Journal of Applied Research and Technology

www.jart.icat.unam.mx

Journal of Applied Research and Technology 18 (2020) 144-153

Original

Disease surveillance in Indonesia through Twitter posts

Mirna Adriani • Fatimah Azzahro* • Achmad Nizar Hidayanto

Defence Faculty of Computer Science, Universitas Indonesia, Depok, 16424, Indonesia

Received 03 26 2020; accepted 05 11 2020
Available online 06 30 2020

Abstract: Social media data has become popular resources for various research topic such as public health. One of the popular research directions is to use social media data to detect if there is an epidemic disease emerging in a certain area. This paper presents a framework for mapping the emergence of disease in Indonesia using data from Twitter. The framework is built upon several methods which consist of classification using SVM, clustering using K-Means, and a named-entity recognizer to extract location names. Our research successfully identifies tweets indicating disease emergence and generates a relatively accurate map visualization. Thus, we believe that using Twitter may help Indonesia government officials to get an overview of the spread of disease in Indonesia in a relatively short time.

Keywords: Twitter, disease map, Indonesia, disease monitoring, classification, clustering

*Corresponding author.
E-mail address: Azzahro.fatimah@cs.ui.ac.id(Fatimah Azzahro).
Peer Review under the responsibility of Universidad Nacional Autónoma de México.

1. Introduction

Social media recently has become a popular channel for the public to share and gain information easily and quickly, including in Indonesia. One of the most widely used social media in Indonesia is Twitter. According to We are Social, Indonesia has 150 million active social media users, and 52% of them routinely use Twitter (We are Social, 2019). Twitter works as citizens' popular source of information because most of the information is not only posted by the news media company, but also by the citizens themselves. Although in the latter case, one should be carefully filtered and differentiate useful information from the uninformative one.

The recent trends have opened a new perspective for researcher where the use of information from social media can help various fields such as crime detection, disaster management, and public health. One of the popular research in the public health sector is the study of monitoring and predicting disease outbreak. In Indonesia, the conventional procedures for monitoring disease outbreak are carried out by collecting data manually from various hospitals, clinics and health centres that are spread throughout Indonesia (Kementerian Kesehatan RI, 2013). This manual monitoring system, although producing more accurate data, is costly and time-consuming. As an archipelago country, this task is awfully challenging considering many health facilities located in the different island and difficult to reach. Thus, a faster disease outbreak monitoring system is essential to be developed so that disease outbreaks can be identified and resolved as early as possible.

Previous studies have used Twitter data to investigate disease spread around the world. Freifeld et al. conducted research to develop a health surveillance system called HealthMap which collects reports of epidemic disease from several sources such as Google News, ProMED Mail, and official reports (Freifeld, Mandl, Reis, & Brownstein, 2008). Using 778 reports, the research aims for a broader and more global health surveillance system and not only limited to a certain country. The research uses 87 types of disease in 89 countries and achieves 84% accuracy for its automatic classification model. Although this research has included Indonesia data, the location accuracy and the completeness of disease that specific to Indonesia can be improved.

This paper presents a study of using Twitter data to develop Indonesia disease map. We focus on identifying five diseases that commonly occurred in Indonesia, namely dengue fever (DHF), diarrhoea, filariasis (elephantiasis), upper respiratory tract infections (URTI), and malaria. These five diseases are chosen because ofdue to the number of tweets available on the Internet. Tweets related to other diseases are available, but the number is not as high as the aforementioned diseases. Moreover, there are several past researches in Indonesia that focuses on iInfluenza-ls Like iIllness. As an attempt to fill the research gaps in other types of diseases, this paper proposes an alternative approach and apply it to the five most common diseases in Indonesia. Additionally, we develop a Location Dictionary NER system as an alternative to identify a location from Twitter posts. The final result of this research is the map visualization of disease spread in Indonesia. The research questions we address are, therefore:

Several efforts that more focus on Indonesia context has been made by Zulfa (2015) and Ramadona, Tozan, Lazuardi, and Rocklöv (2019). In her study, Zulfa used Indonesian tweets to visualize ILI (Influenza Like Illness) spread in Indonesia. The research applies two steps of classification on tweets that potentially contain information regarding ILI symptoms. The classifier used k-nearest neighbours' method to classify tweets that indicate influenza-like illness symptoms. Meanwhile, Ramadona et al., investigated the potential of using Twitter's geotag to predict the spread of dengue disease in Yogyakarta, one of the big cities in Indonesia. Although these efforts are valuable contributions to the fields of public health in Indonesia, the use of geo-tag may suffer from two major biases. First, while Twitter provides automatic location detection, some users prefer to set their geo-tag location manually. Some users may set their fake location to a completely unrelated location for the sake of making jokes or looking cool (Mourad, Scholer, Sanderson, & Magdy, 2018). Hence, the accuracy of geo-tag from Twitter might be compromised. Secondly, not all users want to leave digital traces on the Internet. One might avoid using geo-tag location to maintain their privacy. According to Han et al., less than 1% of tweets used geo-tag (Han, Cook, & Baldwin, 2014). Thus, an alternative method to extract location from Twitter ata should be investigated.

1) How to identify the type of diseases from a tweet?

2) How to visualize the disease spread without using the geo-tag feature on Twitter?

The rest of the paper will be explained as follow. The next section will describe the related works of this research (Section 2). Next, we will discuss the methods that are used to build our framework (Section 3). Then, the paper will discuss the results of our experiment to evaluate our framework performance (Section 4). We then provide our conclusion and future works regarding the framework that has been built (Section 5).

2. Related studies

Several studies have utilized social media data to monitor disease spread around the world. The studies are mostly focused on influenza and influenza-like illness (Nagar et al., 2014). Although, other studies have conducted research for other types of diseases such as H1N1 virus (Signorini, Segre, &

Polgreen, 2011) and cholera (Chunara, Andrews, & Brownstein, 2012). Additionally, A study conducted by Cullota proposed several methods to identify influenza-like illnesses in a population using regression models for his classification approach (Culotta, 2010). He compared his proposed model with the report published by the Centers for Disease Control and Prevention (CDC) and achieved a correlation of .78 with CDC's statistics.

A similar study has been conducted in Indonesia that use Twitter posts to monitor ILI (influenza like illness). The research successfully built a classifier model using the k-nearest neighbours' approach to identify tweets that have indications of influenza-like illness. Zulfa built two classification model and achieved 92.7% accuracy for the first classification model, while the second model acquired 91.7% accuracy (Zulfa, 2015). Meanwhile, research published by Ranovan, Doewes, and Saptono (2018), examined the use of Twitter data to map the tropical diseases in Indonesia. The research used two phases of classification, both utilizing the Multinomial Naive Bayes method. The first step is focusing on separating tweet posts in Bahasa Indonesia and other languages. Next, Ranovan et al. used the proposed classifier to identify tweets containing disease information. Although this study produced an interesting result, it only demonstrated how to classify tweets that may indicate disease occurrence without attempting to identify the disease type.

In the area of text mining, research on named-entity recognition (NER) in Bahasa Indonesia has been conducted by Wahyudi and Budi (2004). The research uses a rule-based approach based on contextual, lexical, and morphological information to develop the InNER (Indonesian Named Entity Recognizer) system. The study produced a system with a recall of 63.43% and a precision of 71.84%. Another research on NERs in the form of Indonesian tweets was also completed by Taufik (2015) using a statistical approach with CRF (conditional random field) method. The results of the study reached the best F-measure of 59.36%.

3. Methods

The purpose of this study is to provide the visualization of disease outbreak in Indonesia by using Twitter data. In this study, we limited our scope by only identifying 5 common diseases in Indonesia, namely, dengue fever, diarrhoea, filariasis, upper respiratory tract infection (URTI), and malaria. To achieve our purpose, this study proposed a series of methods that consist of 7 steps. The general view of our proposed framework is shown in Figure 1. The following subsections will explain the proposed framework; starting with the process of data collection, manual labelling, classification, clustering, location entity recognition and ending with the disease map visualization.

