2021 9th International Conference on Information and Communication Technology (ICoICT) | 978-1-6654-0447-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICoICT52021.2021.9527513

Disaster Tweet Classification Based On Geospatial Data Using the BERT-MLP Method

1st Iqbal Maulana
School of Computing
Telkom University
Bandung, Indonesia
ibalmaulana@student.telkomuniversity.ac.id

2nd Warih Maharani
School of Computing
Telkom University
Bandung, Indonesia
wmaharani@telkomuniversity.ac.id

***Abstract*—as a popular social media in the world and even in Indonesia, Twitter has a variety of popular topics making these topics trending, including the topic of natural disasters that have occurred in Indonesia. The DKI Jakarta flood disaster in early 2020 made a big scene on trending twitter topics. This study aims to classify these tweets into "flooded" and "not flooded" predictions with the tweets and geospatial features. The model proposed for classifying is BERT-MLP. Bidirectional Encoder from Transformers (BERT) is used in the pre-trained model to classify these tweets and Multi Layer Perceptron (MLP) is used to classify geospatial features. The scenario designed for the model focuses on the preprocessing of tweets as follows without stopword removal, without stemming, with both, and without both. Once classified, the tweet will be visualized into a two-dimensional interactive map. The best scenario results have an accuracy of 82% in scenarios without stemming and with stopword removal. This is due to the stemming process eliminates some of the features in tweets around 6%. This study also shows the relationship between the influence of negative context tweets on the "not flooded" class with an orientation of 65% of the total data. However, defining manual stopwords can affect because stopword removal will not delete words that still have context related features to the topic.**

***Index Terms*—geospatial data, tweet, BERT, MLP, deep learning**

I. Introduction

In recent years, we have been faced with a series of natural disasters that have caused tremendous financial, environmental, and human losses [1]. The unpredictable nature of natural disaster behavior makes it difficult to have comprehensive situational awareness to support disaster management. During a disaster, affected individuals often express their feelings by uploading their status to social media, such as Twitter [1]. While using Twitter social media, individuals will find the latest updates from official government accounts or response organizations on Twitter to ask for help or upload information that can be used to raise situational awareness, in particular regarding specific data from the location of the disaster that occurred and how the disaster spread [2].

Geospatial data is data about geographic location, dimensions, size, characteristics of natural or man-made objects that are below, on, or above the earth's surface [2]. On Twitter, geospatial data can be used to find out spatial information (location) which is the location of the source emergency public perceptions disaster on social media [3]. DKI Jakarta floods in early 2020 are a case study in this research. The hashtags # flood2020, #banjir, and #banjirjakarta are perched at the top trending topic on social media Twitter in early January 2020 and late February 2020 with total tweets reaching more than 100,000 [4]. This case study was chosen because the location of the flood disaster in DKI Jakarta in 2020 was not immediately known by the public, so it needed an information search process to determine its exact distribution [2]. The tweets are taken based on their location will be filtered first for tweets uploaded in the DKI Jakarta province. The purpose of this research is to create a system to analyze and classify the locations of natural disasters with geospatial data and tweets obtained from Twitter and visualize them on a digital map so that the location of each point of a disaster can be seen and known.

The model proposed for classifying is BERT-MLP, Bidirectional Encoder from Transformers (BERT) is used in the pre-trained model to classify these tweets, and Multi Layer Perceptron (MLP) is used to classify geospatial features. BERT is a two-way encoding representation of transformers and has a technique for pre-training NLP (Natural Language Processing) developed by Google [5]. BERT was created and published in

2018. The tweet has a geospatial feature in the form of numbers, therefore MLP is used to process geospatial data first. Jacob's research was using Pre-Trained BERT with multi-language against language processing got an accuracy of 86.7% [5]. Yaru Hao's research identified the effectiveness of the BERT method on the visualization of data [6]. In this research, the BERT model will classify Indonesian tweets because the bilingual model consists of 104 languages includes Indonesian text. MLP will be used a geospatial feature in the form of latitude and longitude as decimal input and recombine tweet data and geospatial data in the MLP model by adding layers.

