Twitter Data Classification using Multinomial Naive Bayes for Tropical Diseases Mapping in Indonesia

Romy Ranovan, Afrizal Doewes, and Ristu Saptono
Department of Informatics, Universitas Sebelas Maret (UNS), Surakarta, Indonesia.
afrizal.doewes@staff.uns.ac.id

***Abstract*—Tropical diseases are diseases commonly found in tropical and sub-tropical regions. The goal of this research is to map the tropical diseases based on data from Twitter to help policymakers take essential steps regarding health condition in Indonesia. Tweets classification was conducted in two phases, both using Multinomial Naive Bayes. The first phase is to filter non-Indonesian tweets, and the second phase is to classify the tweets containing diseases information. The result shows the type of the diseases and location with high accuracy supported by map visualization.**

***Index Terms*—Classification; Mapping; Multinomial Naive Bayes; Tropical Diseases**.

I. INTRODUCTION

Tropical diseases are diseases commonly found in the tropical and sub-tropical region [1]. Examples of tropical diseases are malaria, dengue, avian flu, and HIV/AIDS. In this research, news of tropical disease outbreak in Indonesia were collected such as malaria, dengue, and avian flu as in [2], then HIV/AIDS was also added to the list. The news of tropical disease outbreak was collected from social media.

Nowadays, social media are not only limited to social communication between its users. The vast availability of data that consist of opinion and general text had popularized data-based approach among the researchers [3]. Twitter is a social media that allows its users to send and receive short messages in less than 140 characters and is considered as a micro-blogging. Ever since the launch, Twitter had a massive role in news outspread.

Data was collected using Twitter API with the keywords of tropical diseases in Bahasa Indonesia. The data were put into language identifier to differentiate between tweets in Bahasa Indonesia and tweets in other languages. Then, only the tweets in Bahasa Indonesia were classified to find which tweets are considered as diseases outbreak news. Multinomial Naive Bayes was implemented in both classification phases.

In this research, two approaches were employed in addition to Multinomial Naive Bayes (MNB) method; using feature selection method and without feature selection method. Both of this approach was applied in the language identifier and diseases classification. Since the data set ratio is imbalanced with one class having way more data than the other class, oversampling was implemented to the data. The results of the original data were then compared with data that had been over-sampled. There are four aspects a tweet must have to be classified as news of disease outbreak; the name of the disease, the outbreak location, the condition of the victims (dead or still in medical treatment), and optionally the number of cases. After the information is extracted, the result is then visualized in the form of a map.

II. RELATED WORKS

News mining using social networks has attracted researchers due to its openness and easiness to access its data. The multi-class classification research [4] shows that Multinomial Naive Bayes method has the best result out of 4 different methods implemented and according to [5] Naive Bayes has better performance than Support Vector Machine on analysis of micro-blog with short text. On research presented in [6] the authors used the data from Twitter to track public sentiment about the swine flu in the U.S., the analysis shows that Twitter can be used as a measure of public opinion about health-related events.

III. MULTINOMIAL NAIVE BAYES

Multinomial Naive Bayes assign the most likely class to the document example by looking at its feature vector [7], the algorithm is as follows:

$$P(c \mid d) = \frac{P(d \mid c).P(c)}{P(d)} \tag{1}$$

with d as document and c as class. The classification algorithm is as follows:

$$C_{map} = \arg\max_{c=C} P(c \mid d) \tag{2}$$

$$C_{map} = \arg\max_{c=C} \frac{P(d \mid c).P(c)}{P(d)} \tag{3}$$

$$C_{map} = \arg\max_{c=C} P(d \mid c).P(c) \tag{4}$$

with C_{map} is the class with the highest probability. P(d) is removed because the value is constant to all the classes. Therefore, it will not influence the result.

The probability of class algorithm can be broken down into:

$$C_{\max} = \arg\max_{c \in C} P(x_1, x_2, ..., x_n \mid c).P(c) \tag{5}$$

$$P(x_1, ..., x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet ... \bullet P(x_n \mid c) \tag{6}$$

where x_n is the extracted features of a document. This algorithm has a weakness, if a document's feature is not found in any of the class, it will return P(c|d) value as zero. To solve this, Laplace smoothing is applied, the algorithm becomes:

$$P(w_i \mid c) = \frac{count(w_i, c) + 1}{(\sum_{w \in V} count(w, c)) + |V|} \quad (7)$$

where w is a feature of class c and V is vocabulary or the number of unique features of a class.

