International Journal of Electrical and Computer Engineering (IJECE)
Vol. 10, No. 4, August 2020, pp. 3537~3549
ISSN: 2088-8708, DOI: 10.11591/ijece.v10i4.pp3537-3549

Measuring information credibility in social media using combination of user profile and message content dimensions

Erwin B. Setiawan, Dwi H. Widyantoro, Kridanto Surendro
School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Indonesia

Article Info

Article history:

Received Mar 19, 2019
Revised Feb 3, 2020
Accepted Feb 12, 2020

Keywords:

Facebook
Information credibility
Social media
Twitter

ABSTRACT

Information credibility in social media is becoming the most important part of information sharing in the society. The literatures have shown that there is no labeling information credibility based on user competencies and their posted topics. This paper increases the information credibility by adding new 17 features for Twitter and 49 features for Facebook. In the first step, we perform a labeling process based on user competencies and their posted topic to classify the users into two groups, credible and not credible users, regarding their posted topics. These approaches are evaluated over ten thousand samples of real-field data obtained from Twitter and Facebook networks using classification of Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (Logit) and J48 Algorithm (J48). With the proposed new features, the credibility of information provided in social media is increasing significantly indicated by better accuracy compared to the existing technique for all classifiers.

Corresponding Author:

Erwin B. Setiawan,
School of Electrical Engineering and Informatics,
Institut Teknologi Bandung
Ganesha Street no. 10, Bandung, Indonesia.
Email: erwinbudisetiawan@telkomuniversity.ac.id

1. INTRODUCTION

It cannot be denied that the popularity of social media has increased rapidly in recent years. Currently, about 320 million users monthly are active on the micro-blogging site, Twitter. Twitter is a global phenomenon, where 77% of Twitter accounts are outside of the United States and Twitter supports 33 languages. Because of the efficiency, volume, and timeliness of information, Online Social Networking (OSN), for example, twitter.com, has become an important source of information [1]. According to the Twitter blog, about an average of 340 million tweets are generated per day as of March 2012. In addition to receiving information from the people they "follow", people are increasingly looking for relevant topical tweets, which is more than 1.6 billion requests for Twitter search portals per day.

In particular, learning about news is often an important motivation for people to read tweets [2], for example, in order to continuously update information about local emergencies [3]. One of the OSN functions is to become a medium of sharing and searching for information [4, 5]. Each user can act as a source and spreader to the information, either forwarded in full or with modifications and additions. The role of OSN as a source of information is even more prominent in emergencies such as in particular accidents, natural disasters and incidents of terrorism because it provides a faster report than conventional media [6-14].

However, false information that spreads on social media has serious consequences. Thus, a mechanism to automatically determine the credibility of the tweet is required. Morris et al conducted a survey to understand the perceptions of user credibility on Twitter [3]. Morris et al also conducted an experiment with the purpose of uncovering user-based or content-based features used to assess the credibility. Consequently, user-based features can be grouped into three categories: influence, topical

expertise, and reputation. The influence feature includes the number of followers, retweet, and mention. While the topical expertise feature is obtained by searching through the author's homepage, the author's imaging history, outside the web page that discusses the topic the author is conveying, and the author is in a location that is relevant to the topic. The reputation-based feature helps to show the user's familiarity with the Twitter author.

This feature includes the case, either the author is followed by the user, or the author is someone that the user has heard before, or the author's account has been verified by Twitter. The content-based feature that reveals most of the credibility of tweets is if the tweet contains a reputable URL link, some tweets made the same claim as the intended tweet, it uses standard grammar, or it uses its own profile photo image or images related to the topics they are interested in and the structure of the author's username.

A study to analyze how online social media users rated the credibility of tweets has been conducted by Shariff [15]. In this study, 98 evaluators have been empowered to assess the credibility level of 400 tweets that have been used. Shariff reveals that the topic involving politics has a number of tweets with low credibility. In addition, tweets that do not have links, such as URLs, are often difficult for users to recognize. In addition, one of the earliest works that automatically predicted the credibility of the news and tweets has been conducted by Castillo [16]. This work applied two stages of data collection. First, label and save the tweets that are considered newsworthy. Second, use 7 evaluators to label newsworthy tweets with credibility values. To get this annotation, Castillo used Amazon Mechanical Turk and labeled the tweets based on new feasibility and credibility.

Furthermore, the use of SVM ratings and Pseudo-Relevance Feedback (PRF) to rank the credibility of tweets has been done by Gupta [17]. Gupta categorizes its features into two: content-based features or source-based features. The results of the study show that manual labeling has been carried out for the level of credibility related to tweets that propagate fake images of the hurricane Sandy but have not involved the competency of the source/user who spread the tweet. Some key observations about the tweet features which correlate with credibility have been created. The tweets with a large number of unique characters and contain URLs tend to be more trusted.

The latest research was conducted by Ross in 2016 with an aimed at creating and selecting a range of features that would produce a better performance when training and testing data sets originating from two different years with different topics. The data used in this study is the data used by Gupta in two different studies that have been manually labeled namely [18, 19].

Facebook has more challenges in term of information credibility compared to Twitter. Therefore, the research on the information credibility on Facebook is rarely conducted and one of the research was conducted by Saikaew in 2015. The reasons that make Facebook is more challenging because, first, the convenience in accessing Twitter content through Twitter API. Although Facebook has a Graph API with the ability to access content, the access to the information is also limited through the Graph API itself. Second, Facebook has more active users than Twitter. In September 2017, about 2,061 billion users are active in Facebook, while 328 million users are active in Twitter [20]. While Indonesia is ranked second, which is 48%, as the country with the most active social media users. Finally, compared to Twitter, Facebook has richer features, such as features that allow users to simply click and comment easily.

Several researches discussed the credibility of information on popular social networking sites, such as Twitter. However, Saikaew's research is the only research that focuses on calculating the value of information credibility on Facebook that has more users. Saikaew only uses 8 features [21], however we use 54 features to increase accurate of credibility measurement. The labeling is made manually then the rating is updated systemically by the user who can access the application. However, in Saikaew, the user's competence is still not being viewed. Furthermore, this paper applies a different approach, i.e., labeling information credibility based and introduce 17 new features for Twitter and 49 new features for Facebook. Meanwhile, for the feature dimensions, we use two feature dimensions consisting of user profile and message content dimension.

