Topic modeling of weather and climate condition on twitter using latent dirichlet allocation (LDA)

To cite this article: Ahmad Fathan Hidayatullah *et al* 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* **482** 012033

Topic modeling of weather and climate condition on twitter using latent dirichlet allocation (LDA)

Ahmad Fathan Hidayatullah[1], Silfa Kurnia Aditya[2], Karimah[3], Syifa Tri Gardini[4]

Department of Informatics, Universitas Islam Indonesia
Jalan Kaliurang KM 14,5 Sleman Yogyakarta Indonesia

[1]fathan@uii.ac.id, [2]silfa.kurnia.aditya@gmail.com, [3]rima.alkatiri@gmail.com, [4]syifatrigardini@gmail.com

Abstract. This study aims to apply topic modeling approach using LDA for Twitter dataset shared by the official Twitter account of BMKG in Java Island. The topic model result can be seen as the representation about what kind of information that posted by BMKG through Twitter. In addition, it can also illustrate the weather, climate, and disaster trends that occurred in Java Island. Based on the topic modeling result, we found five notable topics from BMKG's Twitter accounts. As for the topics discussed, BMKG's Twitter accounts disseminate the information about weather Information and weather forecast; latest weather information in Yogyakarta Region; weather forecast and warning in Central Java and West Java; earthquake information; latest articles and calendar cropping information. In this study, we also illustrated the trends about weather and disaster by analyzing the most frequent words from each region in Java.

1. Introduction

With the emergence of social media, the dissemination of public information is spreading rapidly. Social media that provides flexibility and simplicity, makes people easier to obtain and share information with others. Therefore, social media is a powerful tool that provides accuracy and immediacy of information happening in real time [1]. Twitter is a good example of social media to spread vital and valid information that currently happening. A lot of organizations around the world, including government organizations in Indonesia, utilize Twitter as one of the media to disseminate important information to audiences [2].

Along with the use of social media among the government agencies in Indonesia, there is an agency called BMKG (Badan Meteorologi Klimatologi dan Geofisika) or the Indonesian Agency for Meteorology, Climatology and Geophysics. BMKG has utilized Twitter to inform the current climate condition and natural disaster to the community. BMKG is one of the government agencies that responsible to monitor and also disseminate the information about climate and natural disaster in Indonesia. The information about weather, climate, and disaster are the most important report to be shared as a warning or a reminder to the public. That information plays an important role in various fields such as transportation safety, flight activity, and disaster information. With the availability of accurate information about weather, climate, and disaster, unwanted events can be anticipated earlier and faster.

This study aims to apply topic modeling approach using LDA for Twitter dataset shared by the official Twitter account of BMKG in Java Island. The topic model result can be seen as the illustration of what information provided by BMKG through Twitter. Furthermore, the result can also represent the weather, climate, and disaster trends that occurred in Java Island.

The remainder of this paper is organized into the following structure. Section 2 describes the related work. Section 3 presents the research methodology. In section 4, the results and discussions are explained. Finally, section 5 describes the conclusion of our research.

2. Related Work

2.1. Latent Dirichlet Allocation

Topic modeling is an unsupervised machine learning method that applies clustering to discover latent variables from a large amount of text data. The most popular method for topic modeling task is Latent Dirichlet Allocation (LDA) which introduced by Blei and Jordan (2003) [3]. LDA is explained as a generative probabilistic model for finding the semantic structure of a corpus based on hierarchical Bayesian analysis of the texts [4]. LDA represents a collection of documents as blend topics which contain words with certain probabilities. Figure 1 illustrates the flow of LDA algorithm.

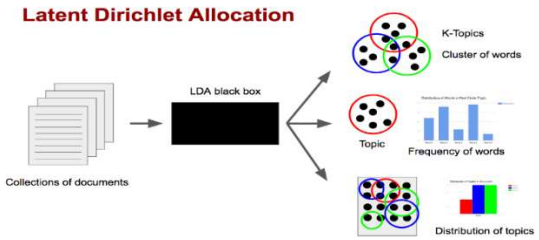

Figure 1. The flow of LDA
(Retrieved from: https://toolbox.kurio.co.id/topic-modeling-696d7ba2592f)

The procedure how LDA works is explained as follows:

1 Initialize some parameters, including the number of documents, the number of topics, and the number of iterations. In LDA, the most important parameter is the number of topics k.
2 Assign each word to a particular topic randomly according to a dirichlet distribution.
3 Repeat the steps for all words in the corpus.