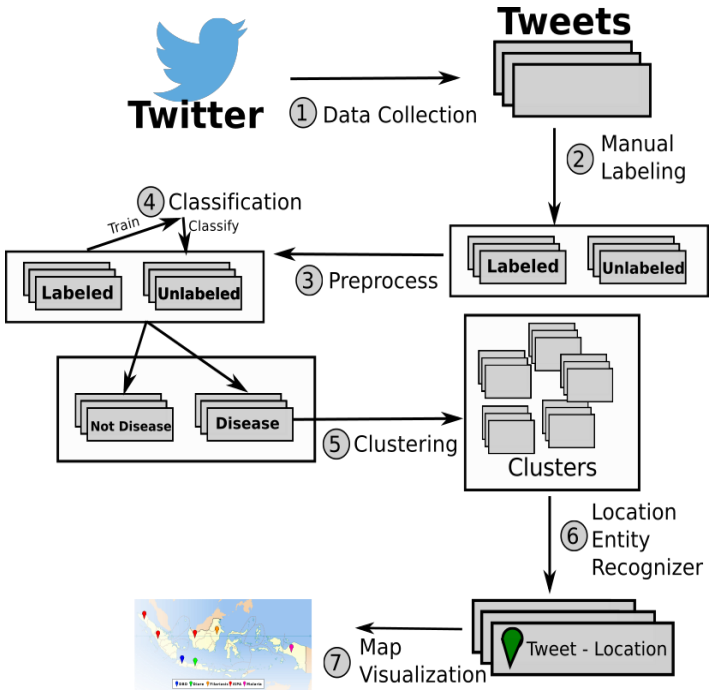

Figure 1. The proposed framework.

3.1. Data collection

We collected 21,227 tweets by using query keywords related to the top five common diseases in Indonesia. The queries are created based on information related to diseases that are found in trusted medical websites and news sites. We consider both the medical and common name of the disease, its synonyms, along with symptoms that are generally associated with the disease. The keywords used for querying the tweets for each disease are shown in Table 1.

Table 1. Query for collecting tweets.

Disease	Query Keywords	Tweet Posts
Dengue	'demam berdarah', dbd, dengue, 'bintik-bintik merah'	2,562
Diarrhea	diare, mencret, 'buang-buang air', muntah	3,394
Filariasis	'kaki gajah', filariasis	2,367
URTI	ispa, 'infeksi saluran pernapasan'	12,355
Malaria	malaria	549

3.2. Manual labelling

After collecting the tweets, we need to determine relevant tweets either by manual labelling or training. In this study, labelling is performed manually by randomly selecting a sample of 400 tweets for each type of disease. Since we use the top five common disease in Indonesia, we have a total of 2,000 manually labelled tweets. For each tweet, we assign three types of labels as follows. First, we assign the 'Disease' and 'Not Disease' label. This label is used to indicate whether a tweet contains information about a disease or not. Next, A disease type label is used to identify whether a tweet contains information about the top five diseases in Indonesia, namely 'dengue fever', 'diarrhoea', 'filariasis', 'URTI', and 'malaria'. Lastly, the Location label is assigned to tweets that contain information about the location where the disease has occurred. A location is considered valid if it contains a valid location name in Indonesia, including island name, province, city, or regency in Indonesia. Both the type of disease label and the location label is only assigned if a tweet is labelled as a 'disease' tweet.

3.3. Data pre-processing

Data pre-processing is an essential step need to be completed before processing the data. A good data pre-processing will result in reducing data inconsistency and improving result quality. In this research, we use five steps of the pre-processing process as explained below. Some tweets may contain common Twitter entities which are not recognized in Bahasa Indonesia, such as URL and mention. Therefore, we need to exclude these features from tweets. Additionally, we use stemming rules for Bahasa Indonesia from Naradhipa and Purwarianti (2012) to help converting informal words into its formal form. Meanwhile, the stop words list in Bahasa Indonesia is adopted from Tala, 2003.

1) Removing punctuation and special characters.
2) Case folding, converting uppercase to lowercase letters.
3) Eliminating twitter special feature: mention and URL.
4) Stemming.
5) Eliminating stop words.

3.4. Classification

The pre-processed data are then classified by using the linear Support Vector Machine (SVM). SVM was chosen as the classification method in this study because it has already empirically proven to have a better result in text classification task compare to other methods (Joachims, 1998). The classification process is performed automatically to identify tweets that indicate disease occurrence and separate it with irrelevant tweets. To use SVM, we need to transform the dataset into their Vector Space Model (VSM) representation. Then, we built our classifier model using SVM on the training dataset consisting of 2,000 tweets that had been manually assigned with 'Disease' and 'Not Disease' label in the prior steps. This training data consist of 703 tweets that are labelled as 'Disease' and 1,297 tweets that are labelled as 'Not Disease'. We used the Python implementation of Linear SVM from scikit-learn library to build our classifier. The results of the classification process are then evaluated using 5-fold cross-validation method and F1 score as our main evaluation metrics.