Our contributions are summarized as follows:

a. The paper introduces new 17 features for Twitter and 49 features for Facebook to increase information credibility
b. We present a labelling process to classify the users into two groups, credible and not credible user groups, depending their posted topics.

The finding in this paper is expected to help organizations and the practitioners to make better decisions, because accurate credibility is achievable due to large number of features. Furthermore, the organizations and the practitioners are informed with the updated topic due to automatically tool.

2. RESEARCH METHOD

The Proposed Information Credibility Model is shown in Figure 1. Dataset are divided into two, i.e., training data and testing data, where training data are labeled manually, and while testing data are pre-processed, including their feature extraction. The result of the feature extraction for training data come into the feature selection process and then move to the credibility classification modeling process and then the modeling result is used to predict the testing data. Finally, the Twitter credibility class with good accuracy is expected to be gained.

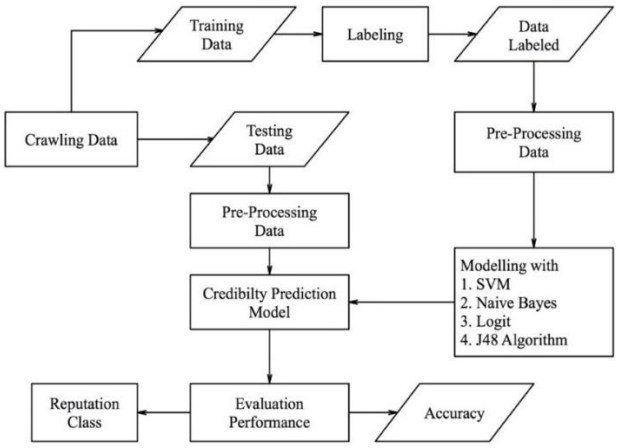

Figure 1. The proposed information credibility model

2.1. Labeling

Labeling is applied based on the compatibility of user competencies and tweet or message. In this paper, we consider of concept stating that posted tweets with a tweet topic correlated to competence of the posting account is a measure to be credible rather than posted tweets with a tweet topic uncorrelated to competence of the posting account. This concept builds a higher probability of posted tweet is credible or not. We also define tweet is message posted in twitter and message is message posted in Facebook. We perform labeling manually for tweet and message categories, while for user competencies, we perform a real survey. The objective of survey is to collect information of user competencies. We made an online survey through the website www.surveymonkey.com in January - March 2017. Respondents were asked questions about their opinion of 256 famous people with each corresponding competence. Information displayed in the survey includes photos, bio profiles, five tweets and five messages having the highest engagement, number of followers, number of tweets, and number of following. The survey has been conducted on 188 respondents, 137 men and 51 women. Where the job distribution is shown in Figure 2. The percentage of four large respondents are 28.19% from private employees, 27.13% for lecturers, 19.15% for students, and 15.43% for self-employed.

Respondent distribution based on education is shown in Figure 3. The largest component of the respondents is 98 respondents (52%) from Bachelor degree, 62 respondents (33%) from Master degree, 13 respondents (7%) from Senior High School level, 4 respondents (2%) from 3-year Diploma, and 1 respondent from pharmacist education. The way to determine whether the user is competent or not is by calculating the highest number of opinion given by the respondent to the provided 256 famous people. The survey is conducted to obtain competencies from 256 famous people, including 115 famous people which the data are taken from Twitter. Competence sample data of 10 people is shown in Table 1.

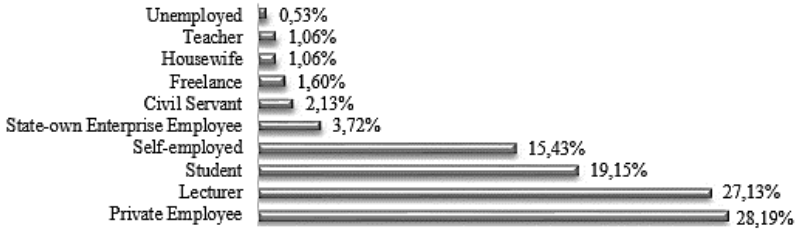

Figure 2. Job distribution of respondents

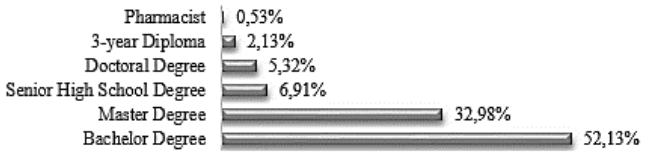

Figure 3. Education distribution of respondents

Table 1. Sample of 10 famous people competencies

No	Name	Competence 1	Competence 2	Competence 3	Competence 4
1	Abdullah Gymnastiar	religious	motivational	social	education
2	Aburizal Bakrie	political	governmental	economic	social
3	Acha Septriasa	entertainment	social	general	advertising
4	Addie MS	entertainment	cultural	social	general
5	Ade Komarudin	political	governmental	social	general
6	Adhicipta R. Wirawan	general	political	financial	economic
7	Adhie M Massardi	political	general	governmental	social
8	Adhyaksa Dault	political	governmental	sport	social
9	Adi Amran Sulaiman	social	political	governmental	general
10	Adib Hidayat	general	entertainment	social	journalism

Two credibility labels are used in this study, i.e., "credible" and "not credible". We define that information is considered as credible when the famous people posts tweet or message appropriate to their competencies. On the other hand, when the tweet or message are posted out of the famous people competencies, the information is considered as not credible. The process is shown in Figure 4

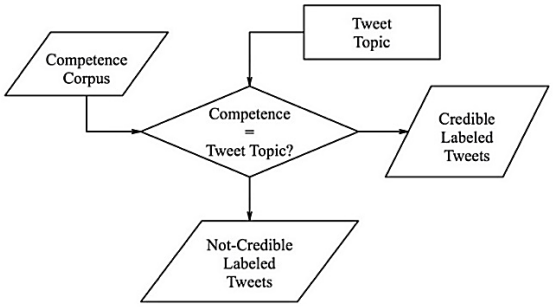

Figure 4. A labelling information credibility process by combining competence corpus and tweet topic

Data resulted from labeling are shown as follows:

a. Twitter social Media

The distribution of information credibility labeling for Twitter social media is shown in Table 2.

b. Facebook social media

The distribution of information credibility labeling for Facebook social media is shown in Table 3.

