TELKOMNIKA, Vol.15, No.1, March 2017, pp. 357~364
ISSN: 1693-6930, accredited **A** by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v15i1.4557

A Framework for Classifying Indonesian News Curator in Twitter

Jaka E. Sembodo[1], Erwin B. Setiawan[2], ZKA Baizal[3]
School of Computing, Telkom University, Telecommunication Street No. 1 Bandung 40257
Corresponding author, e-mail: jecksart@gmail.com[1], setiawanerwinbudi@gmail.com[2], bayzal@gmail.com[3]

Abstract

News curators in twitter are a user, which is interested in following, spreading, giving feedback of recent popular articles. There are two kinds of this user, news curator as human user and news aggregator as bot user. In prior works about news curator, the classification system built using followers, URL, mention and re-tweet feature. However, there are limited prior works for classifying Indonesian News Curator in twitter and still hard for labeling data involve just two features: followers and URL. In this paper, we proposed a framework for classifying Indonesian news curator in twitter using Naïve Bayes Classifier (NBC) and added features such as location, bio profile, and common tweet. Another purpose for analyzing the influential features of certain class, so it's make easier for labeling data of this role in the future. Examination result using percentage split as evaluating system produced 87% accuracy. The most influential features for news curator are followers, bio profile, mention and re-tweet. For news aggregator class are followers, location, and URL. The rest just common tweet feature for not both class. We implemented Feature Subset Selection (FSS) for increasing system performance and avoiding the over fitting data, it has produced 92.90% accuracy.

Keywords*: twitter, machine learning, Indonesian news curator, naïve Bayes classifier*

1. Introduction

Social media is web-based communication tools that made people interacting each other's for giving and consuming information. In this era, there was various social media development, from text-based post, image-based post or mixed content-based post. The instance is twitter [1-3]. Twitter is one of social network that was first launched in July 2006 by Jack Dorsey and now days used by many people in the world. Twitter had the function as a social media micro-blogging type with sum of character in a tweet (name of post in twitter) is maximum 140 characters. The other features of twitter is people can re-post the tweet (re-tweet), made the hash tag, like the others tweet and many more.

In the academics, twitter could be research object that caused its unique and massively used by people around the world that would be update by time period. The instance, research in twitter could be like sentiment knowledge discovery analysis [4], determining trust scope attributes using goodness of fit test for Indonesian user [5], built framework for classifying the user or text in certain object [6-9] or twitter could be research object for helping journalist and news editor in the news development [1-3].

Based on survey [1], 613 journalists over 16 countries reveal that 54% of them using social media (include twitter) for looking the other side of news article development. By the increasing engagement user from journalist and news editor, it increased the number of news crowds by time period in twitter. Another function of twitter from journalist-side or news editor, it was place for knowing user that interested in following, spreading the URL, giving self-opinion or giving feedback of recent popular articles. This role defined as news curator in twitter [1]. News curators consist of two kind of user. First, news curator as the human user, the user that has characteristic who made tweet about news manually not using bot or automatically and made interaction with other users. Second, news aggregator as bot user, the user that made news tweet automatically usually using web-integrated tweet (tweets made after the news articles made) or using bot program and didn't interact with other users [1]. In the context of journalist, this role helped the journalist or news editor for another news feedback perspective. Also for the social public can be like recommendation for news source.

Received September 13, 2016; Revised December 25, 2016; Accepted January 17, 2017

Prior work that introduced news curators role in twitter [1-3], the techniques used are data mining and machine learning. Data mining used for searching and processing data. Machine learning used for the machine to learning data in processing data. The method used is classification. Classification is the method to determine the class of data based on its characteristics/features/attributes, and the researcher used the random forest algorithm with features such as followers, URL, mention and re-tweet was used [1].

Now days, there are limited prior works for classifying Indonesian News Curator in twitter and still hard for labeling data involve just two features: followers and URL. In this paper, we proposed a framework for classifying Indonesia news curator in twitter. We added bio profile, location and common tweet feature and analyzed the influential feature of certain class for labeling data easier in the future.

The paper is organized as follows. We introduced the background and news curator in section 1. We explained our proposed framework design in section 2. Then, we presented the result such as dataset, implementation, examination and analysis in section 3. The last, we made conclusion in section 4.

2. Framework Design

In this paper, we proposed framework for classifying Indonesian news curators in twitter. The frameworks consist of five main steps: first, we collected user include its features using twitter API streaming. Second, we did data preprocessing. Then, we labeled it manually based on labeling data justification in prior work [1]. Then, we classified data using supervise machine learning and evaluated the system using percentage split. The flowchart of framework in this paper shown on Figure 1.