Using the 2,000-training dataset, we assign labels for the remaining 19,282 unlabelled tweets automatically. The iterative steps of the classification model are described below.

1) The initial classifier model that had been built are used to assign labels to 100 tweets which are randomly sampled from all the unlabelled tweets.
2) Based on these 100 newly labelled tweets, we select 10 tweets with the highest confidence label. The confidence measure is the absolute value of the decision function value of the SVM model for each tweet. In the case of SVM, these high confidence samples are the ones whose distance are the furthest from the hyperplane. Thus, among 10 tweets, there will be 5 tweets which are given the 'Disease' label and 5 tweets which are given the 'Not Disease' label by the model.
3) These 10 tweets then are added to the labelled tweets.
4) The classifier model is retrained using the recently added training dataset with the new 10 labelled tweets.
5) Step 1 to step 4 are continually repeated until all the dataset have been labelled.

After applying the iterative label assignment process, based on 19,282 unlabelled tweets, there are 12,852 tweets are labelled as 'Disease' tweets and 6,430 tweets are labelled as 'Not Disease' tweets.

3.5. Clustering

The 12,852 'Disease' labelled tweets from the previous step are then used to perform the clustering process. The objective of the clustering process is to group tweets based on the type of disease. To achieve the objective, we apply k-means clustering method with 5 as the value of k which defines the number of resulted clusters. K-means is a simple yet fast and robust clustering algorithm in producing reliable results in a distinct or non-explicitly labelled dataset (Shi, Liu, & Guan, 2010). Additionally, we used the implementation of the k-means algorithm from the scikit-learn library in the clustering process. For the clustering process, we use data from the manual labelling step that have been assigned to the disease type label. At the beginning of the k-means algorithm, 5 tweets will be chosen randomly as the initial centroids. This random selection causes the results of the cluster will be different each time the algorithm is run. Thus, to fix this issue, the clustering experiments will be carried out 5 times.

Two types of evaluation criteria are utilized to evaluate the result of the clustering process. First, we use internal criteria from the structure of the resulting cluster. We use the within cluster sum of squares (WCSS) as our evaluation metrics. Next, we use external criteria where we measure how well the clustering matches our manually assigned labels. To do so, we compare the cluster results with the manually assigned labels and measure the agreement between them using the adjusted Rand index (ARI). Several methods can be used to evaluate the accuracy of statistical learning algorithms on k-means clustering, including WCSS, BCSS, Elbow method, and ARI. However, WCSS and ARI are used in this research because of its simple yet effective method to measure the accuracy of K-means both from internal and external criteria.

3.6. Location entity recognizer

The next step is to identify the location where diseases have occurred. This step is essential to support the next step which is developing map visualization of disease spread. To complete the task, we developed a named-entity recognition (NER) that focuses on the location name in Indonesia, including island, province, city, or regency name. The NER is built based on three major basic modules which are location names dictionary, tokenization module, and dictionary lookup module. The basic overview of the method is shown in Figure 2.

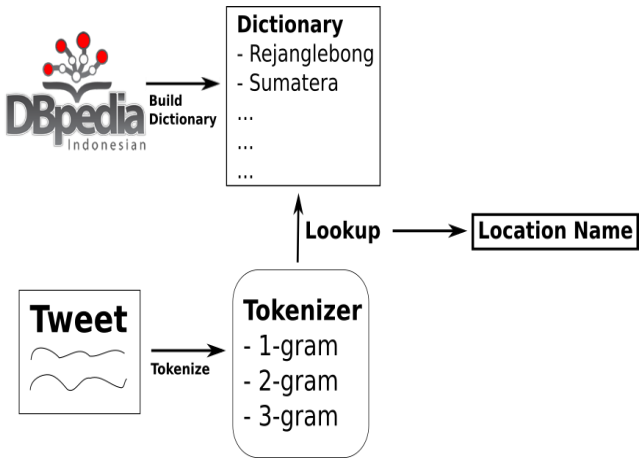

Figure 2. Location entity recognizer.

