Assessment of Tweet Credibility with LDA Features

Jun Ito, Jing Song, Hiroyuki Toda, Yoshimasa Koike, and Satoshi Oyama

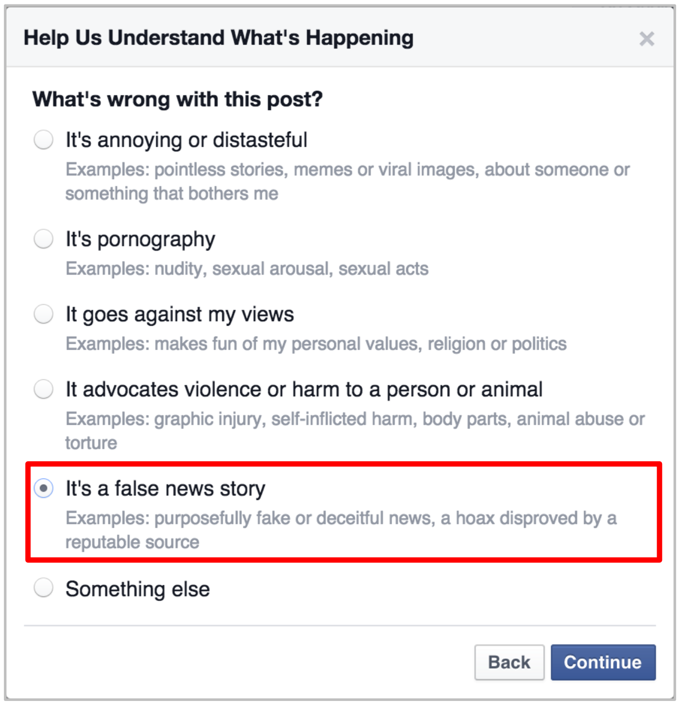
NTT Service Evolution Laboratories, NTT Corporation
Graduate School of Information Science and Technology, Hokkaido Univ.

#RDSM, May 19th, 2015

*) This study has done in one month internship period of the second author (Jing Song).

Background: information credibility is a big issue

Facebook enables their users to report a "false news story."

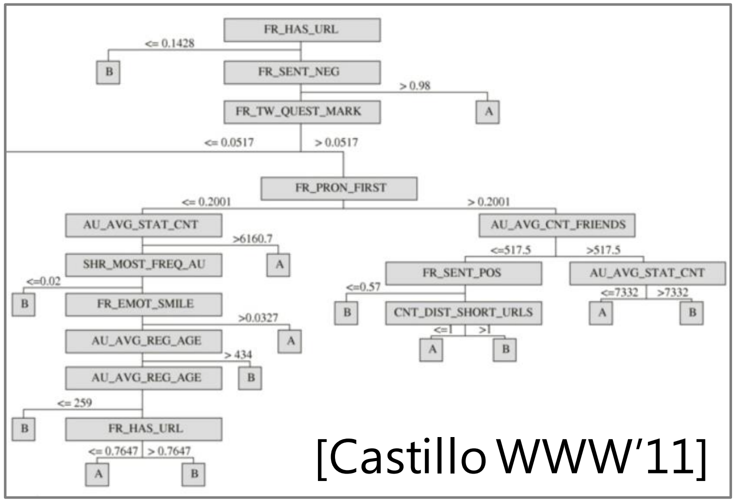


News Feed FYI: Showing Fewer Hoaxes
http://newsroom.fb.com/news/2015/01/news-feed-fyi-showing-fewer-hoaxes/

Twitter does not have a hoax reporting function. But there is a PHEME project.

PHEME: Computing Veracity
http://www.pheme.eu/

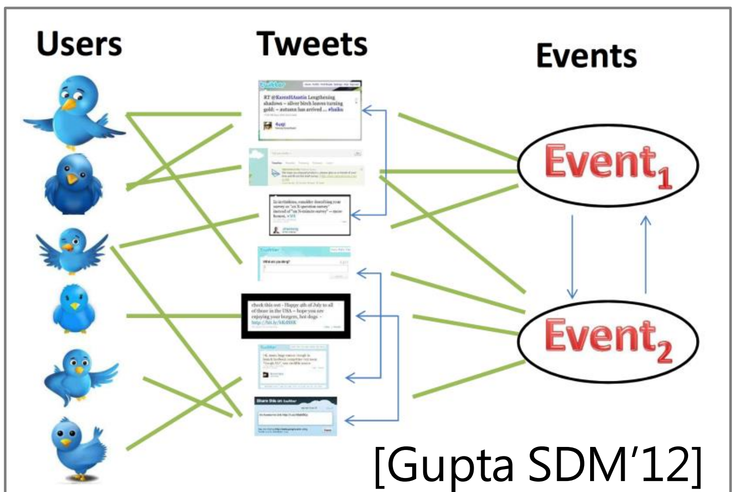
Related Work



[Castillo WWW'11]