Figure 1. Framework for Classifying Indonesian News Curator in Twitter

2.1. Collecting Data

Dataset that we used in this paper is username and tweets involved its feature. First, we searched username that could be indicated as news curators using journalistic keywords from Google and twitter such as: “news”, ‘article”, “journalist” “news anchor” and others. Then, we downloaded data using twitter API streaming that has been built and integrated with PHP programming. We using consumer key, consumer access, access token and accesss secret token for the permission connection that we got from registering in twitter application management. So, we downloaded tweet data include its features and stored it in database automatically [10].

There were 2 groups of feature that we used in this paper. First, group of visible feature, the feature that we can found on user profile. And second, group of tweet activity, the features based on tweet substance that users made. The feature’s detail from group of visible feature shown on Table 1 and from group of tweet activity shown on Table 2.

Table 1. Group of Visible Feature

No	Name of feature	Type of Data	Description
1	Total of Followers	Integer	Number of followers that user have.
2	Bio Profile	Text	Description of user in twitter. Wrote by user-self.
3	Location	Text	Location where the user stay/live. Wrote by user.

Table 2. Group of Tweet Activity Feature

No	Name of feature	Type of Data	Description
1	URL	Boolean	If tweet contained URL the boolean value is “yes”, if not the boolean value is “no”.
2	Re-tweet	Boolean	If tweet is re-tweet the boolean value is “yes”, if not the boolean value is “no”.
3	Mention	Boolean	If tweet contained mention the boolean value is “yes”, if not the boolean value is “no”
4	Common Tweet	Boolean	If the tweet didn’t contain URL, mention and re-tweet, the Boolean value is “yes”, if it’s contained one of them the Boolean feature is “no”.

2.3. Data Pre-processing Data

We did pre-processing data for making it enable in classification system. In general, data preprocessing data described on Figure 2.

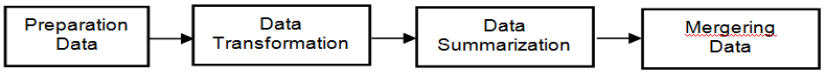

Figure 2. Flowchart of Pre-processing Data

2.3.1. Data Transformation

Transformation data is process to transformed features from visible features group such as: location, follower and bio profile. The proposed made text data being a boolean data (yes or no). For the instance, original data that we used in this paper shown in Table 3.

Table 3. Example Data for Transformation Data

No.	username	location	follower	bio profile
1	@AJIIndo	Jakarta - Indonesia	12K	Aliansi Jurnalis Independen/The Alliance of Independent Journalist (AJI) Indonesia

Our treatment for transform texted features to a boolean features following steps:

Step 1: Location feature, if location consist word of name “Indonesia” or city name of Indonesia like “Jakarta”, “bandung” or another, we annotated as yes. If there was not, we annotated as no.

Step 2: Follower feature, if follower up to 1000 we annotated as yes and if below 1000 we annotated as no.

Step 3: Bio profile feature, if the text in bio profile contained word that related with journalistic (example: “journalistic”, “News Anchor”, “News Portal” etc) we annotated as yes and if was not, we annotated as no.

The instance for transformation data shown in Table 4.

Table 4. The Result of Transformation Data

No.	username	location	follower	bio profile
1	@AJIIndo	Yes	Yes	Yes

2.3.2. Data Summarization

Data summarization is process for calculating sum of each feature in group of tweet activity feature. For the instance, from @AJIindo we downloaded 10 tweets and in total URL was the sum of “yes” in URL, total mention was the sum of “yes” in mention, total re-tweet was the sum of “yes” in re-tweet and the last, total of common tweet was the sum of “yes” in common tweet. For the instance shown in Table 5.

Table 5. Result of Summarization Data

no.	account	total of URL	total of mention	total of re-tweet	total of common tweet
1	@AJIIndo	9	6	3	2

2.3.3. Mergering Data

Mergering data is process merged data from preprocessing of visible features group on (2.3.1) and preprocessing of tweet activity features group on (2.3.2) be a table. The proposed of this process is made the data already for using by classification system. The instance shown on Table 6.