IV. METHODOLOGY

Stages of this research are shown in Figure 1. First, the data is collected using the Twitter REST API then the data are manually labelled into 2 of 4 classes that are class "Bahasa Indonesia" or "not Bahasa Indonesia" and class "Disease" or "not Disease". The data were then preprocessed before applied to the classifier. Two classifications were performed in this study, one for language identification and another for tropical diseases outbreak news classification. After the classification, the result was evaluated to determine the best outcome. Finally, the result was visualized in the form of a map.

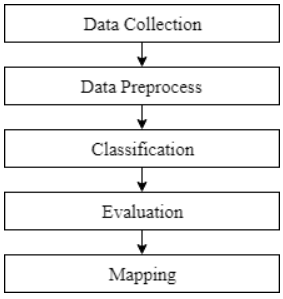

Figure 1: Research Methodology

V. EXPERIMENTAL RESULT

A. Collecting Data

The data was gathered using Twitter REST API from 23 May 2016 to 22 November 2016. The information taken from the metadata consists of "tweet", "date_created", and "username". The data with the total number of 33613 tweets were stored in JSON files.

B. Preprocessing

Preprocessing is a process where raw data is modified into a more structured and more compatible with the next processes [8]. Tweets contain not only plain text. During this process, emoji character, URL, and non-alphanumeric symbols were removed.

There are three steps for preprocessing: case folding, tokenization, and elimination. Case folding is an action to remove the parts of the tweet that is irrelevant for the classification, this includes:

- *mention* (the use of symbol '@' in front of a string usually a username, this username is also removed),
- *URL* (example: https://t.co/rK1EFniTIU),
- String 'RT' (this string shows that the tweet is a re-tweet from another tweet),
- Symbols other than alphanumerical (this includes emoji).

Afterwards, the tweets are tokenized. The elimination was applied to remove duplicate tweets; these duplicates are usually tweets that are re-tweeted which means they have the same features. So if several tweets have the same array of features, only one was kept. An example of the preprocessing results is shown in Table 1.

Table 1
Preprocessing Result

Tweet	Case Folding	Tokenization
Waspada, Sudah 10 orang Warga pekanbaru meninggal akibat DBD https://t.co/lHjwnxbfNT https://t.co/Xnal0j18O8	waspada sudah 10 orang warga pekanbaru meninggal akibat dbd	"waspada", "sudah", "10", "orang", "warga", "pekanbaru", "meninggal", "akibat", "dbd"

C. Classification

Two Classification processes were performed in this research, first is Language Identification and the second is Tropical Diseases Outbreak News Classification. For both classifications, the data was divided into training set and testing set with the ratio of 80:20.

Oversampling approach was applied to the data. Oversampling is a way to deal with imbalanced data set [9]. In oversampling all original data is used and then new data is added, this new data is a result of the replication of minority class data; it is done until the desired data ratio is achieved [10].

For the classification two different approaches were used, both approaches implemented Multinomial Naive Bayes method, but the difference is that the first approach was without feature selection while the second one was with feature selection. The method for feature selection is TF thresholding. Since the data used in this research is a micro-blog, the repetition of a single feature in a data is rarely found. TF thresholding was used in feature selection, it is done by comparing the feature frequency on each class, if the difference is below the threshold then the feature will be ignored or erased on both class and will not be used in the classification. The algorithm to calculate the feature selection is the absolute value from subtraction between the frequency of feature N of class a and the frequency of feature N of class b, divided by the maximum frequency of both class:

$$FS_x = \frac{|f_{xa} - f_{xb}|}{f_{\max}} \quad (8)$$

The details of each classification are as follows:

Language Identification - Language Identification is performed to filter twitter data with Bahasa Indonesia from data with non-Bahasa Indonesia. In this process, the data was labelled into two class: Bahasa Indonesia and Not Bahasa Indonesia. From the labelling, 2312 data are Bahasa Indonesia, and 17284 data are not Bahasa Indonesia. From the classification result, as shown in table 4, the model achieved high accuracy in all approaches.