Information Credibility on Twitter

- Credibility of trends (news events)
- J48 classifier
- Features from text, user, trend, and propagation



[Gupta SDM'12]

Evaluating Event Credibility on Twitter

- Credibility of events
- J48 or KNN
- Features are almost the same as Castillo
- Graph-based optimization after J48/KNN
- Tweets written about the same event have similar credibility score

Our Focus and Contributions

Our Focus

- Tweet credibility of a **trendy news post**.
- Credibility of **every tweet** instead of every trend.
- Considering the **user topic distributions**.

Contributions

- We show basic analysis results that how people judge the credibility of a tweet from the 2,000 trendy tweets in Japan posted on April, 2014.
- We propose the methods to infer information credibility of a tweet by using two new features, the **"tweet topic"** and the **"user topic"**, derived from the LDA (Latent Dirichlet Allocation) model.
- We build two hypotheses based on a user's **"expertness"** and **"bias"** and design four methods to extract additional features.

Data Collection

Data Collection

Innovative R&D by NTT

①

trends/place

Access API every 5 min to get trendy words

②

Google news

Check whether the trendy words also exists in the Google News title

③

Pick up 10 trends

④

streaming sample

Collect 200 tweets with trendy words in each trend

⑤

Annotate tweet's credibility

NTT

0. Earthquake in Chile

1. Tomioka Silk Mill

2. Koakuma Ageha

3. Attack on Titan

4. Sinking of the MV Sewol

Ten trends in our data set

5. Club NOON

6. White collar exemption

7. STAP cells

8. Escort Ship's Curry Grand Prix

9. Sukiyabashi Jiro

How to annotate credibility

Annotator

14 annotators who were widely distributed by age and sex and who were all used to Twitter.

Data

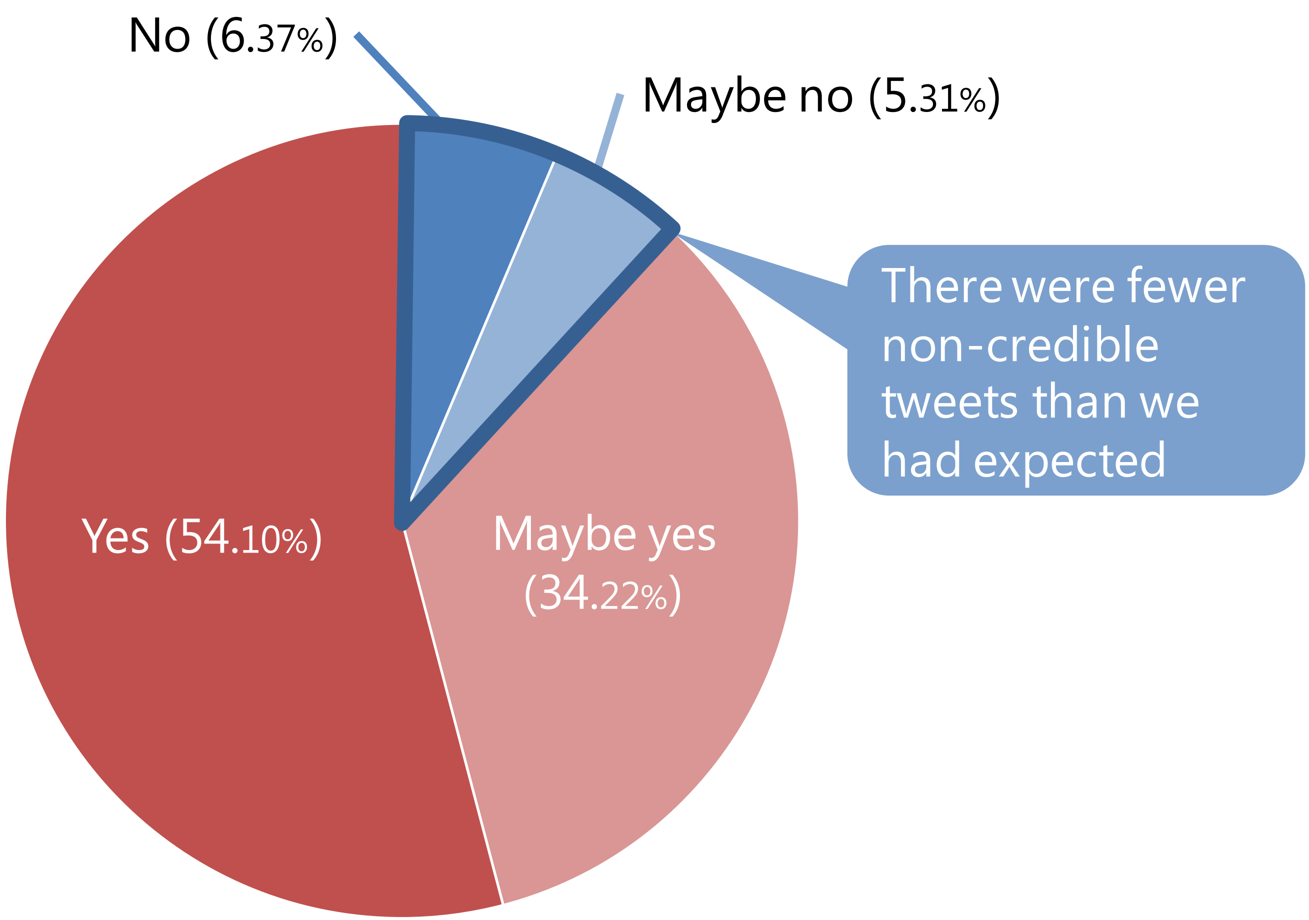
100 tweets w/ URLs and 100 tweets w/o URLs for each trend in ten trends (**2,000 tweets** in total).