Table 6. Result of Merging Data

No	username	Location	Follower	Bio profile	Total of URL	Total of mention	Total of re-tweet	Total of common tweet
1	@AJIIndo	Yes	Yes	Yes	9	6	3	2

2.2. Labelling Data

In this paper, we used 3 classes as output from proposed classification system, there are:

a. News Curator : user that interest for sharing news url and interacted to others.
b. News Aggregator : user that share tweet contain news url and didn't interacted to others.
c. Not Both : user that didn't included in both class.

We labelled data based on justification for news curator in [1] with following steps:

Step 1: we filtered user that had followers up to 1000 that indicated news curators' class.
Step 2: we checked tweet activity from user manually. If user almost tweets about news article include the URL without interacting to the others (mention) we labeled it as news aggregator. If the tweet activity of user tweet news article also there was interacting to the others user, we labeled it as news curators.
Step 3: we labeled the rest user that didn't include in characteristic from step 1 and step 2 as not both class.

2.4. Classification

In this paper, we used supervised machine learning with probabilistic approach, Naïve Bayes Classifier (NBC) for the classification system of Indonesia news curator in twitter. The reason using NBC because the algorithm didn't included the relation between each features and the implementation got time complexity faster than other supervised algorithm [12, 13]. Another prior work argued NBC can be used in classification problem especially with data from twitter and produced higher accuracy than another classification algorithm [6, 7]. The flowchart of classification algorithm shown on Figure 3.

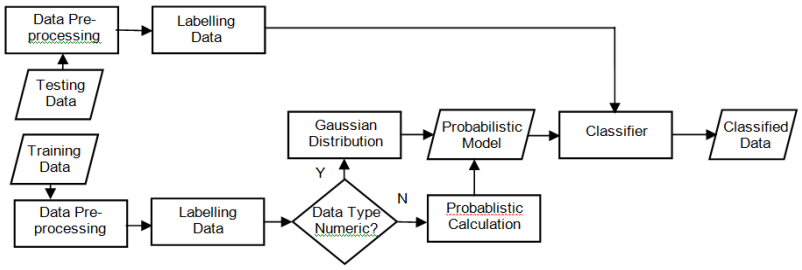

Figure 3. Flowchart of Naïve Bayes Classifier Algorithm

In NBC, the algorithm learns two attributes, hypothesis and evidence. Hypothesis is the supposition of a class and the evidence is the indication from the feature that influences the class. NBC finds the probability of conditional hypothesis based on the measurement of

probability. Dataset split into data training to make mathematics model and data testing use for classification process. For the numerical feature we used the Gaussian distribution and for the category feature we calculate the its probabilistic and the result would produce probabilistic model [7-9]. Classifier made using probabilistic model and classified the testing data that would produce classified class data. In addition, we implemented Feature Subset Selection (FSS) in classifier for trying all of feature combination [14].

2.5. Evaluating System

In this step, we evaluated the performance of system using confusion matrix that produce true positive, false negative, true positive and false negative. The output of confusion matrix such as accuracy, precision, and recall. Accuracy was number of correct classified divided to number of all data. Precision was known as positive predictive value. Precision value got from the result calculation of true positive divided by the sum of true positive and false positive. But recall value got from the result calculation of true positive divided by the sum of true positive and false negative. Then, we made the analysis from the performance of system based of accuracy, precision and recall.

3. Result

In this section, we explain our result based on research framework design that we built on section 3. This section consists of dataset, system implementation, analysis of system performance, and analysis of the features.

3.1. Dataset

Dataset that we used in this paper are user data and tweet data that we crawled on 13-17 April 2016. User data using Indonesian language amounting 300 users with visible features (followers, bio profile and location) that had been given the class. Details of the user data are 100 data with news curator class, 100 data with news aggregator class, and 100 data with not both class. Name in user data get from twitter.com manually, but its features crawling automatically.

Tweet data is tweet using Indonesian language with tweet activity features (URL, mention, re-tweet and common tweet) amounting 55.423 tweets. Tweet data were collected from crawling each user data using Application Programming Interface (API) twitter that had been implemented in PHP programming. Tweet data was stored in tweet corpus. Next, the data set, include user data and tweet data was stored in database using MySQL.

3.2. System Implementation

System implementation from research system design using PHP programming language built with visual code software, and using database from XAMPP server and MySQL. The sceenshot of interface for crlawing data shown on Figure 4.

Figure 4. Screenshot of Implementation System for Crawling Data

In training process, classification process produced probabilistic model for the classifier. In testing process, we inputted testing data to the classifier for getting classified class. For the instance, the result of testing process shown on Table 7.