Table 2. Information credibility distribution in Twitter

Class	Number	%
Credible	12439	64.12
Not Credible	6962	35.88
Total	19401	

Table 3. Information credibility distribution in Facebook

Class	Number	%
Credible	15677	66.74
Not Credible	7812	33.26
Total	23489	

2.2. Pre-processing

By assuming text input from the original tweet (Twitter) or post message (Facebook) content, pre-processing consists of case folding, tokenization, stop-word removal, and stemming. Case Folding is the process by which words or phrases in a text tweet or post message will be converted into lowercase letters (a to z). This is expected to solve problems when words are written in different letters.

Tokenization is applied to cut the input of a tweet or post message from its composing words. In principle, separate each word in the text tweet or post message. This process includes deleting numbers, punctuation, and characters other than alphabetical letters. These characters are considered as word separators so they will be removed to prevent "noise" in further processes. Meanwhile, stop-word removal removes non-topical words that are not considered important such as: "and", "this', "that", "is", "or", "which", "through", and so on. This pre-processing helps reduce irrelevant features in the data. Finally, stemming is the process of finding root words by removing prefixes, infixes, suffixes, and confixes (combination of prefixes and suffixes) in derivative words. By originating, variations in words that have the same root will be considered the same way (feature). It helps improve retrieval performance on Information Retrieval.

2.3. Feature extraction

This section elaborates the feature extraction on Twitter and Facebook. The feature distribution, in both Twitter and Facebook, is attached, while the user profile dimension feature and message content dimension feature are also presented.

2.3.1. Features used on twitter

This paper uses two dimensions of features, namely the user profile dimension and message content dimensions. The most popular old features used by previous works have also been summarized in this study. In total, 33 features obtained from 5 different papers are discussed in this paper. The collection of features from works using classifiers is performed to predict credibility [3, 15, 16, 22, 18]. Furthermore, 17 new features are proposed in Table 4 indicated by underlined bold features.

Table 4. Feature distribution used in Twitter

No	Feature	Castillo (2011)	Morris (2012)	Gupta (2014)	Syariff (2014)	Ross (2016)	The Proposed
1	display_name				V		V
2	age_account_day	V		V			V
3	**check_web_institution**						**V**
4	has_bio	V	V			V	V
5	**words_desc**						**V**
6	**#positive_desc**						**V**
7	**#negative_desc**						**V**
8	**#sentiment_desc**						**V**
9	**numPosWordDesc**						**V**
10	**numNegWordDesc**						**V**
11	**Check_personal_web**						**V**
12	**Check_location**						**V**
13	is_verified	V	V	V		V	V
14	number_*follower*	V	V	V		V	V
15	number_*statuses*	V	V	V		V	V
16	number_*following*	V	V			V	V
17	**NumFollowingNumFollower**						**V**
18	***#likes_user***						**V**
19	**NumLikesNumFollower**						**V**
20	length_*tweet*	V		V		V	V
21	#words_tweet	V		V		V	V
22	#stock_char			V		V	
23	hasStockChar					V	
24	#colon_char			V		V	
25	hasColonChar					V	
26	#char			V			V
27	NumCharPanjangTweet	V				V	V
28	NumCharNumKata	V				V	V
29	#mention	V	V	V	V	V	V
30	#hashtag	V	V	V	V	V	V
31	#url	V	V	V	V	V	V
32	#emot_happy	V		V			V
33	has_happy					V	V
34	#emot_sad	V		V			V
35	has_sad					V	V
36	**check_spam**						**V**
37	**source**						**V**
38	is_url	V		V	V	V	V
39	is_mention	V	V		V		V
40	is_hashtag	V	V		V		V
41	is_retweet	V	V	V	V	V	V
42	***#like_tweet***						**V**
43	retweet_counted	V	V	V	V	V	V
44	#pos_tweet	V		V		V	V
45	#neg_tweet	V		V		V	V
46	**ratioPosNumTweet**						**V**
47	**ratioNegNumTweet**						**V**
48	#sentimen_tweet					V	V
49	sentiment_tweet					V	V

From 49 available features, only bout 45 features are used. Besides, its dimensions are divided into two dimensions, namely 19 features of the user profile dimension and 26 features of the message content dimension. The most widely used tweets feature for measuring credibility in tweets are retweeting, tweet length, number of words, number of mentions, number of hashtags, number of URLs, tweets having URLs, number of retweets, having happy emoticons, having sad emoticons, and value sentiments [22]. The description of each of the 45 features is shown in Tables 5 and 6.

Table 5. User profile dimension feature on Twitter

No	Feature	Description	New Feature
1	display_name	Whether the display name use the real name of the account owner or not. This is closely related to the level of trust.	No
2	age_account_day	In this feature, the age of the user's account can be seen. The longer the age of someone's account the higher the level of trust	No
3	check_web_institution	Having a URL that connects to the original website of the user's institution and it can be used to see the credibility	Yes
4	has_bio	If there is a description of the user's authenticity in the profile, then it can be a basis for assessing the user's credibility.	No
5	words_desc	The number of words which gives an explanation of whether the user explains the bio profile. A detailed explanation will make it easier for us to assess a person's credibility	Yes
6	#positive_desc	The number of positive sentiment words from an account's bio profile	Yes
7	#negative_desc	The number of negative sentiment words from an account's bio profile	Yes
8	#sentiment_desc	Number of sentiments from an account's bio profile	Yes
9	numPosWordDesc	The ratio of the sentiments number is positive towards the number of words in an account's bio profile. The value of the ratio is bigger equal to the value of the account's credibility.	Yes
10	numNegWordDesc	The ratio of the sentiments number is negative towards the number of words in an account's bio profile. The value of the ratio is getting smaller compared to the value of the account's credibility.	Yes
11	check_web_personal	Having a URL that connects to the user's original website and it can be used to see the credibility.	Yes
12	check_location	Having a location in the description can guarantee the authenticity of the user's original area.	Yes
13	is_verified	A verified account is an official account that has been authenticated by Twitter.	No
14	number_*follower*	The number of followers can help to find out how many other users want to see/follow the trail of information from the user. The number of followers can become an indication of the user's information credibility level, the more followers the higher the level of trust.	No
15	number_*statuses*	The number of statuses can inform the level of user's activity in using Twitter. Users who do more activities will have more credibility.	No
16	number_*following*	From the number of Following, it can be seen that the user has many friends who might be giving more sources of information. The number of Following shows many sources of information.	No
17	numFollowingNumFollower	The ratio of Following to the number of Followers of an account	Yes
18	*#likes_user*	The number of likes can show how active the user is in using Twitter. The number of likes can also indicate the number of truths of tweets that are liked by users.	Yes
19	numLikesNumFollower	The ratio of the number of Like to the number of an account's followers.	Yes