II. Related Works

In previous research, Amit Agarwal analyzed geospatial sentiment twitter data for the UK-EU referendum. This study tries to find out the British politicians who were being talked about the most and what people think of them sentimentally. However, the research only analyzed in terms of geospatial data and did not classified tweets [3]. Martin Werner classified building types into commercial and residential from geospatial text mining Twitter with SVM, Naive Bayes and Neural Network. The feature space of the two classes produced by geotagged Los Angeles Twitter text messages. However, the features that are classified are based on tweets only, not with geospatial features in them [7]. Mark Kibanov classified geospatial data using the adaptive KNN method. However, the accuracy of geospatial San Francisco crimes dataset did not increase after continuing training because of imbalance data distribution. Moreover, there is no added features or other weights at each of these geospatial points [8].

A. Geospatial Data

Geospatial data is data about geographic location, dimensions, size, space and characteristics above the earth's surface. The space aspect shows the location and position of an object expressed in a specific reference coordinate system [2]. In Lee's research, geospatial data can be used for several case studies such as fuel and time savings, income generation, urban planning, health care and even disaster mapping [2]. In this research geospatial data is the location point or longitude and latitude of the uploaded tweet to be identified.

B. Bidirectional Encoder from Transformers

BERT was first introduced by the Google AI team in 2018 [5]. BERT is based on the Transformer architecture which is a attention mechanism that studies contextual relationships between words in a text. BERT's goal is to generate language model. Bidirectional means BERT learns information from the left and right side of the token context during training. In transfer learning, BERT has two stages in carrying out pre-training, namely Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), in MLM researchers replace 15 % words in each sequence with the [MASK] token [5]. The model then tries to predict the original value of the word [MASK] based on the context given by other words which are not covered in the sequence. During the NSP, the model accepts sentence pairs as input and learns to predict if the second sentence in the pair is the next sentence in the original document. Sentiment analysis is the utility of BERT method by adding [CLS] token at the beginning of the first sentence and [SEP] token inserted at the end of each sentence [5]. More details can be seen in the Figure 1.

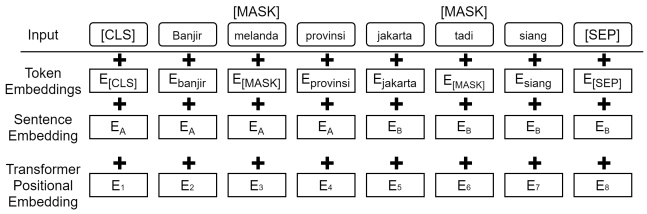

Fig. 1. Embeddings illustration on BERT

In sentence embedding, the sentence is divided into sentence A or sentence B which is added to each token. Positional embedding is appended to each token to indicate its position in the order. The output from the [CLS] token is converted into a 2×1 vector shape using the layer classification. In classification tasks such as sentiment analysis, it is carried out similarly to the next sentence classification, by adding a layer classification above the transformer output for tokens [CLS] as will be done in this research.

III. Methodology

The system development is described into four stages : the crawling of the tweet, preprocessing, training, performance measurement and data visualization

A. System Overview

Twitter data is retrieved in the form of tweets using Twitter's Application Program Interface (API), then the data is stored in excel format. The types of preprocessing scenarios are : without stemming, without stopword removal, with both, and without both. The tweets pass

through the BERT model meanwhile the longitude & latitude features are included in the MLP model. Geospatial data and tweet data that have been entered BERT-MLP model will be reclassified by the MLP method by adding several layers of data will be visualized based on the geospatial coordinates on a two-dimensional map.