Table 7. The Instance of Testing Data from the Result of Classification Process

Account	Probility of News Curator	Probability of News Aggregator	Probability of Not Both	Original Class	Classified Class	Desc.
@afrizan1	9.0E-10	0	2,11E-10	Not Both	Not Both	Correct
@belengjabar	7.0E-10	9.1E-9	0	News Aggregator	News Aggregator	Correct
@Berita1Sports	5.0E-10	4,71E-02	0	News Aggregator	News Aggregator	Correct
@BeritaCirebonID	0	2,17E-04	0	News Aggregator	News Aggregator	Correct
@Dhandy_Laksono	2.49E-8	0	0	News Curator	News Curator	Correct
@defapratama	2.0E-8	0	7.26E-8	Not Both	Not Both	Correct

Based on Table 7, the sample of data user, @Dhandy_Laksono account has the original class as news curator. Then, after the classification process, each class has each probabilistic value. In @ Dhandy_Laksono, we noticed the probabilistic value of news curator is 2,49E-8 (2,49 x 10^{-8}) , not both is 0 and news aggregator is 0. So, based on the result, predictive class for @Dhandy_Laksono is news curator, its correct prediction because it has similar class between original and predictive class.

4.3. Performance of System

We evaluated system using percentage split to get the performance of system. We built confusion matrix for measuring accuracy, precision and recall. In percentage split, we exanimated dataset proportion (training data: testing data) such as 90:10, 80:20, 70:30: 60:40 and 50:50 in our proposed framework and made analysis of the result. In additon, we implemented Feature Subset Selection (FSS) in classifier because we got the indication of over fitting, the condition when dataset produced probabilistic model with high standard variance is almost complex and its decrease the accuracy of system [14]. For the instance of overfitting indication, we noticed accuracy decreased from dataset proportion 70:30 to 80:20, from 90% to 88,33%, that shown on Figure 5.

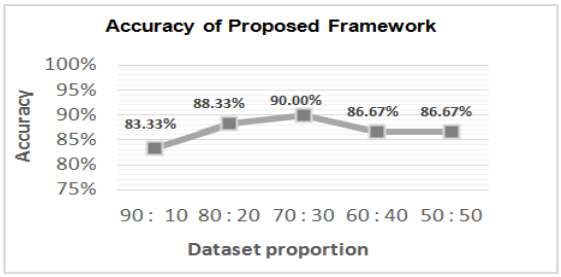

Figure 5. The instance of Accuracy in Each Dataset Porportion

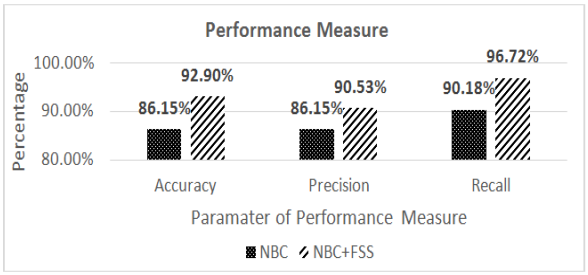

Figure 6. System Performance using Accuracy, Precision and Recall Parameters

Because of that indication, we exanimated dataset with two scenario examinations, first only using Naïve Bayes Classifier (NBC) for our proposed framework, second we added FSS after NBC process for the optimization. We calculated average of accuracy, precision and recall from all dataset proportion and produced the result that shown on Figure 6.

Based on Figure 6, we noticed that proposed framework using NBC plus FSS produced better performances than only using NBC. Accuracy in NBC+FSS higher 6,75% that only NBC, its mean FSS could be alternative ways for optimizing NBC. In the other parameter, precision and recall, proposed framework using NBC+FSS produced higher 4,13% and 6,54% that only using NBC. Its caused by FSS exanimated all the combination of features into the proposed framework and produced the optimal feature combination. There were 128 combinations exanimated from FSS, almost the result of each dataset proportion produced the optimal combination feature using followers and URL. We made analysis followers feature be optimal feature because its already defined in prior work [1], and for the URL features it because the characteristics of news curator and news aggregator in twitter is sharing tweet about news that contained URL of news article.

4.4. Analysis of Features

In this step, we calculated the probability of each features based on the examination result. The proposed of this step, we gave another perspective for labeling Indonesian News Curator in future work. The result shown on Figure 7.