Table 6. Message content dimension feature on Twitter

No	Feature	Description	New Feature
1	length_*tweet*	The existence of which length of characters or words that could explain whether the user gives a short or long message to influence the perception of others.	No
2	#words_tweet	Which of the number of words that could explain whether the user gives a short or long message to influence the perception of others.	No
3	#char	Number of character in a tweet	No
4	numCharLengthTweet	The ratio of the number of characters compared to the length of a tweet	No
5	numCharNumWords	The ratio of the number of characters compared to the number of words from a tweet.	No
6	#mention	The number of mention from a tweet	No
7	#hashtag	The number of *hashtag* from a tweet. By clicking the #Hashtag in Twitter, the same information with the same hashtag will appear so that people will be assisted to find the information uniformity to digest the truth of the information with detail and clear history.	No
8	#url	The number of URL in a tweet	No
9	#emot_happy	The number of happy emoticons	No
10	has_happy	The existence of emoticon that contains happy expression	No
11	#emot_sad	The number of sad emoticons	No
12	has_sad	The existence of emoticon that contains sad expression	No
13	check_spam	To see whether a tweet has some words listed in spam.	Yes
14	Source	The means used to share a tweet can be divided into two, via a smartphone or PC Client.	Yes
15	is_url	A tweet with URL helps deliver more information so it can provide trust by giving the tweet source. The more in number of URLs given in a tweet the more credible the information is.	No
16	is_mention	Tweet contains Mention it means where its source was taken from someone else to provide better source certainty. The mention can indicate whether the mentioned user mentioned provides evidence of the news authenticity, for example, the user included photos of the evidence.	No
17	is_hashtag	The existence of *#hashtag* helps to ensure and view the news history in order to be able to seek information credibility. By clicking the #Hashtag in Twitter, the same information with the same hashtag will appear so that people will be assisted to find the information uniformity to digest the truth of the information with detail and clear history.	No
18	is_retweet	To know whether the tweet is posted by themselves or reposted (re-tweet) from others.	No
19	#like_*tweet*	The number of users' likes to a tweet	Yes
20	retweet_counted	The number of users who re-tweet a tweet.	No
21	#pos_tweet	The number of positive sentiments words from a tweet.	No
22	#neg_tweet	The number of negative sentiments words from a tweet.	No
23	ratioPosNumTweet	The ratio of the number of positive sentiments to the number of words in a tweet. The value of the ratio is bigger equal to the value of the account's credibility.	Yes
24	ratioNegNumTweet	The ratio of the number of negative sentiments to the number of words in a tweet. The value of the ratio is getting smaller equal to the value of the account's credibility.	Yes
25	#sentiment_tweet	The number of sentiments from a tweet's bio profile.	No
26	sentiment_tweet	The existence of positive, neutral, and negative sentiments to select the information that its credibility level is going to be seen. The positive sentiments are usually describing more credible information.	No

2.3.2. Features used in facebook

This paper successfully develops Facebook API application with 54 features (8 user profile dimension features, 46 message content features). Besides, 49 new features have been added from previous research [21]. Table 7 shows the user's dimension features in Facebook, while Table 8 shows the message content dimension features in Facebook.

Table 7. User profile dimension features in Facebook

No	Feature	Description	New Feature
1	check_bio	The authenticity description in the user's profile can become a basis to know the user's credibility.	Yes
2	#word_bio	The number of words in describing the user's profile (bio profile). A detailed description can make it easier to know someone's credibility.	Yes
3	length_bio	The length of character and words that explain whether the user gives a short or long message that could influence someone's perception.	Yes
4	#positive_desc	The number of positive sentiment words in an account bio profile	Yes
5	#negative_desc	The number of negative sentiment words in an account bio profile	Yes
6	sentiment_desc	The existence of positive, neutral, and negative sentiments to select the information that its credibility level is going to be seen. The positive sentiments are usually describing more credible information.	Yes
7	#url_institution	Having a URL that connects to the original website of the user's institution and it can be used to see the credibility	Yes
8	engagement_count	The number of engagement shows the number of other users who want to see/follow the user's trail of information. The number of engagement can become an indication of the user's information credibility level. The more engagement the higher the trust.	Yes