2.2. Topic Modeling using LDA

Since the emergence of topic models, researchers have introduced this approach into the various fields by applying Latent Dirichlet Allocation (LDA) method to build the topic models. LDA has been applied to illustrate the topic regarding traffic information in Java that posted by the Traffic Management Center via Twitter [2]. Wang, et al. (2014) applied topic modeling using LDA on health topic to detect health-related topics discussed in Chinese social media [5]. Liu, et al. (2016) studied the application of topic modeling using LDA and PLSA (Probabilistic Latent Semantic Analysis) in the bio-informatics domains [6]. Neill, et al. (2017) focused on the use of topic model for legislative texts in Britain that could assist in easier tracking of important domain terms that correspond to compliance related issues [7]. LDA has applied to help the historians to identify the most important and interesting topics in newspaper during a specific time period [8].

3. Research Methodology

3.1. Data Retrieval

The data set was retrieved from the official Twitter account of BMKG in Java Island. In this research, we did not collect tweet from all official Twitter account of BMKG because there are several BMKG Twitter accounts that no longer active post the tweets. Therefore, we only grab tweets from the official Twitter account of BMKG that still active to post the tweets until April 2018. The tweet was collected using GET status/user_timeline method on Twitter API v1.1. The user_timeline method provided by Twitter, can reach until 3200 tweets from an account. We also use an open source library in Python

called tweepy to access the Twitter API. Total 26,420 tweets were collected from total 10 official Twitter account of BMKG in Java Island.

3.2. Data Pre-Processing

Pre-processing Twitter data is necessary due to the unstructured text in the tweets [2]. Twitter texts also have some unique characteristics such as character limitation (no more than 140 characters), hashtags (#), RT (retweet), and @username. Therefore, those characters will be removed from the tweets because they do not have any impact to the desired result. Overall, the pre-processing steps in this study follow the previous research [9]. There are some tasks that carried out in pre-processing such as case folding, tokenization, removing URLs, removing punctuations, removing symbols, and removing stop words.

Moreover, we also identify the occurrence of word pairs by joining two words into one term/token because this research will model the topic from a token. To find the word pairs, we use bigrams, a library provided by the Natural Language Toolkit (NLTK) in Python. We then analyze and select which pair of words is an important phrase that often appears in the corpus. For example, the occurrence of the word *"prakiraan"* (forecasts) and *"cuaca"* (weather) arise together 8925 times in the clean tweets. Therefore, we join both words into one token *"prakiraancuaca"* (weather forecasts). Table 1 shows a sample of 10 major pairs of two words from the tweets.

Table 1. Sample of 10 Major Word Pair

No	Word Pair	Number	No	Word Pair	Number
1	*prakiraan cuaca* (weather forecasts)	8925	6	cerah berawan (partly cloudy)	957
2	*citra radar* (radar image)	3204	7	kalender tanam (planting calendar)	920
3	*peringatan cuaca* (weather warning)	2463	8	katam terpadu (integrated planting calendar)	917
4	Jawa Tengah (Central Java)	2024	9	hujan ringan (light rain)	847
5	*info gempa* (earthquake info)	973	10	Jawa Barat (West Java)	785

3.3. Topic Modeling

The process of topic modeling on the whole tweet data is performed by the Latent Dirichlet Allocation (LDA) method. Moreover, this research uses LDA from the Gensim library in the Python programming language. These are the topic modeling steps that conducted in our research:

1 Building term dictionary and corpus.
Dictionary is created by assigning each unique token to an index. The dictionary result will then be used to create a corpus by converting the document list into a document term matrix.
2 Building LDA model object.
The LDA model object is created with the Gensim library and then runs it by training the LDA model on the document term matrix and define the number of topic as a parameter.
3 Visualization.
We use PyLDAvis library to visualize and analyze the topic result. It provides interactive visualization for topic modeling that makes it easy to interpret the generated topic model [10].

4. Result and Discussion

4.1. Topic Modeling

This research grouped tweet topics into 5 topics. The visualization of topical modeling results is shown in figure 2. The left panel illustrates the distance between topics from the generated topic model. Based on the five pieces of the resulting topics, it appears that the topics 1 and topic 3 are mutually interconnected. It shows that there are similarities and closeness of the topic objects discussed between the two groups of topics. As for other topics, the topics 2, 4, and 5 seem to have a considerable distance from each other indicating that among the three groups of topics there is no topic closeness and similarity.

