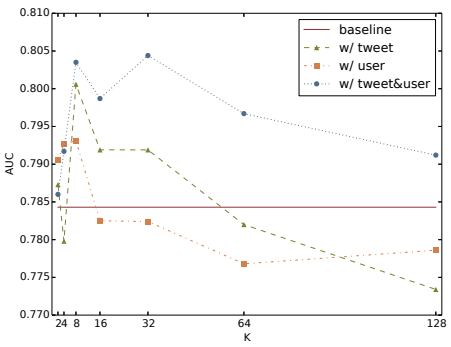


**Figure 7: Performance of four sets of features.**

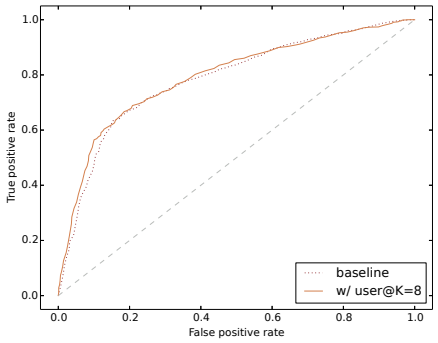


**Figure 8: ROC curves comparing the baseline with the w/ user.**

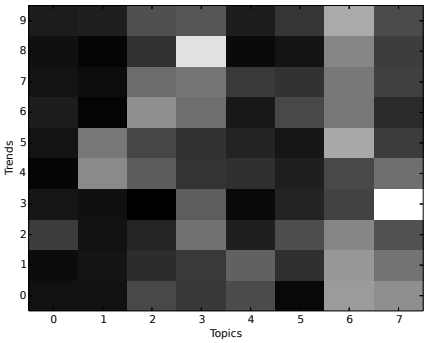


**Figure 9: Probability lattice diagram of user topics and trends.**

**Table 5: Performance of the "expertness". Bolded score means over the "bias".**

| K | JSD | TOP1 | RMSE | SE |
|---|---|---|---|---|
| 2 | **0.7873** | 0.7867 | 0.7865 | **0.7880** |
| 4 | **0.7893** | **0.7889** | **0.7886** | 0.7805 |
| 8 | 0.8033 | 0.7987 | 0.8003 | 0.7986 |
| 16 | 0.8037 | **0.8000** | 0.8010 | 0.7899 |
| 32 | 0.8010 | 0.8038 | 0.8003 | 0.8033 |
| 64 | **0.7979** | **0.7986** | 0.7961 | 0.7947 |
| 128 | 0.7929 | 0.7957 | 0.7946 | 0.7874 |

**Table 6: Performance of the "bias". Bolded score means over the "expertness" and the * and ** are significance level of 5% and 1%, respectively.**

| K | JSD | TOP1 | RMSE | SE |
|---|---|---|---|---|
| 2 | 0.7840 | **0.7895** | **0.7871** | 0.7854 |
| 4 | 0.7872 | 0.7857 | 0.7886 | **0.7845** |
| 8 | **0.8063** | **0.8039*** | **0.8044** | **0.8061*** |
| 16 | **0.8045** | 0.7983 | **0.8030** | **0.7992**** |
| 32 | **0.8034** | **0.8039** | **0.8027** | **0.8086** |
| 64 | 0.7973 | 0.7966 | **0.7976** | **0.7970** |
| 128 | **0.7969*** | **0.7964** | **0.7967** | **0.7954*** |

a significant difference, was the SE with "bias" when $K$ was 8; the AUC score was 0.8061 with a 5% significance level (p-value of 0.017). This score is approximately a 3% improvement over the baseline. Among the four methods with "bias", SE appears to the best one because it showed good performances with a significant difference many more times than the others. This is because SE has the features of $K$ size, and consequently it supplied more information than the other methods.

## 6. CONCLUSION

We collected trendy tweets in Japan and analyzed how people judge whether a tweet is credible or not. In our analysis, we found that the most important factor in making this judgment is whether a tweet has an information source. Two other factors, whether the topic of a tweet is a serious one and whether the user of a tweet is reliable, also attracted people's attention.

On the basis of analysis results, we proposed new methods to assess the information credibility of a tweet, both of which utilize the "tweet topic" and "user topic" features obtained from the Latent Dirichlet Allocation (LDA) model. Experiments we conducted showed that both of them are effective when the topic size is appropriate, and the performance is enhanced by using them together. The reason these topics work is that they can recognize reliable trendy topics and users, e.g., a news tweet of an earthquake posted by a user who is not a bot.

Furthermore, we presented two additional features, the "expertness" and the "bias", derived from two hypotheses we built. Since the "bias" worked better than the "expertness" in our experiments, the hypothesis, that *biased users who diverge much from the average topic distribution of all users tend to post non-credible tweets*, might be the more convincing of the two.

## 7. REFERENCES


[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[3] C. Castillo, M. Mendoza, and B. Poblete. Information Credibility on Twitter. In *WWW*, pp. 675–684, 2011.

[4] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 44(3):837–845, 1988.

[5] B. Fogg and H. Tseng. The Elements of Computer Credibility. In *CHI*, pp. 80–87, 1999.

[6] M. Gupta, P. Zhao, and J. Han. Evaluating Event Credibility on Twitter. In *SDM*, pp. 153–164, 2012.

[7] A. Java, X. Song, T. Finin, and B. Tseng. Why We Twitter: Understanding Microblogging Usage and Communities. In *WebKDD/SNA-KDD*, pp. 56–65, 2007.

[8] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a Social Network or a News Media? In *WWW*, pp. 591–600, 2010.

[9] J. Lin. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37:145–151, 1991.

[10] M. Mendoza, B. Poblete, and C. Castillo. Twitter Under Crisis: Can we trust what we RT? In *SOMA*, pp. 71–79, 2010.

[11] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz. Tweeting is Believing?: Understanding Microblog Credibility Perceptions. In *CSCW*, pp. 441–450, 2012.

[12] J. O'Donovan, B. Kang, G. Meyer, T. Höllerer, and S. Adalı. Credibility in Context: An Analysis of Feature Distributions in Twitter. In *SocialCom/PASSAT*, pp. 293–301, 2012.

[13] R. Thomson, N. Ito, H. Suda, F. Lin, Y. Liu, R. Hayasaka, R. Isochi, and Z. Wang. Trusting Tweets: The Fukushima Disaster and Information Source Credibility on Twitter. In *ISCRAM*, pp. 1–10, 2012.

[14] X. Xia, X. Yang, C. Wu, S. Li, and L. Bao. Information Credibility on Twitter in Emergency Situation. In *PAISI*, pp. 45–59, 2012.