Tropical Diseases Outbreak News Classification - In this process, the data was labelled into two class: Disease and Not

Disease. Disease class means the data from this class contain news about the tropical disease outbreak, while the data from Not Disease class do not. From the labelling, 219 data are Disease, and 2093 data are Not Disease. Some tweet examples from both classes are shown in table 2. Aside from the approaches stated above, this process also applied 5-Fold Cross Validation to compensate the small dataset.

In Cross Validation, also often called rotation estimation, the data is split into two groups, training set and test set [11]. The data in training set will be used to train the model which is used to predicts the data in the test set. This research use k-fold cross-validation in which the data is split into k-parts evenly, in this case, 5-parts. Then the model is trained and tested k-times with each of the parts would become test set exactly once and included in the training set k-1 times.

In this process, feature selection was applied in each fold. As a result, the features used in classification in each fold might be different. To address this problem, two approaches were used; the first one is to use the union of features from all folds, the second is to use the intersection of features, which mean only features appeared on all folds were used.

From the classification result, as shown in Table 5, the model achieved high accuracy in all approaches. The result from both classifications shows decreasing accuracy at the approach with feature selection compared to the ones without feature selection, except for the approach with 1:1 ratio over-sampled data set, which in this case the result for the ones with feature selection is slightly better.

Table 2
Tweet Labeling for Disease Class

Original Tweet	Class
Cegah Angka DBD, Tim Kesehatan Fogging Kuta Lhoksukon - Waspada Online https://t.co/4pDp1Nqt0E #Kuta (To Prevent Dengue Fever, Health Team to Fog Kuta Lhoksukon - Waspada Online https://t.co/4pDp1Nqt0E #Kuta)	Not Disease
Dinkes Tanjabbar Temukan 16 Kasus HIV/AIDS https://t.co/xlWzcCW59L https://t.co/WkTK1WB9td (Tanjabbar Public Health Office found 16 cases of HIV/AIDS https://t.co/xlWzcCW59L https://t.co/WkTK1WB9td)	Disease
Satu Keluarga Terserang DBD, Puskesmas Cijeungjing Fogging Bojongsari Ciamis https://t.co/DmMN5FuWVI. (One Family Fell Ill to Dengue Fever, Cijeunjing Health Service to Fog Bojongsari Ciamis https://t.co/DmMN5FuWVI.)	Disease
@HANAbear_ lagi musim dbd lagi ya..… (@HANAbear_ dengue fever season again huh….)	Not Disease

D. Evaluation

The results were evaluated to measures the performance of Multinomial Naive Bayes method used in this research. Confusion matrix is used to calculate accuracy, precision, and recall. Confusion matrix, also called error matrix, is a table used to visualize the performance of an algorithm [12] or, in this study, the performance of the classifier. The resulting confusion matrix could look like the format in Table 3.

Table 3
Confusion Matrix

Confusion Matrix		Prediction	
		TRUE	FALSE
Actual	TRUE	TP	FN
	FALSE	FP	TN

Accuracy, precision, and recall are calculated using the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (9)$$

$$\text{Pr}ecision = \frac{TP}{TP + FP} \quad (10)$$

$$\text{Re}call = \frac{TP}{TP + FN} \quad (11)$$

In this research, three datasets were used: (1) data without oversampling, (2) over-sampled data with 1:2 ratio (the minority class is over-sampled to 50% of the number of the majority class), and (3) over-sampled data with 1:1 ratio (the minority class is over-sampled to the same number of the majority class). Three different feature selection thresholds were used: 0.1, 0.3, and 0.5. The best result means the approach still has high accuracy despite having its feature reduced the most after feature selection.

In language identification, as shown in table 4, for the approaches without feature selection, the results are 87.26% accuracy, 48.05% precision, and 98.7% recall. Overall, the approach with the best accuracy and precision is the one with 1:1 ratio over-sampled data and thresholds 0.3. Best recall value was achieved in 1:1 ratio over-sampled data, without feature selection.

In tropical diseases outbreak news classification, as shown in table 5, for approach without feature selection, the results are 93.68% accuracy, 62.9% precision, and 98.63% recall. There is no result difference between approaches with the union of features and approaches without feature selection. It shows that all features are selected at the end of all k-fold. Overall, two approaches got the best result in accuracy and precision despite having the least number of features. Those two approaches are the ones with 1:1 ratio over-sampled data, using 0.3 and 0.5 thresholds. Both of them had 1478 of their features eliminated.