Table 8. Message content features in Facebook

No	Feature	Description	New Feature
1	type	The classification of post message types (photo, link, status, note, video, event)	Yes
2	#url_post	The number of URL in a post message	No
3	#char	The number of character in a post message	Yes
4	ratioCharLenghtWordPost	The ratio of the character number compared to the length of a post message	Yes
5	ratioCharNumWord	The ratio of the character number compared to the number of words in a post message	Yes
6	#mention	The number of mention in a post message	Yes
7	#hashtag	The number of the hashtag in a post message. By clicking the #hashtag in Twitter, the same information with the same hashtag will appear so that people will be assisted to find the information uniformity to digest the truth of the information with detail and clear history.	No
8	#emot_happy	The number of happy emoticons	Yes
9	has_happy	The existence of emoticon that contains happy expression	Yes
10	#emot_sad	The number of sad emoticons	Yes
11	has_sad	The existence of emoticon that contains sad expression	Yes
12	#word	Which of the number of words that could explain whether the user gives a short or long message to influence the perception of others.	Yes
13	length_message	Which of the length of the character and word that could explain whether the user gives a short or long message to influence the perception of others.	Yes
14	check_spam	To see whether the post message contains the words included in the spam list	Yes
15	check_full_picture	To check the presence or absence of the full picture in a post message	Yes
16	link_domain	The presence of a post message with URL helps deliver more information so it can provide trust by giving the post message source. The more in number of URLs given in a post message the more in certainty to the credibility of the information.	Yes
17	post_published	The age of a post message on the number of days is based on when the last post message was taken	Yes
18	likes_count_fb	The number of like count for a post message	No
19	likes_count_fb_per_day	The number of like count for a post message in the number of days based on the age of the post message	Yes
20	comments_count_fb	The number of comments in a post message	No
21	comments_count_fb_per_day	The number of comments for a post message in the number of days based on the age of the post message	Yes
22	reactions_count_fb	The number of short response activities with certain icons (like, none, love, wow, haha, sad, angry, thankful) in a post message	Yes
23	reactions_count_fb_per_day	The number of short response activities with certain icons (like, none, love, wow, haha, sad, angry, thankful) in the age of a post message	Yes
24	shares_count_fb	The number of users who share a post message	No
25	shares_count_fb_per_day	The number of users who share a post message each day based on the age of the post message	Yes
26	engagement_fb	The number of interaction in a post message (share, like, comment)	Yes
27	engagement_fb_per_day	The number of interaction in a post message (share, like, comment) each day based on the age of the post message	Yes
28	comments_retrieved	The number of comments in a post or by the user	Yes
29	comments_base	The number of basic level comments	Yes
30	comments_replies	The number of comment level replying	Yes
31	comment_likes_count	The number of like in a comment of a post message	Yes
32	rea_NONE	The number of short response activities by NONE in a post message	Yes
33	rea_LIKE	The number of short response activities by LIKE in a post message	Yes
34	rea_LIKE_per_day	The number of short response activities by LIKE in a post message each day based on the age of the post message	Yes
35	rea_LOVE	The number of short response activities by LOVE in a post message	Yes
36	rea_WOW	The number of short response activities by WOW in a post message	Yes
37	rea_HAHA	The number of short response activities by HAHA in a post message	Yes
38	rea_SAD	The number of short response activities by SAD in a post message	Yes
39	rea_ANGRY	The number of short response activities by ANGRY in a post message	Yes
40	rea_THANKFUL	The number of short response activities by THANKFUL in a post message	Yes
41	#positive	The number of positive sentiment words in a post message	Yes
42	ratioPosNumWord	The ratio of the number of positive sentiments to the number of words in a post message. The value of the ratio is bigger equal to the value of a post message credibility	Yes
43	ratioNegNumWord	The ratio of the number of negative sentiments to the number of words in a post message. The value of the ratio is getting smaller equal to the value of a post message credibility	Yes
44	#negative	The number of negative sentiments in a post message	Yes
45	#sentiment	The number of sentiments in a post message	Yes
46	sentiment	The existence of positive, neutral, and negative sentiments to select the information that its credibility level is going to be seen. The positive sentiments are usually describing more credible information.	Yes

In addition, this paper also applies a new approach related to the spam prediction and sentiment prediction described as follows:

a. Spam prediction (*check_spam*)

We use two corpuses related to the spam words or phrases that are 200 English spam words or phrases and 100 Bahasa Indonesia spam words or phrases as used in our previous study [23]. The two corpuses are developed based on Indonesia spam-words. Table 9 describes 12 examples of Bahasa Indonesia spam words or phrases [23].

b. Sentiment prediction

This paper uses a corpus which contains the list of sentiments words consists of 354 words [24]. This sentiment is obtained by searching for words that are categorized as negative, positive and neutral. Some sample data are shown in Table 10 [24].

Table 9. Samples of 12 Indonesian spam-words

No	Indonesian Spam-words
1	*kredit dp*
2	*paket kredit*
3	*cicilan ringan*
4	*dp ringan*
5	*cash/kredit*
6	*dana tunai*
7	*proses cepat*
8	*dana cepat*
9	*pinjaman uang*
10	*pinjaman dana*
11	*pinjaman*
12	*gadai*

Table 10. Ten data survey of sentiment

No	Word	Positive (%)	Negative (%)	Neutral (%)	Quality
1	*buruk*	0	78.3	21.7	3
2	*jelek*	0	78.3	21.7	3
3	*lama*	4.3	30.4	65.3	0
4	*lamban*	4.3	78.3	17.4	3
5	*lambat*	13	52.2	34.8	1
6	*baik*	82.6	0	17.4	4
7	*berani*	82.6	0	17.4	4
8	*benar*	82.6	0	17.4	4
9	*sudah*	56.5	0	43.5	1
10	*ayo*	65.2	4.3	30.5	2

2.4. Classification algorithm

The four learning algorithms that will be explored are Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (Logit) and J48. As illustrated in Fig. 1, the four algorithms are used to model the topic classification of tweets during the training phase. The topic model of tweets is then used to classify the credibility of new information, using the same algorithm as that used to model the classification. The following is a description of each algorithm.

a. Naive bayes (NB)

Naive Bayes (NB) is a classification model in the form of probability values for each attribute to the class, and the classification of new data is done by looking at classes that have the maximum probability based on attribute data [25]. Naive Bayes has the advantage of construction easiness which does not require several complex parameters, and it is scalable. In addition, this method is expressed as an algorithm that has the properties of simplicity, elegance, robustness, and high accuracy [26].

b. Support vector machine (SVM)

The idea of Support Vector Machine (SVM) for classification is to find the optimal hyperplane (line/boundary field) that separates data into two classes in the data n-dimensional feature space. With this concept, the optimal hyperplane solution in SVM does not have a local optimum, and as a result, the solution will be unique [25]. SVM can be implemented easily and is one of the right methods used to solve high-dimensional problems within the limitations of existing data samples.

c. Logistic regression (Logit)

Logistic Regression (Logit) is a probability classification model with a real value input vector. The input vector dimensions are called features. There are no restrictions imposed for correlated features. Logistic Regression is used every time we need to set input to one of several classes. The logistics function is a linear combination of features. The output is usually binary, but Logistic Regression can also be applied to multiclass classification problems [25].

d. J48 algorithm (J48)

J48 is a development of the ID3 algorithm. J48 is an implementation of the C4.5 algorithm that produces a decision tree. This algorithm can classify data with decision tree methods that have the advantage of being able to process numerical (continuous) and discrete data, can handle missing attribute values, and produce rules that are easily interpreted. Each data from an item is based on the value of each attribute. Classification can be seen as a mapping of a group of sets of attributes from a particular class. Decision tree classifies the data given using the value of the attribute [27].

3. RESULTS AND ANALYSIS

This section provides the results and analysis of the data set and labeling scheme for Twitter and Facebook.

3.1. Data set for experiment

The use of Twitter data containing Indonesian language is the same as in [28], involving 115 accounts with 19401 tweets. Table 11 provides a sample labeling of tweet topics from Law, Politics, and Entertainment [28]. Table 12 shows the distribution of Twitter data. It consists of 19 topics where the distribution is not balanced ranging from 0.2% to 15.3% [28].Facebook data used in this study consists of 56 accounts with 23489 messages. Due to the absence of a user account, not all accounts on Twitter (115 accounts) can be retrieved. Table 13 describes the distribution of Facebook data used, consists of 19 topics, which shows that the distribution is also unbalanced, ranging from 0.17% to 18.38%.