Method

- **Seven randomly assigned annotators** to answer 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).

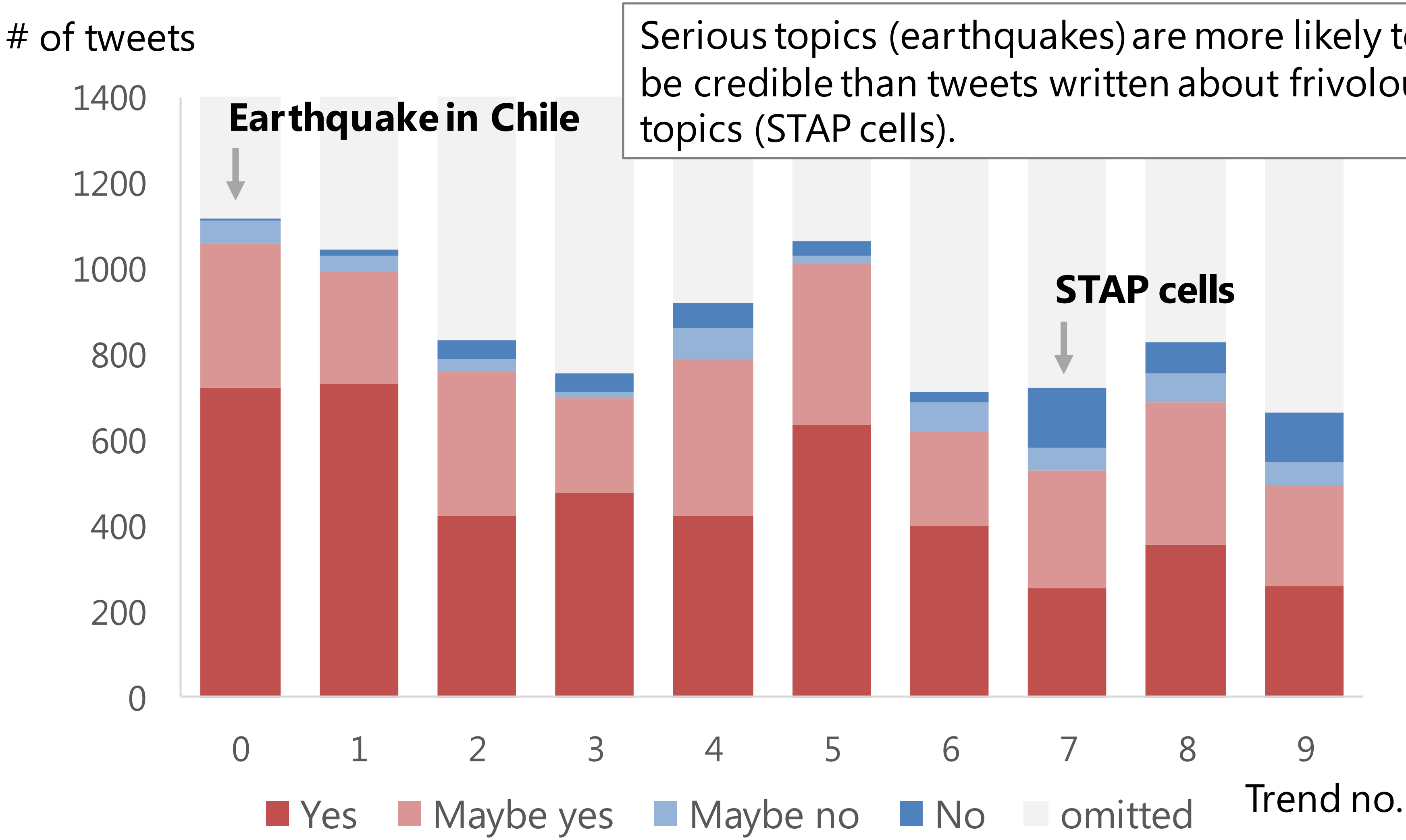
Answer Results and Analysis

Is this tweet credible?



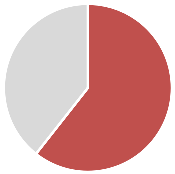
Credibility for each trend

Serious topics (earthquakes) are more likely to be credible than tweets written about frivolous topics (STAP cells).



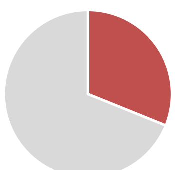
Why do you think this tweet is credible?

Top 3 reasons to think this tweet as credible



I know about it (60.61%)

It has an information source (54.30%)



The information source is credible (31.11%)

- The **presence of an information source** is important.
- The **reliability of the tweet's writer** is also important.
    - Popular news media, a person who was right there when the incident happened, etc.

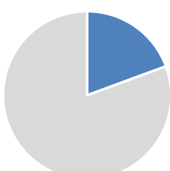
Why do you think this tweet is **not** credible?

Top 3 reasons to think this tweet as non-credible

Otherwise (free description) (32.54%)

- Most annotators pointed out that a tweet from an unfamiliar writer did not seem to be credible.

It has no information source (30.07%)



It is a joke tweet (19.39%)

- The **presence of an information source** is important.
- The **reliability of the tweet's writer** is also important.