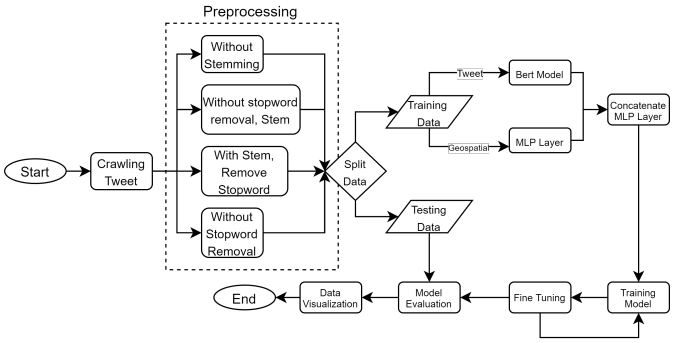

Fig. 2. System Overview

TABLE I
EXAMPLES OF THE TWEETS

No	Tweets	Geo	Label
1	RT @soleh Huj4n Sebentar Menghasilkan Banjir di Cengkareng,, #banjir #banjirjkt	106.74,-6.14	Flooded
2	Rumah Murah Type Petakan Type 30 SHM Bebas Banjir	106.86,-6.18	Not Flooded

TABLE II
PREPROCESSING

Before	After
rt @soleh huj4n menghasilkan banjir di cengkareng,, yg berlumpur!! g pernah dikuras2 oleh @dkijakarta #banjir #banjirjkt bit.ly/votedki	hujan menghasilkan banjir di cengkareng yg berlumpur tidak pernah dikuras oleh banjir banjir

B. Crawling Tweets

The crawled tweets are related to the flood disaster phenomenon in DKI Jakarta Province in early 2020. R programming language with R Studio tools and the rtweet library[1] and twitteR library[2] are used for crawling data, because already covers the procedures and functions written in the twitter API documentation, so the use of procedures and functions can be used immediately. We labelled the data manually. Unlabeled tweets will be labeled by three annotators and assessing the tweet based on two stages of assessment. The first stage will be considering in terms of tweets related to the flood disaster in DKI Jakarta and the second stage related to location where the tweet was uploaded compared to the location mentioned in the tweet. The third stage is to compare the location of the tweet with the location where it actually occurred on the petabencana website.

C. Preprocessing

The initial stage of the preprocessing data normalization includes converting tweets to lower cased, remove links / URLs, RT symbols (retweet), symbols (hashtags) and numeric scaling is performed for geospatial data. In this research, the scenario designed for the model focuses on the preprocessing of tweets as follows without stopword removal, without stemming, with both, and without both. More details can be seen in the illustration of Table II

[1]MIT license library
[2]Artistic 2.0 license library

D. Network Architecture BERT-MLP

in the previous experiment, this study used the SVM method to classify. However, the application of the tweet input model (text) and geospatial data (longitude latitude) was not accepted by the SVM method due to differences in data types, the output of text feature extraction in the form of vectors and geospatial data as raw numeric. This can be overcome by using deep learning (BERT-MLP) and configuring the layers by adding a concatenate layer and also having its own weighting mechanism. In this BERT-MLP model, there are two stages used to create a network architecture, namely pre-trained and fine tuning.

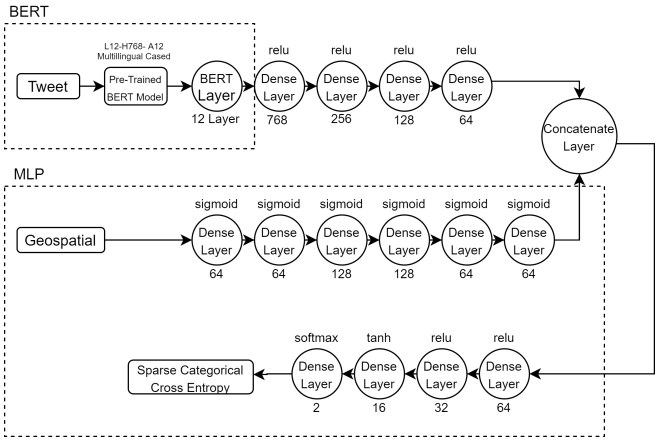

Fig. 3. Network Architecture BERT-MLP

1) Pre-Trained Model: This study uses pre-trained data for the BERT-Base model3, Multilingual Cased or multi-language (multi-cased L-12 H-768 A-12) sourced from Google AI [5]. This model allows for large scale data and billingual language model with total of 104 languages including Indonesian as well. The model has 12 layers, 768 hidden, 12 heads and 110 M parameters. Tokenizing tweet obtained from the pre-trained model tokenizer by adding [CLS] token at the beginning of a tweet and a [SEP] token inserted at the end of each tweet with maximum number of sequence of 256.