Table 11. Some samples of category labeling in Twitter

Tweet	Label
Yg disoal Saripin cm apakah KPK berwenang sidik #BG, dugaan korupsinya tdk diusik. Bagi saya, BG tetap "tersangka" mestinya #JKW jg demikian	Law
DPR Akan Gelar Paripurna Sahkan Revisi UU Pilkada Hari Ini http://t.co/jcxclL9faO @detikcom	Political
Studio Denny JA, MTV dan Mizan bersama HanungBramantyo membuat 5 film layar lebar bertema Islam Cinta: http://t.co/BrdHfhBsub	Entertainment

Table 12. Twitter data distribution by topic/category

No	Label	Number	Percentage
1	Religion	1025	5.28%
2	Business	460	2.37%
3	Culture	235	1.21%
4	Economy	235	1.21%
5	Entertainment	1742	8.98%
6	Law	1557	8.03%
7	Advertisement	485	2.50%
8	Journalism	2420	12.47%
9	Health	74	0.38%
10	Finance	35	0.18%
11	Motivation	927	4.78%
12	Sports	431	2.22%
13	Government	1935	9.97%
14	Education	466	2.40%
15	Transportation	149	0.77%
16	Political	2959	15.25%
17	Social	1238	6.38%
18	Technology	1218	6.28%
19	General	1810	9.33%
	Total	**19401**	

Table 13. Facebook data distribution by topic/category

No	Category	Number	Percentage
1	Religion	1952	8.31%
2	Business	1267	5.39%
3	Culture	87	0.37%
4	Economy	1421	6.05%
5	Entertainment	2977	12.67%
6	Law	1329	5.66%
7	Advertisement	331	1.41%
8	Journalism	96	0.41%
9	Health	313	1.33%
10	Finance	41	0.17%
11	Motivation	1169	4.98%
12	Sports	495	2.11%
13	Government	3613	15.38%
14	Education	775	3.30%
15	Transportation	43	0.18%
16	Political	4287	18.25%
17	Social	630	2.68%
18	Technology	2022	8.61%
19	General	641	2.73%
	Total	**23489**	

3.2. Experiment

We consider three objectives of performing experiment, i.e., (i) to compare the proposed technique with previous research on Twitter and Facebook about information credibility, (ii) evaluate the effect of adding new features in Twitter and Facebook, and (iii) evaluate the effects of feature dimensions used both on Twitter and Facebook. Our experiments used a comparison of training data versus data testing, with a composition of 80:20.

3.2.1. Twitter social media

In this study, each cell describes an average of 5 times of the accuracy taking for each testing vs twitter composition taken randomly. The results of the proposed method and the previous research are shown in Table 14. Table 14 shows that this paper succeeded in increasing the accuracy of previous studies in almost all classifiers. When compared to previous studies, it can be seen that the highest accuracy is 88.42% achieved by using J48 classifier with the lowest increase of 5.93% and the highest of 27.17%.

Table 14. The proposed and previous research results in Twitter

Classifier	Percentage (%)							
	Castillo (2011)	Morris (2012)	Gupta (2014)	Syariff (2014)	Ross (2016)	The Proposed Feature		
						User Profile	Message Profile	User Profile + Message Profile
NB	60.79	58.41	60.79	65.95	60.52	66.97	62.93	66.42
SVM	77.70	82.36	76.73	67.86	77.13	82.70	73.41	87.36
Logit	77.26	77.24	76.27	66.25	76.56	78.25	64.57	78.04
J48	83.09	83.47	83.00	69.53	82.16	82.77	79.36	**88.42**

Table 14 also shows a comparison of the accuracy value between the user profile dimension and message content dimension in 4 different classifiers. The user profile dimension accuracy is higher than the message content dimension accuracy for all classifiers. The highest accuracy value on the user profile dimension using the J48 classifier is 82.77%. All the merging the features of the both dimensions used in this study increase accuracy for SVM and J48 classifiers, while the two other classifiers, i.e., NB and Logit classifiers, provide a decrease on the accuracy.

The new features are classified according to the influence of them on the accuracy. The features that increase the accuracy after it added to the baseline features are classified as increased group, the features that decrease the reverse are classified as decreased group, while the features that not effect are classified as mixed group, in this paper. Here, the baseline features represent the set of feature used in Ross and Thirunarayan [22].

Table 15 shows the effect of the 17 new features proposed, consist of 12 features based on user profile and 5 features based on message content dimension, in each classifier on Twitter. All features proposed on Twitter in both feature dimensions provide an increase on the accuracy of each classifier. For influence on all classifiers, all new feature increase on the accuracy of 6.60%, with 6.67 % for features based on user profile, and 6.45% for features based on message content dimension. The biggest average for feature is 8.55%, achieved by the NumFollowingNumFollower feature. In the terms of the effect in each classifier, #sentiment_desc feature provides the highest improvement of accuracy of +13,41 % was achieved on SVM classifier.

Table 15. New features distribution by influence of accuracy based on features dimension on Twitter

Influence of accuracy	Feature dimension	
	User Profile	Message Content
Increased	check_web_institution, #sentiment_desc, numPosWordDesc, check_web_personal, NumFollowingNumFollower, NumLikesNumFollower, word_desc, #positive_desc, #negative_desc, check_location, #likes_user, numNegWordDesc	source, ratioNegNumTweet, #like_tweet, check_spam, ratioPosNumTweet
Decreased	-	-
Mixed	-	-

3.2.2. Facebook social media

This paper has carried out two developments. First, developing Facebook API that can retrieve datasets online. Second, adding more features to 49 new features based on users and content. Table 16 shows the highest accuracy increase compared to Saikaew's study. This paper succeeded in increasing the accuracy of previous studies in almost all classifiers. The increase is 9.91% with an accuracy value of 78.61% by using J48 Classifier. Table 16 also shows a comparison of the accuracy value between the user profile dimension and message content dimension in 4 different classifiers. The user profile dimension accuracy is higher than the message content dimension accuracy for all classifiers. The highest accuracy value on the user profile dimension using the SVM classifier is 76.50%. All the merging the features of the both dimensions used in this study increase accuracy for only J48 classifiers, while the three other classifiers provide a decrease on the accuracy.