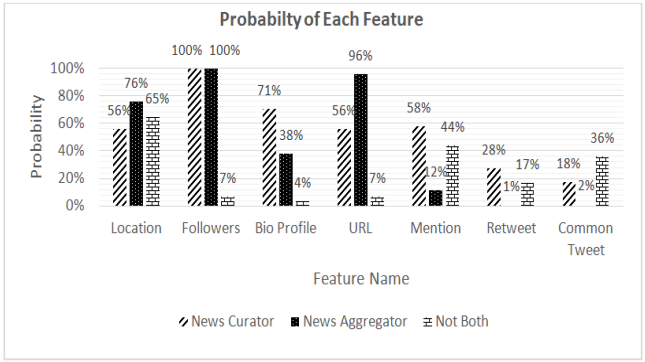

Figure 7. Probability of Each Feature for the Result in Each Class

Based on Figure 7, we know that location, followers, and URL features have the highest probability that influence to news aggregator class with the value 76%, 100% and 96%. While followers, bio profile, mention, and re-tweet features have the highest probability that influence to news curator class with the value 100%, 71%, 58% and 28%. In addition, common tweet has the highest probability that influence to not both classes with the value 36%.

6. Conclusion

After the implementation, examination and analysis of our proposed framework for classing Indonesian news curator in twitter, we made conclusion that Naïve Bayes Classifier (NBC) can be used in our porposed framework as classifier that produced 86,15% average of accuracy. For the optimization, we can used Feature Subset Selection (FSS) that exanimated all the feature combinations into the classifier and produced higher 6,75% accuracy than only using NBC. As the result of feature analysis, we concluded that the most influence feature for news curator class are followers, bio profile, mention and re-tweet. While for news aggregator class are followers, location and URL. And for the not both class is common tweet class. This feature could be used for labeling data Indonesia news curators in future work.

References

[1] J Lehman, C Castillo, M Lalmas, E Zuckerman. *Finding News Curators in Twitter*. WWW Workshop on Social News on the Web (SNOW). Rio de Janeiro, Brazil. 2013.
[2] N Diakopoulos. *Finding and Assesing Social Media Information Sources in the Context of Journalism*. CHI'12. Austin, Texas, U.S.A. 2012.
[3] J Lehman, C Castilo, M Lalmas, E Zuckerman. *Transient News Crowds in Social Media.* Seventh International AAAI Conference on Weblogs and Social Media (ICWSM). Cambridge, Massachusetts. 2013.
[4] T Pramiyati, I Supriana, A Purwarianti. Determining Trust Scope Attributes Using Goodness of Fit Test:A Survey. *TELKOMNIKA*. 2014; 13(2): 654-660.
[5] L Ruhwinaningsih, T Djatna. A Sentiment Knowledge Discovery Model in Twitter's TV Content Using Stochastic Gradient Descent Algorithm. *TELKOMNIKA*. 2014; 14(3): 1067-1076.
[6] SF Rodiyansyah. Klasifikasi Postingan Twitter Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive Bayessian Classification. Indonesian Computer, Electronics and Instrumentation Support Society (IndoCEISS). Yogyakarta: Universitas Gajah Mada. 2014.
[7] RS Pradana. Pengkategorian Pesan Singkat Berbahasa Indonesia pada Jejaring Sosial Twitter dengan Metode Klasifikasi Naive Bayes. *Repositori Jurnal Mahasiswa PTIIK UB*. 2013; 1(3).
[8] N Amalia. Penerapan Teknik Data Mining untuk Klasifikasi Ketepatan Waktu Lulus Mahasiswa Teknik Informatika Universitas Telkom Menggunakan Algoritma Naive Bayes Classifier. *Repositori Telkom Open Library.* 2015.
[9] H Frebruriyanti. Rancang Bangun Sistem Layanan Informasi Bencana Melalui Twitter Menggunakan Basis Data XML. *Unisbank Repository*. 2013.
[10] JE Sembodo, EB Setiawan, ZKA Baizal. *Data Crawling Otomatis pada Twitter.* Indonesian Symposium on Computing (Indo-SC). 2016: 11-16.
[11] J Han, M Kamber, J Pei. Data Mining Concepts and Techniques. Third Edition. San Fransisco: Morgan Kauffman Publishers. 2012.
[12] P Tan, M Steinbach, V Kumar. Introduction to Data Mining. Boston: Pearso Education. 2006.
[13] I Rish. *An Empirical Study of The Naive Bayes Classifier*. International Joint Conference on Artificial Intelligence. California. 2006.
[14] R Kohavi, GH John. Artificial Intelligence. *Wrapper for Feature Subset Selection.* 1996: 273-324.