3.6.1. Building location names dictionary

We built the location names dictionary using data from the Wikipedia page which lists names of province, regencies and cities in Indonesia. Aside from the official name of each location entity, we also included the alias names and geographic coordinates that we queried from DBPedia Indonesia[3]. The alias names were extracted from the value of 'is dbpedia-own:wikiPageredirects of' property from the DBPedia entity. Meanwhile, the geographic coordinates were extracted from the geo:lat and geo:long property. Table 2 lists some example of regency names with their alias names and geographic coordinates.

3.6.2. N-Gram tokenization

Many of the location names in Indonesia consist of more than 1 word, thus we need to apply n-gram tokenization to recognize location names. Thus, in this research, we extract unigrams (one word), bigrams (2 words), and trigram (3 words) from each tweet's text. These n-grams or tokens will be used for performing dictionary lookup on the location names dictionary.

3.6.2. N-Gram tokenization

The main part of our NER system is a method to associate location names found on the text to location names in the dictionary. The method that we used is adapted from Norvig's project to build a simple spelling corrector (Norvig, 2009). The algorithm basically uses the concept of edit distance to give word suggestion based on input that contains incorrect spelling.

This spelling correction method was adapted in this study by examining the name of a location that is not exactly matched

Table 2. Regencies name with alias and geographic coordinates.

Regency	Alias	lat	long
Kabupaten Aceh Barat	Aceh Barat	4.45	96.1833
Kota Bukittinggi	Bukittinggi Bukit Tinggi	-0.531389	100.466942

with the dictionary and have a small edit distance to the word in the dictionary. To illustrate the dictionary lookup step, we use the example of input 'Rejanglebong'. The actual string stored in the dictionary is 'rejang lebong'. We ignore case sensitivity to reduce the complexity of the method and also by considering that tweets in Indonesian rarely apply a proper rule to capitalize the first letter of a location name.

1) After getting the input of 'Rejanglebong', the method will list words which could possibly be generated from 'Rejanglebong'. To do so, the method will edit one character from the word 'rejanglebong'. This list, for example, will consist of 'rejanglaebong', 'rejanglebofng', 'rejanglybong' and so on.

2) Next, the method list words which could possibly be generated by editing two characters from the word 'rejanglebong'. This list of words can be obtained by searching all words with which could possibly be generated by editing one character from each word from the word list that was generated in the first step.

3) Finally, we look up each word from both lists in the dictionary and assume that the first match (e.g. 'Rejang Lebong') as the correction for 'rejanglebong'.

The proposed NER system is evaluated by measuring the accuracy when applied to our 2,000 tweets which were previously labelled. We compare the detected location names with the location label that we assigned manually on the manual labelling process. The accuracy is calculated from the percentage of tweets' locations that are correctly recognized by our NER system. The result of our evaluation will be compared with the performance of an already available NER system called Polyglot-NER. Polyglot-NER is a named-entity recognition system that can be used on 40 languages, including Bahasa Indonesia (Al-Rfou, Kulkarni, Perozzi, & Skiena, 2015).

3.7. Visualization

The final step of our framework is to generate a map visualization from tweets that indicate disease spread in a certain location. Tweets to be mapped must be associated with their geographic coordinates which consist of latitude and longitude. We extracted the coordinate from DBPedia Indonesia as explained previously. In the case of no location names were recognized from certain tweets, we tried to apply the following rules to associate the tweet with their geographic coordinates.

1) If a tweet has a location name that is identified by the NER, then the coordinate would be looked up from the dictionary.
2) If there is no location name that is recognized by the NER, we extract the geo latitude and geo longitude property of a tweet from the Twitter API.
3) If a tweet does not have either of the information above, the tweet would not be included in the visualization step.

Additionally, we perform the elimination of duplicate tweets by applying deduplication based on their edit distance as a similarity measure. A pair of tweets will be considered duplicate if their edit distance is 0 after omitting mentions, urls, and hashtags. If a pair of tweets were found as duplicate, then the one that would be included in the visualization process is the one which was posted earlier. This could be known by comparing the 'created at' property from a tweet. A library called Folium was used to generate the map visualization of disease spread in Indonesia. Folium is a library for Python that can generate a web page that contains an interactive map. We used the tileset from OpenStreetMap to show a map of Indonesia.