Table 16. Saikaew vs the proposed in Facebook

Classifier	Percentage (%)			
	Saikaew (2015)	The Proposed Feature		
		User Profile	Message Profile	User Profile + Message Profile
NB	65.02	66.58	62.32	65.39
SVM	71.10	76.50	71.38	71.83
Logit	69.93	73.41	70.54	72.57
J48	71.52	76.46	74.61	**78.61**

Table 17 shows the effect of the 49 new features proposed, consist of 8 features based on user profile and 41 features based on message content dimension, in each classifier on Facebook. Here, the baseline features used as the comparison is representing the set of feature used in Saikaew [21]. All proposed features based on user profile dimension provide an increase on the accuracy of each classifier, whereas for message content based only 27 features or equal to 65.85% which give an increase in accuracy, remaining is 14 features or 34.15% provide mixed results.

Table 17. New features distribution by influence of accuracy based on features dimension on Facebook

Influence of Accuracy	Features dimension	
	User Profile	Message Content
Increased	check_bio, #word_bio, length_bio, num_positive_desc, num_negative_desc, sentiment_desc, #url_institution, engagement_count	type, #char, ratioCharLengthWordPost, ratioCharNumWord, #mention, #emot_happy, has_happy, #emot_sad, #word, length_message, check_spam, check_full_picture, link_domain, post_published, likes_count_fb_per_day, reactions_count_fb_per_day, engagement_fb, engagement_fb_per_day, comments_retrieved, comments_base, rea_LIKE, rea_LIKE_per_day, rea_SAD, num_positif, ratioPosNumWord, #sentiment, sentiment
Decreased	-	-
Mixed	-	has_sad, comments_count_fb_per_day, reactions_count_fb, shares_count_fb_per_day, comments_replies, comment_likes_count, rea_NONE, rea_LOVE, rea_WOW, rea_HAHA, rea_ANGRY, rea_THANKFUL, ratioNegNumWord, num_negative

For influence on all classifiers, all new feature increase on the accuracy of 0.57%, with 2.64 % for features based on user profile, and 0.17% for features based on message content dimension. The biggest average for feature is 7.26%, achieved by the engagement_count feature. In the terms of the effect in each classifier, engagement_count feature also provides the highest improvement of accuracy of +11,98% was achieved on J48 classifier.

The additional new feature on Twitter and Facebook are found to provide the best accuracy value and are influencing the credibility of the information, where the results are shown in Tables 14 and 16. It is clearly shown that user profile dimension is having a higher accuracy compared to message content dimension for all classifiers. Based on these results, it can be concluded that the credibility of information can be seen from the Twitter users. In searching for the information from Twitter, making users who provide content or tweets as the source of information can add the credibility and trust. This result confirm that purpose concept is practical and reliable. Finally, the effect of two feature dimensions, user profile dimension and message content dimension, on Twitter and Facebook are also found to provide the best accuracy value and are influencing the credibility of the information, where the results are shown in Tables 15 and 17. It is clearly shown that user profile dimension is more consistent increasing accuracy than message content dimension for all classifiers.

4. CONCLUSION

In this study, a method to measure the credibility of information on social media, i.e., Twitter and Facebook, has been proposed using labeling process and additional new features. We introduced 17 new features for Twitter and 49 new features for Facebook. We also used 4 classification methods, i.e., NB, SVM, Logit and J48 Algorithms. By adding new features, we obtained an accuracy of measurement about 88.42% for Twitter and 78.61% for Facebook, which is better than the previous results for all classifiers. In terms of the two feature dimensions, the user profile dimension accuracy is found to be better than the message content dimension for all classification conditions. Finally, the effect of new features to accuracy, all features proposed on Twitter in two feature dimensions provide an increase of accuracy for all classifiers. Furthermore, in Facebook, all the proposed features based on user profile dimension provide an increase of accuracy for all classifiers. However, in Facebook, from the view point of message content dimension, only 27 features (65.85%) provided an increase in accuracy. On the other hand, the remaining 14 features (34.15%) provided mixed results. For all conditions, we found that the user profile dimension is more consistent to increase the accurate measurement rather than the message content dimension for all classifiers. We are expecting that these results can provide contributions to the future development of information credibility on social media.

ACKNOWLEDGEMENTS

The authors would like to thank PDD Hibah Dikti 2018 and BPPDN RISTEKDIKTI for the support to this research. The authors would also would like to thank Dr. Eng. Khoirul Anwar for the discussions to improve this paper.

REFERENCES

[1] H. Kwak, C. Lee, H. Park, S. Moon, "What is Twitter, a Social Network or a News Media? Categories and Subject Descriptors," *Proc. 19th Int. World Wide Web Conf*, Raleigh, North Carolina, USA, pp. 591–600, Apr. 2010.

[2] J. Teevan, D. Ramage, and M. R. Morris, "#TwitterSearch: a comparison of microblog search and web search," *Proc. fourth ACM Int. Conf. Web search data Min. - WSDM '11*, Hong Kong, China, pp. 35-43, Feb. 2011.

[3] M.R. Morris, *et al.*, "Tweeting is Believing? Understanding Microblog Credibility Perceptions," *CSCW '12 Proc. ACM 2012 Conf. Comput. Support. Coop. Work*, Seatle, Washington, USA, pp. 441–450, Feb. 2012.

[4] A. Java, X. Song, T. Finin, and B. Tseng, "Why We Twitter: Understanding Microblogging," *Proc. 9th WebKDD 1st SNA-KDD 2007 Work. Web Min. Soc. Netw. Anal*, San Jose, California, USA, pp. 56–65, Aug. 2007.

[5] M. Naaman, M. Naaman, J. Boase, J. Boase, C. H. Lai, and C. H. Lai, "Is it Really About Me? Message Content in Social Awareness Streams," *Dot-Me.of-Cour.Se*, Savannah, Georgia, USA, pp. 0–3, Feb. 2010.

[6] M. Mendoza, B. Poblete, and C. Castillo, "Twitter Under Crisis: Can we trust what we RT?," *Work. Soc. Media Anal*, Washington, DC, USA, pp. 9-17, Jul. 2010.

[7] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors," *WWW '10 Proc. 19th Int. Conf. Worldwide web*, Raleigh, North Carolina, USA, pp. 851-860, Apr. 2010.