E. Mapping

The result of the classification was presented in the form of a map as shown in figure 2. The data about the news outbreak were grouped with the corresponding cities, with all disease combined. The circles in the map show the amount of the tweets, the bigger the circle indicates more tweets about the tropical diseases outbreak in particular locations.

VI. CONCLUSION

In this paper, a classification model using Multinomial Naive Bayes method was performed to classify news about tropical disease outbreak from Twitter data. Also, oversampling was applied to the data and TF thresholding was used as feature selection method. The model was broken down into two parts: language identification and Diseases or Not-Diseases classification. Language identification is used to pick only data in Bahasa Indonesia. In tropical diseases classification, 5-fold cross-validation was applied. This study was able to map the tropical diseases in Indonesia based on data from Twitter with high accuracy.

Table 4
Language Identification Result

Approach	Class Ratio (minority to majority)	Feature Selection Threshold	Features Before Feature Selection	Features After Feature Selection	Accuracy	Precision	Recall
MNB	None (1 to 7.48)	-	44150	44150	87.26%	48.05%	98.70%
	OS (1 to 2)	-	44894	44894	92.73%	82.12%	99.94%
	OS (1 to 1)	-	44894	44894	98.28%	96.67%	100.00%
MNB and Feature Selection	None (1 to 7.48)	0.1	44150	42322	84.00%	42.34%	98.70%
		0.3	44150	42308	82.13%	39.65%	98.70%
		0.5	44150	42308	82.13%	39.65%	98.70%
	OS (1 to 2)	0.1	44894	42868	91.99%	80.66%	99.94%
		0.3	44894	42850	91.67%	80.03%	99.94%
		0.5	44894	42850	91.67%	80.03%	99.94%
	OS (1 to 1)	0.1	44894	42876	98.73%	97.62%	99.88%
		0.3	44894	42852	98.80%	97.79%	99.86%
		0.5	44894	42852	98.77%	97.73%	99.86%

Table 5
Tropical Diseases Outbreak News Classification Result

Approach	Cross-Validation	Class Ratio (minority to majority)	Feature Selection Threshold	Features Before Feature Selection	Features After Feature Selection	Accuracy	Precision	Recall
MNB	-	None (1 to 9.56)	-	6069	6069	93.68%	62.90%	98.63%
		OS (1 to 2)	-	6069	6069	95.85%	89.34%	99.52%
		OS (1 to 1)	-	6069	6069	97.18%	95.43%	99.09%
MNB and Feature Selection	Union	None (1 to 9.56)	0.1	6069	6069	93.68%	62.90%	98.63%
			0.3	6069	6069	93.68%	62.90%	98.63%
			0.5	6069	6069	93.68%	62.90%	98.63%
		OS (1 to 2)	0.1	6069	6069	95.85%	89.34%	99.52%
			0.3	6069	6069	95.85%	89.34%	99.52%
			0.5	6069	6069	95.85%	89.34%	99.52%
		OS (1 to 1)	0.1	6069	6069	97.18%	95.43%	99.09%
			0.3	6069	6069	97.18%	95.43%	99.09%
			0.5	6069	6069	97.18%	95.43%	99.09%
	Interception	None (1 to 9.56)	0.1	6069	4603	89.83%	51.23%	99.09%
			0.3	6069	4592	89.78%	51.30%	99.09%
			0.5	6069	4591	89.78%	51.30%	99.09%
		OS (1 to 2)	0.1	6069	4595	95.76%	89.10%	99.52%
			0.3	6069	4591	95.76%	89.10%	99.52%
			0.5	6069	4591	95.76%	89.10%	99.52%
		OS (1 to 1)	0.1	6069	4593	97.37%	95.78%	99.09%
			0.3	6069	4591	97.37%	95.78%	99.09%
			0.5	6069	4591	97.37%	95.78%	99.09%