In addition, the right panel illustrates the 30 important words that appear in the corpus. The right panel illustrates about the dominant words discussed topic from the dataset. Based on the right panel visualization, the appearance of the term *'prakiraancuaca'* (weather forecasts) becomes the most frequent word appear in the corpus. Therefore, we infer that the official BMKG Twitter accounts are very often to spread information concerning weather forecast to the society. There are also some other terms that related to weather information for example *"cuaca"* (weather), *"peringatancuaca"* (weather warning), *"infocuacajogja"* (Jogja weather information), *"citraradar"* (radar image), *"kelembaban"* (humidity), *"suhu"* (temperature), and *"angin"* (wind).

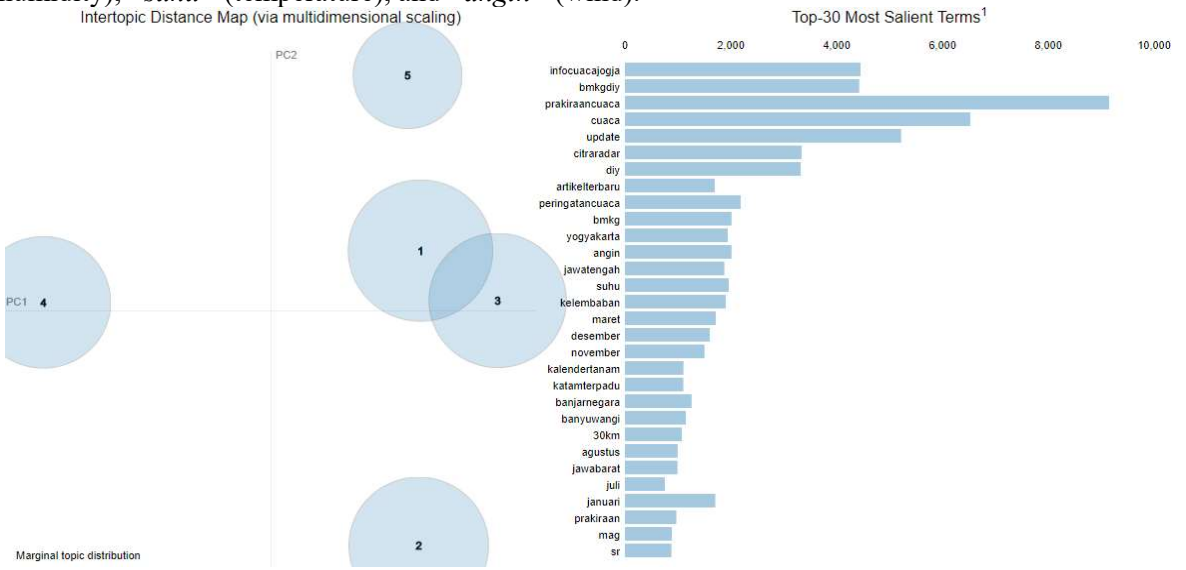

Figure 2. PyLDAvis Visualization

Moreover, we discuss about the interpretation of the meaning of the ten models of topics that have been obtained. The interpretation of the meaning of a topic is viewed based on the 10 most commonly occurring and dominating words in each topic. In addition, we also view the interrelationship between the words of the 10 most frequently appeared from a topic as shown in Table 2.

Table 2. Topic Model Result

Topics	Terms
Topic #1: Weather forecasts	*cuaca* (weather), *prakiraancuaca* (weather forecasts), *angin* (wind), *suhu* (temperature), *kelembaban* (humidity), *banyuwangi*, 30km, *cerahberawan* (partly cloudy), *timur* (east), *hujanringan* (light rain)
Topic #2: Weather information update in the Special Region of Yogyakarta	*infocuacajogja* (jogja weather information), *bmkgdiy* (BMKG of special region of Yogyakarta), *update*, *cuaca* (weather), *citraradar* (radar image), *diy* (Daerah Istimewa Yogyakarta/Special region of Yogyakarta), *Yogyakarta* (Yogyakarta), *januari* (January), *oktober* (October), *februari* (February)
Topic #3: Weather warning for Central Java and West Java Region in a few months	*Prakiraancuaca* (weather forecasts), *peringatancuaca* (weather warning), *jawatengah* (Central Java), *maret* (March), *desember* (December), *November* (November), *update, januari* (January), *agustus* (August), *jawabarat* (West Java)
Topic #4: Earthquake	*bmkg* (BMKG), *banjarnegara, mag* (Magnitude), *sr* (Richter scale), *stasiunmeteorologi* (Meteorological Station), *infogempa* (earthquake info), *lok* (location), *bt (*east longitude), *ls* (south latitude), *jateng* (Central Java)
Topic #5: Recent articles, crop calendars, and weather forecasts	*artikelterbaru* (latest article), *kalandertanam* (planting calendar), *katamterpadu* (integrated planting calendar), *prakiraancuaca* (weather forecasts), *juli* (July), *september* (September), *juni* (June), *mei* (May), *musimkemarau* (dry season), *terimakasih* (thank you)

The most suitable topic for topic #1 is weather forecasts. It can be inferred from the term probability as shown by column number 1 in Table 2. There are two most salient terms of the term probability, *'cuaca'* (weather) and '*prakiraancuaca'* (weather forecast). Other words, such as *'angin'* (wind), *'suhu'* (temperature), *'kelembaban'* (humidity), *'cerahberawan'* (partly cloudy), and *'hujanringan'* (light rain), describe the terminologies that reinforce information about weather forecasts.