[8] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging during two natural hazards events: what Twitter may contribute to situational awareness," *CHI 2010 Cris. Informatics April 10–15, 2010, Atlanta, GA, USA*, Atlanta, Georgia, USA, pp. 1079–1088, Apr. 2010,

[9] K. Starbird, *et al.*, "Chatter on the Red: What Hazards Threat Reveals about the Social Life of Microblogs Information," *Proc. Conf. Comput. Support. Coop. Work*, Savannah, Georgia, USA, pp. 241–250, Feb. 2010.

[10] A. L. Hughes and L. Palen, "Twitter adoption and use in mass convergence and emergences," PAISI 2012. *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, vol. 7299, 2012.

[11] Grier, C., Thomas, K., Paxson, V., and Zhang, M., "@ spam : The Underground on 140 Characters or Less Categories and Subject Descriptors," *Proceedings of the 17th ACM conference on Computer and communications security*, http://doi.org/10.1145/1866307.1866311, pp. 27-37, 2010.

[12] Gupta, A, "Twitter Explodes with Activity in Mumbai Blasts! A Lifeline or an Unmonitored Daemon in the Lurking?," Indraprastha Institute of Information Technology, Delhi, pp. 1-7, 2011. [Online], Available: http://precog.iiitd.edu.in/Publications_files/AG_PK_TR_2011.pdf,

[13] Xia, X., Yang, X., Wu, C., Li, S., & Bao, "L. Information credibility on twitter in emergency situation" *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, LNCS, vol. 7299, pp. 45-59, 2012. http://doi.org/10.1007/978-3-642-30428-6_4.

[14] B. De Longueville and R. S. Smith, "OMG, from here, I can see the flames!': a use case of mining Location Based Social Networks to acquire spatio-temporal data on forest fires," *ACM LBSN '09*, Seattle, WA, USA, pp. 73–80, Nov. 2009.

[15] S. M. Shariff, X. Zhang, and M. Sanderson, "User Perception of Information Credibility of News on Twitter," ECIR 2014: Advances in Information Retrieval, *Lecture Notes in Computer Science, © Springer International Publishing Switzerland*, vol 8416, pp. 513–518, 2014.

[16] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," *Proc. 20th Int. Conf. World wide web - WWW '11*, Hyderabad, India, pp. 675–684, March 2011.

[17] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking Sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy," *Proc. 22nd of International World Wide Web Conference Committee (IW3C2)*, ACM 978-1-4503-2038-2/13/05, Rio de Janeiro, Brazil, pp. 729–736, May 2013.

[18] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier, "TweetCred: Real-Time Credibility Assessment of Content on Twitter," *SocInfo 2014: Social Informatics, Lecture Notes in Computer Science, Springer International Publishing Switzerland*, vol. 8851, pp. 228-243, 2014. https://doi.org/10.1007/978-3-319-13734-6_16.

[19] A. Gupta and P. Kumaraguru, "Credibility ranking of tweets during high impact events," *Proc. 1st Work. Priv. Secur. Online Soc. Media PSOSM 12*, Lyon, France, pp. 2–8, Apr. 2012.

[20] M. Brandt, "80% Of Twitter's Users Are Mobile, Statista," wearesocial.com, 2017. [Online], Available: https://wearesocial.com/special-reports/digital-southeast-asia-2017 (accessed October 10, 2018),

[21] K.R. Saikaew, C. Noyunsan, "Features for Measuring Credibility on Facebook Information," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 9, no. 1, pp. 174–177, 2015.

[22] J. Ross, K. Thirunarayan, "Features for Ranking Tweets Based on Credibility and Newsworthiness," *International Conference on Collaboration Technologies and Systems Features*, Orlando, FL, USA, pp. 18–25, 2016.

[23] E.B. Setiawan, D.H. Widyantoro, and K. Surendro, "Detecting Indonesian Spammer on Twitter," *6th Int. Conf. Inf. Commun. Technol. (ICOICT)*, paper 10, 2018.

[24] E.B. Setiawan, D.H. Widyantoro, and K. Surendro, "Feature Expansion for Sentiment Analysis in Twitter," *Proc. EECSI 2018*, Malang, Indonesia, pp. 16–18, Oct. 2018.

[25] T.M. Mitchell, "Machine Learning," *McGraw-Hill Science*, 1st Edition, 1997.

[26] X. Wu, *et al.* "*Top 10 algorithms in data mining"*, *International Journal of Knowledge and Information Systems*, Springer-Verlag, vol. 14, no. 1, pp. 1–37, 2008. https://doi.org/10.1007/s10115-007-0114-2.
[27] I.H. Witten and E. Frank, "Data Mining - Practical Machine Learning Tools and Techniques (The Morgan Kaufmann Series in Data Management Systems)," *Elsevier, 2nd Edition*, San Francisco: Morgan Kaufmann Publishers, 2005.
[28] E.B. Setiawan, D.H. Widyantoro, and K. Surendro, "Feature Expansion using Word Embedding for Tweet Topic Classification," *10th International Conference on Telecommunication Systems Services and Applications (TSSA)*, Denpasar, Indonesia, pp. 1-5, 2016.

BIOGRAPHIES OF AUTHORS

Erwin Budi Setiawan is a doctoral student in School of Electrical Engineering and Informatics Institut Teknologi Bandung (ITB), Bandung, Indonesia. He has more than 10 years Research and Teaching experience in the domain of Informatics. Currently, he is a Senior Lecturer (equivalent to Associate Professor) with the Faculty of Informatics, Telkom University. His research interests are machine learning and social media analysis.

Dwi Hendratmo Widyantoro is currently a professor of the Department of Informatics at the Institut Teknologi Bandung (ITB), Bandung, Indonesia. He completed his master's study and PhD from Texas A&M University, USA. He has more than 10 years Research and Teaching experience in the domain of Computer Science. His research focuses on machine learning, information retrieval, information extraction and information retrieval. He currently also holds the position of deputy dean of academic at School of Electrical Engineering and Informatics, Institut Teknologi Bandung. He is also a member of boards of international conferences and as chief editor in Journal of ICT Research and Applications.

Kridanto Surendro is currently an assistant professor of the Department of Informatics at the Institut Teknologi Bandung (ITB), Bandung, Indonesia. He has more than 10 years Research and Teaching experience in the domain of Computer Science. He received Ph.D. degree in Computer Science from the Keio University, Japan, in 1999. His research focuses on information science, software engineering, information systems (business informatics).