- Interestingly, 3rd factor was whether the tweet seemed a **joke**.

Analysis Summary

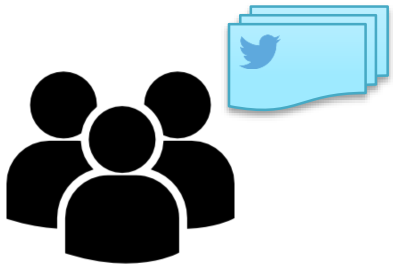
- The **presence of an information source** is the most important factor in a person's deciding that information has credibility.
- The **writer's reliability** is also important.
- The level of tweet credibility may **differ from topic to topic**.

Our Methods

Basic Features

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.

Tweet and User Topics



Past tweets of user *u* are concatenated as a doc *d*.

topic-word distribution ϕ_{tw}

doc-topic distribution θ_{dt}

tweet topic

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

user topic

$$P_u(d_u) = \theta_{d_u t}$$

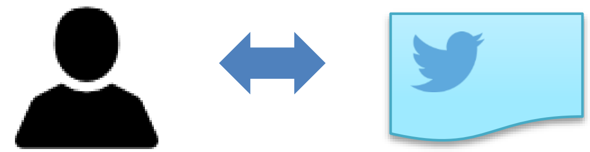
Given a target tweet *x*, composed of a word set *W* and posted by user *u*, we create a feature vector *v* as

$$v_x = \text{BasicFeatures}(x) + \underline{P_t(W) + P_u(d_u)}$$

Expertness and Bias

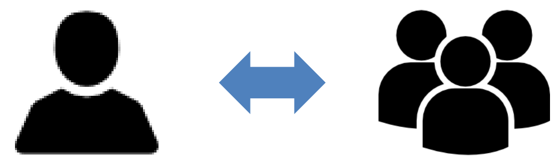
For further inspection of "user topic", we hypothesized

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.

Expertness and Bias

We tried **four methods** to calculate the **distance** (⟷) of two given distributions.
The **distance is added as new features** to the existing features.