Topic #2 illustrates about the weather information update in the Special Region of Yogyakarta. On this topic, the terms that illustrate about weather in Yogyakarta appear frequently, such as *infocuacajogja'* (Jogja weather information), *'bmkgdiy'* (BMKG of Special region of Yogyakarta), *'update', 'cuaca'* (weather), *'citraradar'* (radar image), *'diy'* (Daerah Istimewa Yogyakarta/Special Region of Yogyakarta), and *'yogyakarta'*. Moreover, there are few words that illustrate regarding the month of the weather information, for instance *'januari'* (January), *'oktober'* (October), and *'februari'* (February).

The appearance of the term *'prakiraancuaca'* (weather forecast) in topic #3 is very high. Therefore, the most dominant information for this group is about weather forecasts. Besides, there are some terms, such as *'peringatancuaca'* (weather warning), *'jawatengah* (Central Java), *'jawabarat'* (West Java), *'maret'* (March), and *'desember'* (December), which illustrate the weather warning for Central Java and West Java Region in a few months.

The term *'bmkg'* has the most probability in topic #4. The words, such as *'mag'* (magnitude), *'sr'* (*skala Richter*/Richter scale), *'infogempa'* (earthquake information), *'lok'* (lokasi/location), *'bt'* (*bujur timur*/east longitude), and *'ls'* (*lintang selatan*/south latitude) illustrate the information about earthquakes. Two other words, Banjarnegara and Jateng (Central Java) represent the earthquake location.

Finally, topic #5 illustrates several topics, including recent articles, crop calendars, and weather forecasts. The occurrence of the term *'artikelterbaru'* (recent articles), is quite dominating on topic 5. The terminology *'kalendertanam'* and *'katamterpadu'* represent the information about cropping calendar. The term *'prakiraancuaca'* illustrates the information about weather forecasts.

4.2. Weather and Disaster Trends

Furthermore, we investigate the information from each BMKG's Twitter accounts to see the brief illustration of the weather and disaster information from each region. The investigation is observed based on the occurrence of the terms which related to weather and disaster. From our dataset, there are twelve terms which indicate the weather and disaster information, such as *angin kencang* (strong winds), *badai* (storm), *banjir* (flood), *berawan* (cloudy), *cerah* (bright), *cerah berawan* (bright cloudy), *gelombang tinggi* (high waves), *gempa* (earthquake), *hujan* (rain), *hujan lebat* (light rain), *hujan ringan* (heavy rain), *kilat* (lightning), *petir* (lightning), *siklon* (cyclone), and *tsunami*.

Figure 3 shows that the BMKG of Banyuwangi and Jatiwangi have the most varied weather and disaster information among the BMKG's Twitter accounts. However, BMKG of Jatiwangi more frequently shared about bright cloudy and light rain. In addition, the appearance of the word *hujan* (rain) is quite evenly in all regions except Surabaya (BMKG of Juanda). The most frequent words appeared from BMKG of Juanda is *gempa* (earthquake). Moreover, the information about earthquakes is also frequently shared by BMKG of Bandung Twitter's account.

5. Conclusion

This study has implemented topic modeling using LDA on Indonesian Twitter messages posted by the official Twitter account of BMKG in Java. LDA has promised a good algorithm in topic modeling research. We decided to choose 5 topics as a parameter for LDA method. Based on our study, it can be concluded that discussed topics are about general weather Information and weather forecast; latest weather information in Yogyakarta Region; weather forecast and warning in Central Java and West Java; earthquake information; latest articles and calendar cropping information. From the topic modeling result, BMKG Twitter accounts in Java were mostly used to provide information to the public about the weather condition, which the primary functionality of BMKG itself. The rest topics were discussed about the earthquake information, recent articles, and calendar cropping information. In addition, we also

illustrated the trends about weather and disaster by analyzing the most frequent words from each region in Java. In the future works, it will be interesting if we can use the topic modeling result to predict the weather condition and disaster. This will be useful for the community as an early warning and precaution. In addition, further study can be held by using the improvement of LDA or other topic modeling method in to obtain better topic modeling results.