4. Experiment results

In this research, we separately evaluate the proposed classification, clustering, and the NER system. Overall, the result of the proposed system is quite satisfying. The following sections will discuss the result of the evaluation for each system.

4.1. Classifier evaluation

The proposed classification model was evaluated using the 5-fold cross-validation. The training dataset consisting of 2,000 tweets were divided into 5 folds, and the classifier was trained 5 times using each fold as the test dataset. The result of the cross-validation is shown in Table 3. As presented in Table 3, the highest F1 score was acquired on the third training which is 0.967, while the average of the F1 score acquired is 0.828.

4.2. Clustering evaluation

The clustering process was performed five times to the 2,000 tweets that manually assigned the disease type label. These labels were used as our gold standard and were used to measure the quality of the clustering result. Table 4 shows the

clustering result of one of the five experiments with words that frequently occurred in each cluster. Tweets indicating URTI type disease are shown to be clustered to two clusters (#0 and #2) which are represented by the keyword ISPA (URTI in English). The clustering also fails to separate diarrhoea and malaria tweets, which are grouped in c#4.

Table 3. 5-fold cross-validation on the classifier model.

Iteration (k)	Precision	Recall	F1 Score
1	0.864	0.901	0.882
2	0.970	0.454	0.618
3	0.985	0.950	0.967
4	0.783	0.979	0.870
5	0.979	0.679	0.802
Average	0.916	0.792	0.828

Table 5 shows the inner evaluation criteria called within-cluster sum of squares (WCSS) which is the sum of the distance of each tweet to the cluster centre. From Table 5, we can see that the value of WCSS is relatively large. However, the clustering method still failed to group the tweets into 5 different types of common diseases in Indonesia. Thus, we performed the further analysis by using the adjusted Rand index (ARI). The result shows that the average of ARI score acquired is only 0.496 indicating that the clustering method results in poor quality.

4.3. Location entity recognizer evaluation

For the purpose of evaluating the location entity recognizer, we only included tweets that have valid location names in Indonesia and omitted tweets that indicate disease spread outside the country. The accuracy of our dictionary-based location entity recognizer is compared to the performance of Polyglot-NER applied to the same dataset. The Polyglot-NER is a similarly built NER based on data from Wikipedia, although it is a more general NER and support 40 languages, including Bahasa Indonesia.

The comparison results between the performance of our developed NER and Polyglot-NER applied to the same dataset is shown in Table 6. We can see that our system acquired an accuracy of 79.8%, while the Polyglot-NER only achieved 76.45%. Thus, our NER system has worked relatively as well as the Polyglot-NER which has been validated from an earlier research.

Table 7 shows some examples of location identification result of some tweets by using Polyglot-NER and our Location Dictionary NER. Both systems recognized the location name ”Pelalawan” from the first tweet, while from the second and third tweet, only one of the two methods recognized the location names. For the second tweet, our dictionary-based system could not recognize” Ciomas”, because ”Ciomas” is a district, and we have not added districts’ names to our dictionary. There is also the case where Polyglot-NER could not recognize the location name “Meranti” from the third tweets, while our system successfully recognized it. Additionally, there is also the fourth example which shows an example of a case where neither our system nor Polyglot-NER can recognize”Mandau” district as a location name.

4.4. Map of disease spread in Indonesia

We generated a map visualization of disease spread in Indonesia by using a Phyton library called Folium. We represent each tweet as a marker which colour is based on the disease type indicated by the tweet. The blue label represents tweets about dengue fever, while green, red and purple labels represent diarrhoea, filariasis (elephantiasis), and upper respiratory tract infections (URTI). Lastly, the orange label is used for malaria. The result of the visualization step is shown in Figure 3.

Based on the map, the most occurrences disease is URTI which is marked with purple labels. The results of our study found that the incidence of URTI mostly occurred in Sumatra and Kalimantan islands. This is reinforced by data released by BNBP (Indonesian National Board for Disaster Management) that there are 6 provinces that have experienced significant increases in URTI disease, namely Kalimantan Tengah, Kalimantan Barat, Kalimantan Selatan, Sumatera Selatan, Riau and Jambi (Dinda, 2019; Yulika, 2019). Residents of these six provinces are prone to URTI because of forest and land fires incidents (Dinda, 2019).

Table 4. Clustering result.