Expertness **Bias**

P *Q* *P* *Q′*

Jensen-Shannon Divergence (JSD)

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

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

TOP1

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

Root Mean Squared Error (RMSE)

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

Squared Error (SE)

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

Experiments and Results

Experiments

Exp. 1. Effectiveness of Tweet and User Topics
Exp. 2. Effectiveness of Expertness and Bias

Data

- Labeled 2,000 tweets
  - **Class 1 (positive)**: The tweets labeled "**Yes**" or "**Maybe yes**" by at least <u>four of seven annotators</u>
    Class 0 (negative): Otherwise
- Past tweets of users in labeled tweets

Tools

- GibbsLDA++
  - Only <u>nouns</u> with appearance frequency <u>over ten</u> are used
- scikit-learn (RandomForestClassifier)
- MeCab (Japanese part-of-speech and morphological analyzer)

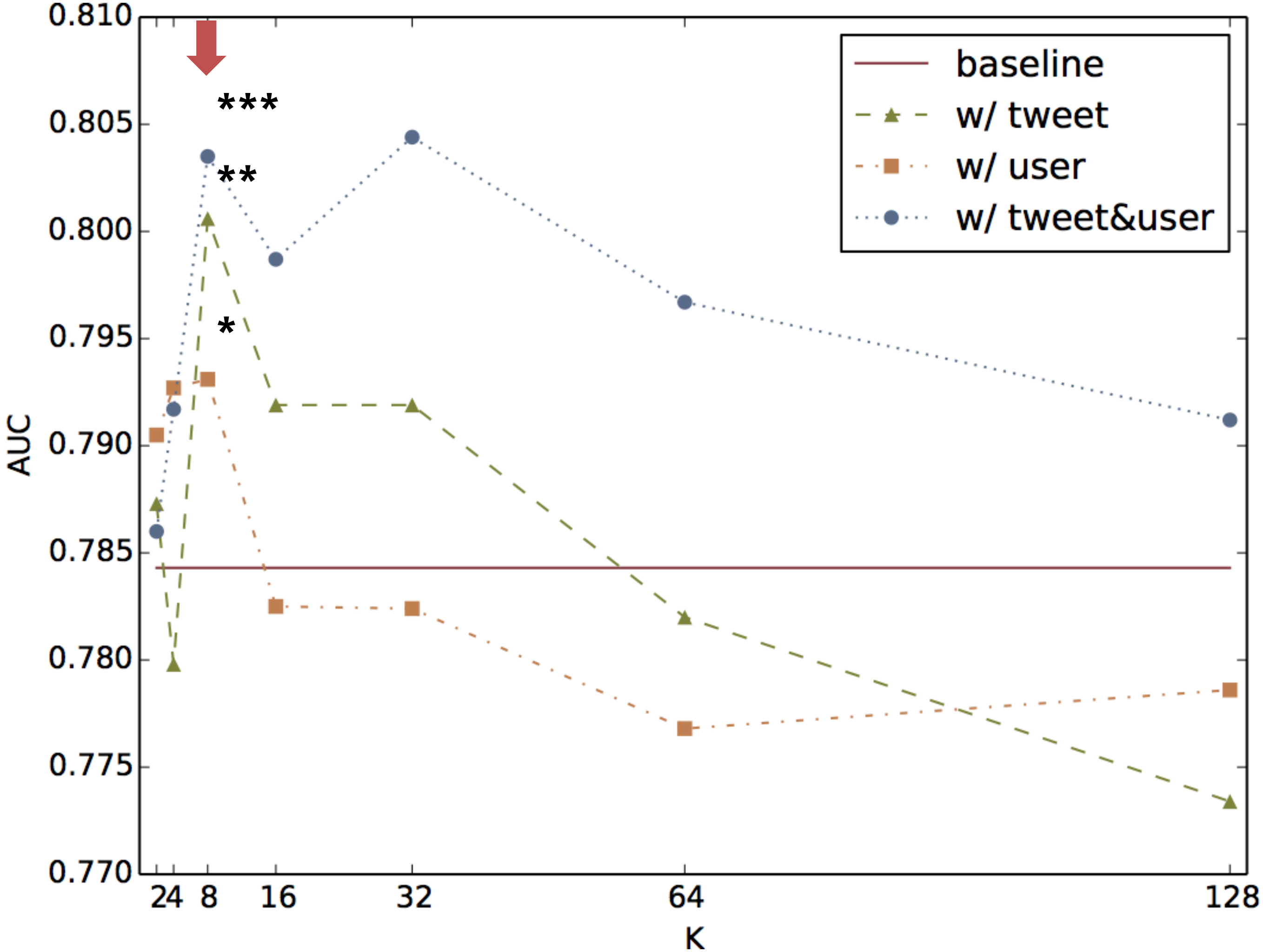
Evaluation

AUC (Area Under Curve) for whole prediction outputs of 10-fold cross validation.