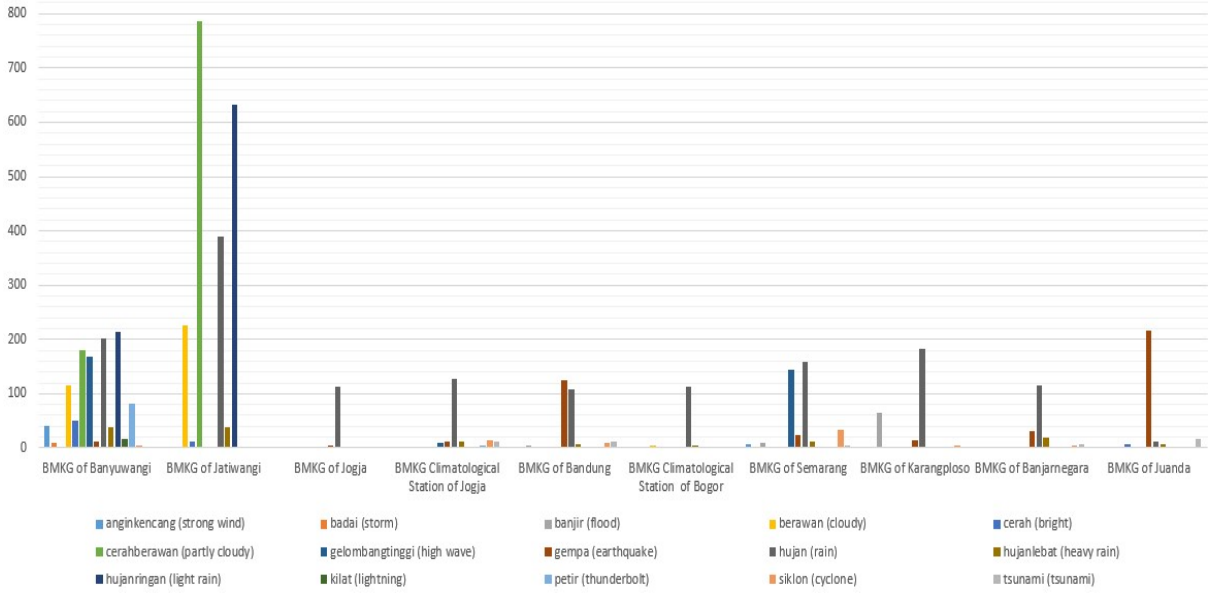

Figure 3. The Illustration of Weather Condition in Each Region

6. Acknowledgement

The authors would like to thank the Department of Informatics, Universitas Islam Indonesia, for technical and financial support.

References

[1] Palen L and Vieweg S 2008 The emergence of online widescale interaction in unexpected events *Proc. of the ACM 2008 Conf. on Computer Supported Cooperative work* pp 117–26
[2] Hidayatullah AF and Ma'arif MR 2017 Road traffic topic modeling on twitter using latent dirichlet allocation *Int. Conf. on Sustainable Information Engineering and Technology (SIET)* pp 47–52
[3] Blei DM, Ng AY and Jordan MI 2003 Latent dirichlet allocation *J. Mach. Learn. Res.* vol **3** pp 993–1022
[4] Blei DM, Edu BB, Ng AY, Edu AS, Jordan MI and Edu JB 2003 Latent dirichlet allocation *J. Mach. Learn. Res.* vol **3** pp 993–1022
[5] Wang S, Paul MJ and Dredze M 2014 Exploring health topics in Chinese social media: An analysis of Sina Weibo *AAAI Workshop on the World Wide Web and Public Health Intelligence* vol **31** pp 20–3
[6] Liu L, Tang L, Dong W, Yao S and Zhou W 2016 An overview of topic modeling and its current applications in bioinformatics *Springerplus* vol **5** no **1** p 1608
[7] O'Neill J, Robin C, O'Brien L and Buitelaar P 2017 *An Analysis of Topic Modelling for Legislative Texts ASAIL*
[8] Yang TI, Torget AJ and Mihalcea R 2011 Topic modeling on historical newspapers *Proc. of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* pp 96–104
[9] Hidayatullah AF and Ma'arif MR 2017 Pre-processing tasks in indonesian twitter messages *IOP Conf. Series: J. of Phys.* 801
[10] Sievert C and Shirley K 2014 LDAvis: A method for visualizing and interpreting topics *Proc. of*

the Workshop on Interactive Language Learning, Visualization, and Interfaces pp 63–70