Cluster #0	Cluster #1	Cluster #2	Cluster #3	Cluster #4
warga ispa terserang kabut riau	orang wabah taiwan meninggal berdarah	ispa penderita orang kabut meningkat	gajah kaki penyakit warga menderita	diare malaria papua demam tni

Table 5. WCSS and ARI score of the clustering result.

Experiment	WCSS	ARI Score
1	3744.056	0.547
2	3679.819	0.543
3	3645.229	0.770
4	3724.660	0.305
5	3697.809	0.317
Average	3698.315	0.496

Table 6. NER performance comparison.

Method	Correct	Error	Total	Accuracy
Polyglot-NER	435	134	569	76,45%
Location Dictionary	454	115	569	79,80%

Table 7. NER result examples.

Tweet	Polyglot-NER	Location Dict. NER
In a month, 1.893 residents of Pelalawan Infected with URTI	Pelalawan	Pelalawan
Three Patients Infected with Filariasis Died in Ciomas	Ciomas	-
Public Health Office Recorded 40 Residents of Meranti Suffers Filariasis	-	Meranti
Smog, 8,248 Residents of Mandau Infected with URTI (RiauPos) #MelawanAsap	-	-

Figure 3. Tweets visualization on map of Indonesia.

Meanwhile, orange labels were detected in Papua and Bengkulu provinces, indicating the presence of malaria. This is in accordance with Indonesia Ministry of Health's findings that Papua still ranks first as the area with the most malaria-affected population based on the distribution of API (Annual Parasite Incidence) (Levi, 2015; Utami & Efendi, 2019). In western Indonesia, Bengkulu province was ranked first in the case of Malaria (Usmin, 2016). In the meantime, dengue fever marked with a blue label was detected in Jogjakarta province of Java island. Additionally, elephantiasis marked with red labels is found in Central and West Java. This is supported by the findings of the West Java Health Office who discovered elephant foot disease in several areas, such as Tasikmalaya and Depok (Suryarandika & Aminah, 2017).

5. Conclusion and future works

Our research successfully identifies tweets indicating disease emergence and generates a map visualization using a series of methods consisting of classification, clustering, and a named-entity recognition system. We identified 335 tweets which indicate disease emergence in a certain location in Indonesia. We compared our findings with the official annual epidemic report from Indonesian's Ministry of Health and popular news sites in Indonesia. We found that our findings are in line with the reports that published several weeks later. Thus, we believed that using Twitter may help the government officials to get an overview of the spread of disease in Indonesia in a relatively short time. However, the results of clustering disease type are not good enough for real-world application, showing there is still a big room for improvement. Furthermore, we believe for future work, researchers can try to use data from other social media platform, such as Instagram considering the increasing of Instagram's popularity, especially in Indonesia. Using Instagram or other picture-based social media, future works can explore other methods for clustering with higher accuracy for clustering pictures. Additionally, future works that use text-based social media may explore other classification methods such as DBSCAN or fuzzy clustering to try achieving higher accuracy.

References

Al-Rfou, R., Kulkarni, V., Perozzi, B., & Skiena, S. (2015). POLYGLOT-NER: Massive Multilingual named entity recognition. In SIAM International Conference on Data Mining 2015, SDM 2015 (pp. 586–594). https://doi.org/10.1137/1.9781611974010.66

Chunara, R., Andrews, J. R., & Brownstein, J. S. (2012). Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *American Journal of Tropical Medicine and Hygiene, 86*(1), 39–45. https://doi.org/10.4269/ajtmh.2012.11-0597

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing Twitter messages. In SOMA 2010 - *Proceedings of the 1st workshop on social media analytics* (pp. 115–122). https://doi.org/10.1145/1964858.1964874

Dinda, D. (2019). *Penderita ISPA Akibat Karhutla Tembus 919 Ribu Orang.* Retrieved January 5, 2020, from https://www.cnnindonesia.com/nasional/20190923160933-20-433052/penderita-ispa-akibat-karhutla-tembus-919-ribu-orang

Freifeld, C. C., Mandl, K. D., Reis, B. Y., & Brownstein, J. S. (2008). HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. *Journal of the American Medical Informatics Association, 15*(2), 150–157. https://doi.org/10.1197/jamia.M2544

Han, B., Cook, P., & Baldwin, T. (2014). Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research, 49*, 451–500 https://doi.org/10.1613/jair.4200

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 1398, pp. 137–142). https://doi.org/10.1007/BFb0026683

Kementerian Kesehatan RI. (2013). Pedoman sistem kewaspadaan dini dan respon. Guidelines for early alert and response systems. Jakarta.