2) Fine Tuning: The model architecture is based on the two types of feature input text (tweets) and numeric (geospatial). In the tweet data, the model has a loaded weight on the previous pre-trained models. A layer of multilingual BERT is added, totaling 12 layers and 768 hidden. The next process is followed by a dense layer with the Rectified Linear Unit (ReLu) activation function, the role of activation relu does not activate all neurons because all negative values will be in the 0 range [0, infinite] so that the process is efficient [9]. In numeric or geospatial data across the dense layer of sigmoid activation, considering that geospatial data is a number of coordinates (longitude, latitude), has the purpose of possibility to checking occurrence whether is flooding or not. The sigmoid function has a range [0,1] which is a scale probability. The two layers will be merged with the concatenate layer. Next the layer goes through dense layer activation function relu and tanh which are differentiable with range [-1,1] and end with function layer softmax activation not only has a range [0,1] but predicts a multinomial probability distribution[13]. The network architecture can be seen in Figure 3.

E. Data Visualization

Before evaluate the model, testing data are already preprocessed. Numeric scaling of geospasial will be inverse and returned to its original form. Once the metric is evaluated, the test data will be visualized into a two dimensional map. The data that is ready to be visualized will be labeled "Flooded" and "Not Flooded" in the map pattern so that it can show the difference in the exact location of the coordinates which is given. The data will be entered on a two-dimensional map using the foliu library[3] and located precisely in the map image of the Province of DKI Jakarta with latitude and longitude positions. Tweets are indicated by a marker, red mark is flooded location and blue mark indicates it is not flooded. For example can be seen in Figure 4.

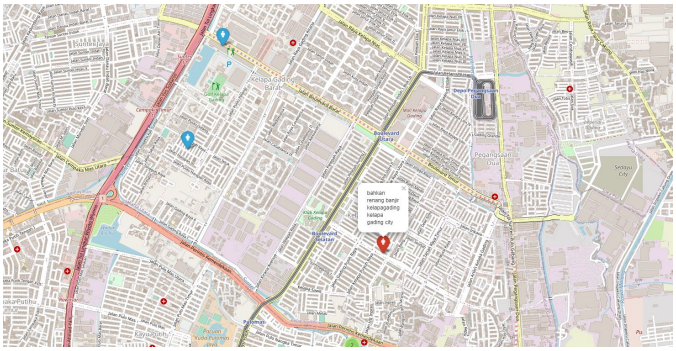

Fig. 4. An interactive map showing users affected by the disaster

IV. Evaluation

In this research, the measurement of the implementation of the model is carried out to assess how well the BERT-MLP is applied to the case of several scenarios. This research conducted a study with four scenarios that focus on preprocessing tweets. Every tweet and geospatial data goes through a preprocessing before entering the classification model. The model trained using Adam's optimizer, with learning rate lr = 0.0005, trained for 9 epochs, each epoch divided by into 15 batches [10]. The loss function used is sparse categorical cross entropy and metrics measurement by sparse categorical accuracy. In the training process a scheduler is applied with a maximum learning rate $1e-5$, end learning rate $1e-7$ and warming up epoch by 3. Early stopping is used to avoid overfitting with monitoring loss validation which has $\Delta = 0.001$ and patience for the stopping up to 3 times the epoch. Accuracy is used for evaluating the model, which is the ratio of the predicted true or false (positive and negative) to the whole data, with equations

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (1)$$

A. Experimental Result

The data crawling stage obtained 72,476 tweets from the results of the crawling period at the end of February 2020, filter by taking the geospatial data (location) obtained amounted to 21,103. The data are splitted to train and testing set is 80:20. Amount of the training data is 16,882 and the testing data is 4,221. The results of the training and four scenarios can be seen in Table III.