NTT

Exp. 1. Effectiveness of Tweet and User Topics

Both tweet topic and user topic are useful to evaluate the credibility of a tweet, when the topics are clustered by appropriate size (K=8).

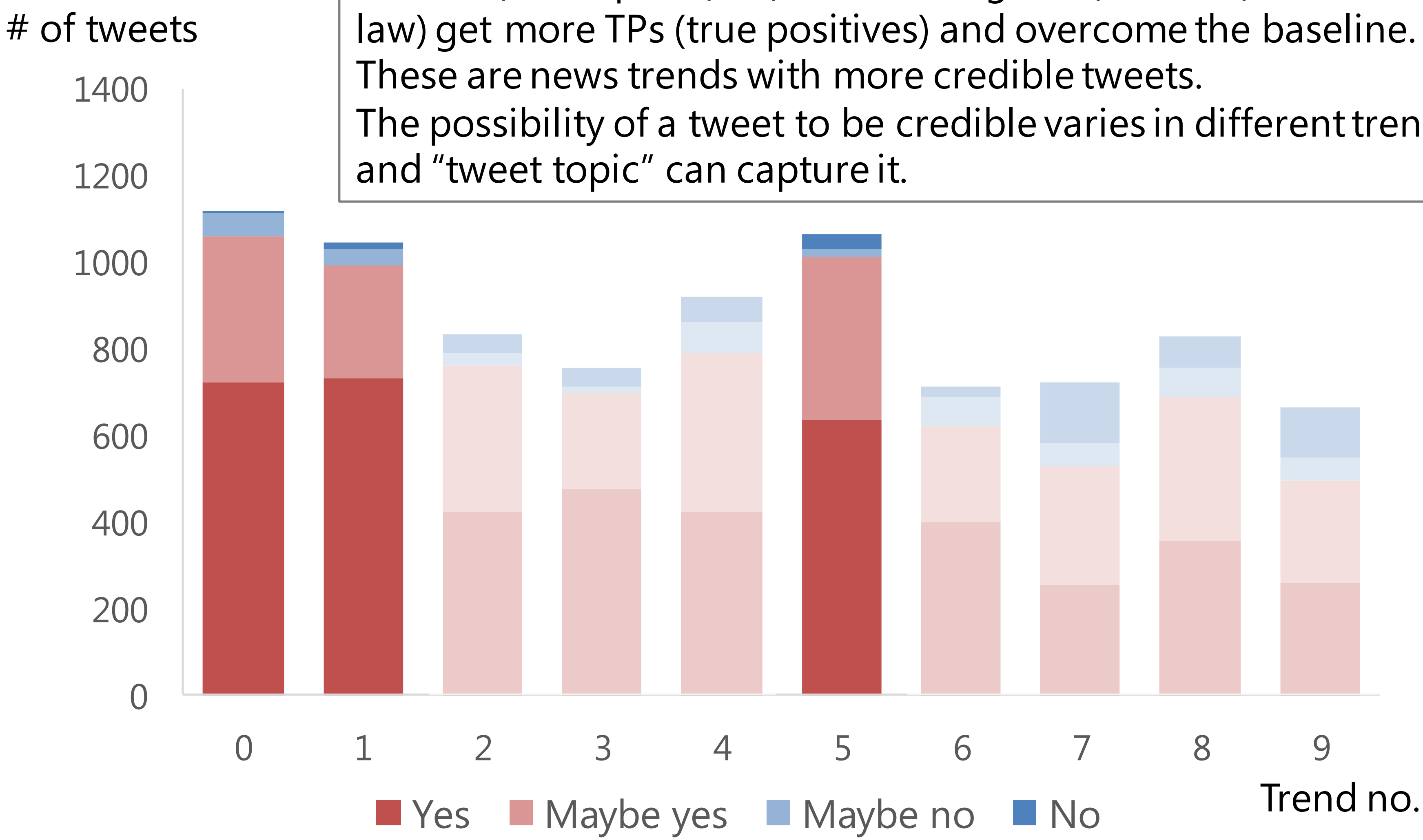


Significance level

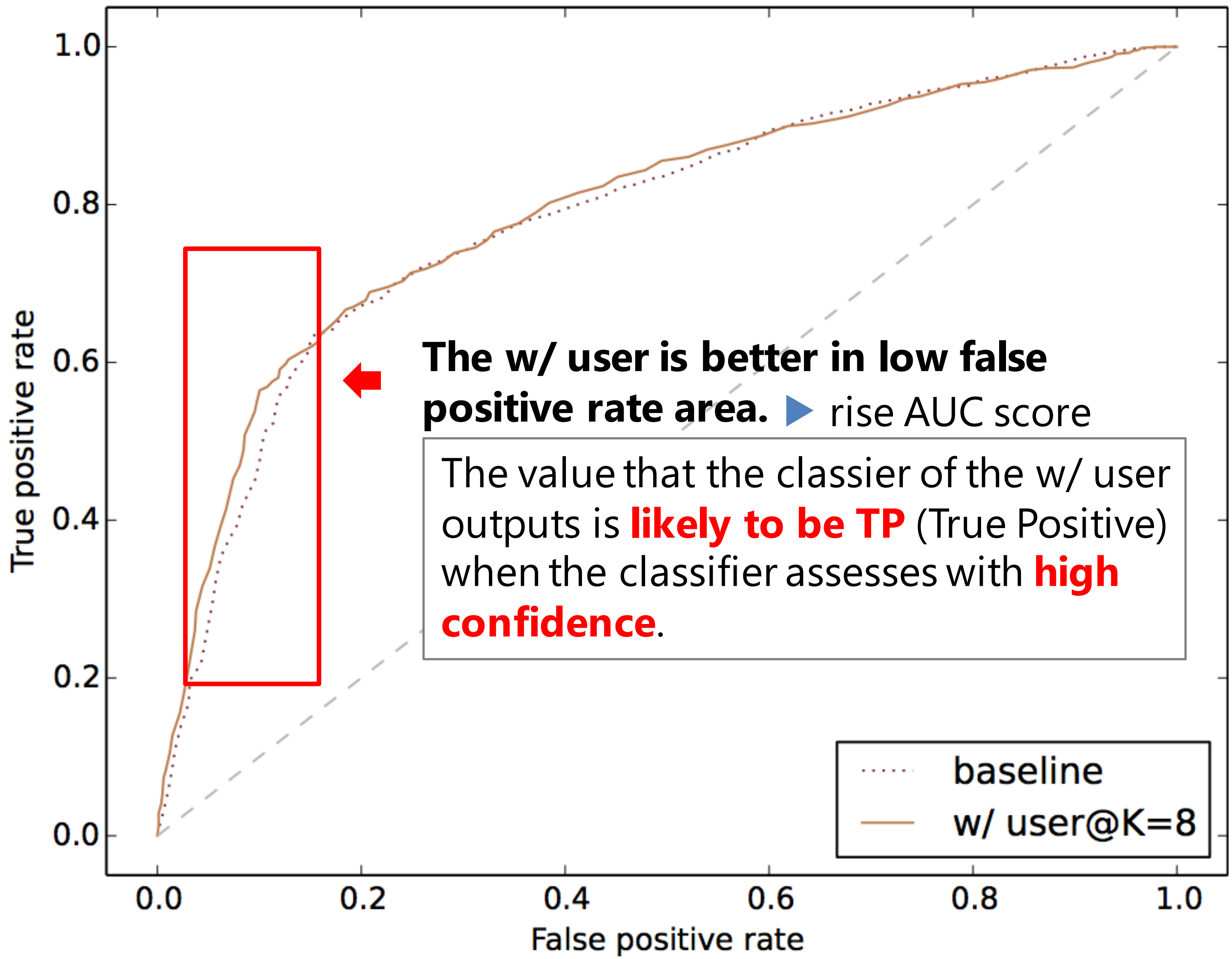
*** 1%
** 5%
* 10%

Why “tweet topic” works?

Trend 0 (earthquake), 1 (world heritage site), and 5 (anti-dancing law) get more TPs (true positives) and overcome the baseline. These are news trends with more credible tweets.
The possibility of a tweet to be credible varies in different trends, and “tweet topic” can capture it.



Why “user topic” works?



The w/ user is better in low false positive rate area. ▶ rise AUC score

The value that the classier of the w/ user outputs is **likely to be TP** (True Positive) when the classifier assesses with **high confidence**.

Exp. 2. Effectiveness of Expertness and Bias

- Out of the 28 combinations, the bias worked better than the expertness 20 times.
- SE appears to the best one because it showed good performances with a significant difference many more times than the others.