Levi, C. (2015). Penderita Malaria di Papua Masih Tertinggi. Retrieved January 5, 2020, from https://nasional.tempo.co/read/660785/penderita-malaria-di-papua-masih-tertinggi

Mourad, A., Scholer, F., Sanderson, M., & Magdy, W. (2018). How well did you locate me? effective evaluation of Twitter user geolocation. In Proceedings of the 2018 *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018* (pp. 437–440). https://doi.org/10.1109/ASONAM.2018.8508701

Nagar, R., Yuan, Q., Freifeld, C. C., Santillana, M., Nojima, A., Chunara, R., & Brownstein, J. S. (2014). A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives. *Journal of Medical Internet Research, 16*(10), e236. https://doi.org/10.2196/jmir.3416

Naradhipa, A. R., & Purwarianti, A. (2012). Sentiment classification for Indonesian message in social media. In *2012 International Conference on Cloud Computing and Social Networking (ICCCSN)* (pp. 1-5). IEEE https://doi.org/10.1109/ICCCSN.2012.6215730

Norvig, P. (2009). How to Write a Spelling Corrector. Retrieved December 16, 2015, from http://norvig.com/spell-correct.html

Ramadona, A. L., Tozan, Y., Lazuardi, L., & Rocklöv, J. (2019). A combination of incidence data and mobility proxies from social media predicts the intraurban spread of dengue in *Yogyakarta, Indonesia. PLoS Neglected Tropical Diseases, 13*(4). https://doi.org/10.1371/journal.pntd.0007298

Ranovan, R., Doewes, A., & Saptono, R. (2018). Twitter data classification using multinomial naive bayes for tropical diseases mapping in Indonesia. *Journal of Telecommunication, Electronic and Computer Engineering, 10*(2–4), 155–159.

Shi, N., Liu, X., & Guan, Y. (2010). Research on k-means clustering algorithm: An improved k-means clustering algorithm. In *3rd International Symposium on Intelligent Information Technology and Security Informatics, IITSI* 2010 (pp. 63–67). https://doi.org/10.1109/IITSI.2010.74

Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). 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). https://doi.org/10.1371/journal.pone.0019467

Suryarandika, R., & Aminah, A. N. (2017). *Kasus Kaki Gajah Tersebar Merata di Kabupaten Tasik*. Retrieved January 5, 2020, from https://www.republika.co.id/berita/nasional/daerah/17/07/13/ot0ujw-kasus-kaki-gajah-tersebar-merata-di-kabupaten-tasik

Taufik, N. (2015). Myner: Recognizing Mamed-entity on Bahasa Indonesia Tweets. *Myner: Pengenalan Entitas Bernama pada Tweet Bahasa Indonesia.* Bachelor Thesis. Universitas Indonesia.

Usmin. (2016). *70 Persen Wilayah Bengkulu Belum Aman Malaria*. Retrieved January 5, 2020, from https://www.beritasatu.com/nasional/362496/70-persen-wilayah-bengkulu-belum-aman-malaria

Utami, S. H., & Efendi, D. A. (2019). Masih Peringkat Pertama, 2030 Kemenkes Targetkan Papua Bebas Malaria. Retrieved January 5, 2020, from https://www.suara.com/health/2019/08/22/202243/masih-peringkat-pertama-2030-kemenkes-targetkan-papua-bebas-malaria

Wahyudi, G., & Budi, I. (2004). Pengenalan entitas bernama berdasarkan informasi kontekstual, morfologi, dan kelas kata. *Jurnal Ilmu Komputer Dan Teknologi Informasi*, 4(1), 33–39.

We are Social. (2019). Digital 2019: Indonesia. Jakarta.

Yulika, N. C. (2019). Jumlah Penderita ISPA di 6 Provinsi Mencapai 500 Ribu Orang. January 5, 2020.

Zulfa, I. (2015). *Sistem pemantau influenza like illness dan visualisasinya memanfaatkan twitter*. Universitas Pendidikan Indonesia.