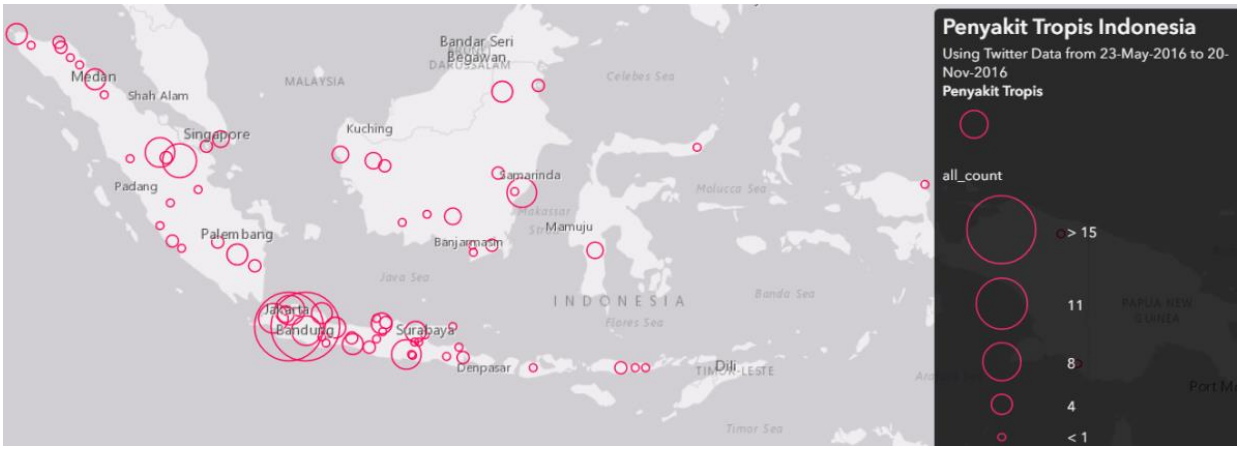

Figure 2: Map Result

REFERENCES

[1] J. Farrar, P. Hotez, J., T. Junghanss, G. Kang, D. Lalloo, and N. White. *Manson's tropical diseases*. Elsevier Health Sciences, 2013.

[2] F. Wulandini and A. S. Nugroho,. "Text Classification Using Support Vector Machine for Webmining Based Spatio Temporal Analysis of the Spread of Tropical Disease, in *International Conference on Rural Information and Communication Technology*, 2009.

[3] B. Pang, and L Lee, "Opinion Mining and Sentiment Analysis". *Foundations and Trends in Information Retrieval*, vol. 2, pp. 1-135, 2008.

[4] V. Prieto, S. Matos, M. Alvarez, F. Chaceda, and J. L. Oliveira, "Twitter : A Good Place to Detect Health Condition". *PLoS ONE*. vol. 9, num. 1, pp. e86191, 2014.

[5] A. Bermingham, and A. Smeaton, "Classifying Sentiment in Microblog : Is Brevity an Advantage?". *19th ACM International Conference on Information and Knowledge Management*, 2010.

[6] S. Allesio et al. "The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic". *PLoS ONE 6(5): e19467, doi:10.1371/journal.pone.0019467*, 2010.

[7] I. Rish, "An Empirical Study of the Naive Bayes Classifier", in *International Joint Conference on Artificial Intelligence, 2001*.

[8] R. V. Imbar, A. M. Adelia, and A. Rehatta, "Implementasi Cosine Similarity dan Algoritma Smith-Waterman untuk Mendeteksi Kemiripan Teks". *Jurnal Informatika*, vol. 10 no. 1, Juni 2014.

[9] N. V. Chawla, K.W. Bowyer, L. O. Hall, and W.P Kegelmeyer, "SMOTE : Synthetic Minority Over-sampling Technique". *Journal of Artificial Intelligence*, vol. 16, pp. 321-357, 2002.

[10] R. Dubey, J.Y. Zhou, Y. Wang, P.M. Thompson, and J.P. Ye. "Analysis of Sampling Technique for Imbalance Data: An N=648 ADNI Study". *Neuroimage*, vol. 87. pp. 220-241, 2014.

[11] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", in *International Joint Conference of Artificial Intelligence*, 1995.

[12] D. M.W. Powers, "Evaluation : From Precision, Recall, and F-Factor to ROC, Informedness, Markedness, and Correlation". *Technical Report SIE-07-001*, 2007.