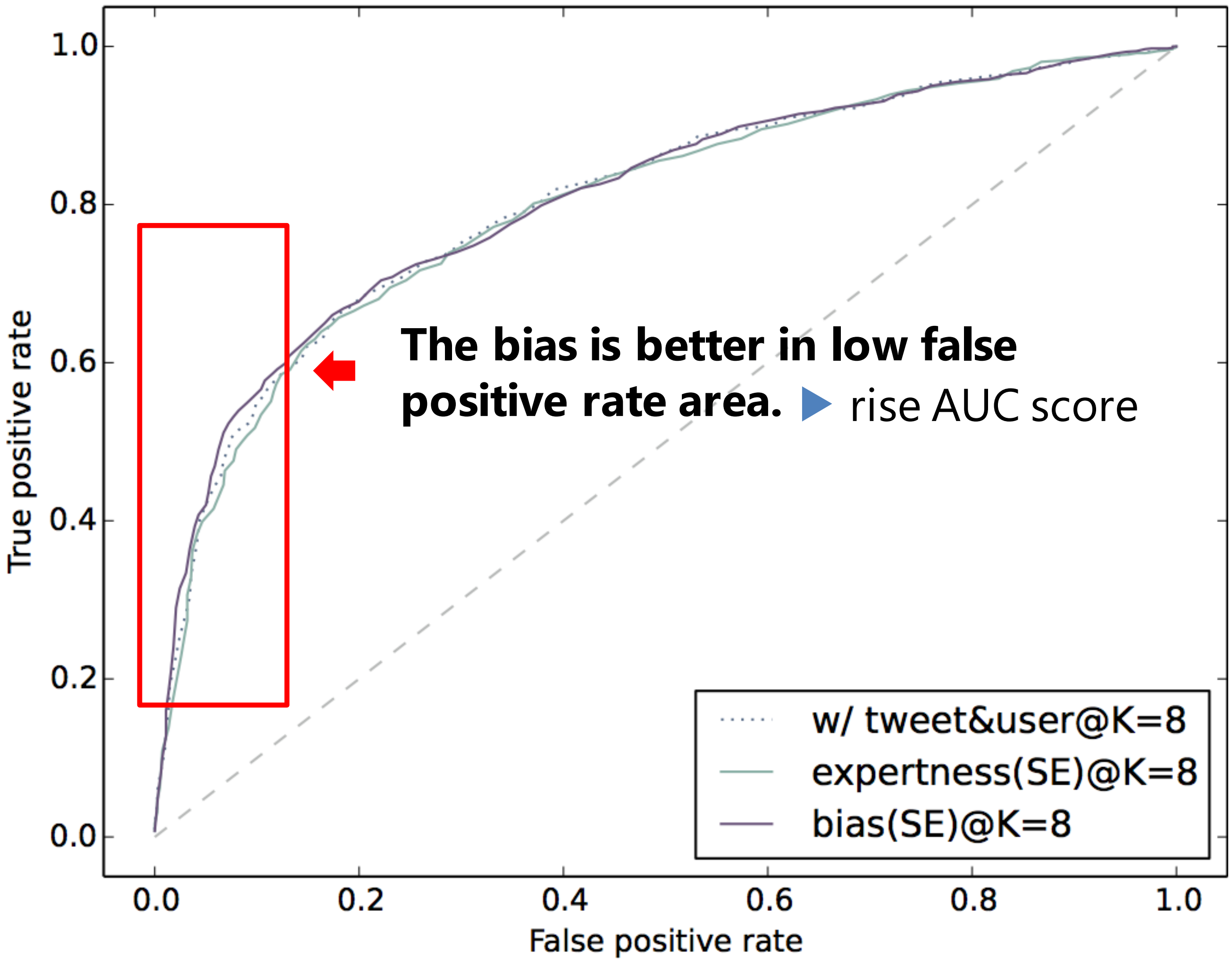
K	JSD	TOP1	RMSE	SE
2	0.7840	**0.7895**	**0.7871**	0.7854
4	0.7872	0.7857	0.7886	**0.7845**
8	**0.8063**	**0.8039****	**0.8044**	**0.8061****
16	**0.8045**	0.7983	**0.8030**	**0.7992*****
32	**0.8034**	**0.8039**	**0.8027**	**0.8086**
64	0.7973	0.7966	**0.7976**	**0.7970**
128	**0.7969****	**0.7964**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

Bold: Over the "expertness" in Tab. 6.

**, ***: Significance level of 5%, and 1%, respectively.

NTT

Why "bias" works?



The bias is better in low false positive rate area. ▶ rise AUC score

Conclusions

- "Tweet" topic works
  - The possibility of a tweet to be credible varies in different trends (e.g. earthquakes or gossips).
- "User" topic works
  - Users categorized in some topic (e.g. daily life) tend to appear in trends with more credible tweets.
- "Bias" works
  - The effect of "user" topic is enhanced by adding the "bias" features.