The training process was carried out with 9 epochs, can be seen in Figure 5. In the stemming and stopword

[3]MIT license library

TABLE III
TRAINING RESULT

Scenario	*Train Acc(AVG)*	*Valid Acc(AVG)*
No stemming & stopword removal	82.28 %	79.83 %
No stemming	**82.53 %**	**82.75 %**
No stopword removal	82.31 %	80.25 %
Stemming & stopword removal	80.65%	80.95 %

removal scenario, the accuracy of validation and training increases periodically parallel, but they did not increase again after 9 epochs at 80% accuracy. Whereas, those who did not use a stopword increased by 82% in training accuracy but validation accuracy remained at 80%. The result of this scenario is due to the influence of the stemmer on the training process which cuts off many affixes in important words which make them an unknown word. If the process does not use stopword removal, there are still many common words that will be included in the model training process.

The scenario that has the highest accuracy in the training process is a preprocessing scenario without stemming but including stopword removal with 82,53% training accuracy and 82,75% validation accuracy. The model is then evaluated using confussion matrix and testing data yields an accuracy of 83%.

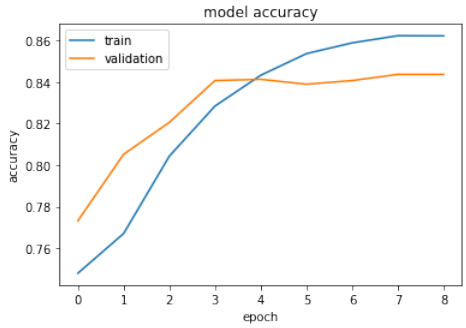

Fig. 5. Accuracy

B. Analysis

The analysis is provided by identifying the misclassification of each tweet by comparing it to the actual class. After obtaining the right parameters in looking for a classification error, the factors that have an influence in the classification process are obtained. These factors are :

1) Stemming and Stopword Removal: In a previous study the stemming process did not have a significant impact on increasing theconstellations in Indonesian text classification [11]. Stemming serves to cut prefix and suffix for take the root word, but the process does not consider the derived root causes the emergence of new words that are not in the corpus, such as the detected proper noun example "*Bekasi*" produces a "*bekas*" output and "*kerepotan*" has a "*kerepot*" output. Stemmer removes up to 130,438 characters in the whole tweet dataset which is 6% of the total number of originals. Stopword deletion is based on the utility of classification model, therefore the stopword corpus is created manually considering the words that need to be deleted and not deleted in the context of the 2020 DKI Jakarta floods. Words like "*disini*", "*dimana*" or words representing a "flood" or "not flooded" state are not used. Based on this analysis, this research produced the best pre-process tweet model using without stemming and using stopword removals scenario.

TABLE IV
NEGATIVE TWEET EXAMPLES

Tweets
Jakarta bakal banjir lagi nih..entah harus bagaimana menyadarkan anies baswedan agar memprioritaskan antisipasi banjir ketimbang Formula E
AniesTenggelamkanDKI #BanjirJakarta2020 #AzabGabenerBodong #banjir #BanjirDKIJakarta

2) Negative context tweets: This negative tweets refers to expression that denigrates a person or subject, given the case study of DKI Jakarta's flood tweets in domination by users who provide criticism and attacks on local governments [12]. In this research the BERT model is modified and uses the pre trained IndoBERT model [13] to classify sentiments positive and negative. The results obtained from the dataset contained 65% negative tweets. Negative tweets had an influence on the location based text classification model due to all the tweets negative tweet sentiment does not have a major relation to the "not flooded" label. Examples of negative tweets in the dataset can be seen in the Table IV.

3) Non Impacted user keyword: We analyzed the unique words that were in the "not flooded" class and ranked them according to their occurrence frequency. The top five is 'anies', 'formula', 'ahok', 'aman', 'president', with the highest word 'anies' having 201 total words. There is also the word 'aman' which represents a location that is not flooded with a frequency of 69, and

the rest has negative content. However, there are several occurrences of words that represent flooding, such as 'berenang' (swim) with a frequency of 41 'air' (water) with a frequency of 31, 'bencana' with a frequency of 49. These words should be in tweets in the flood class for example the word "berenang" means swimming, one of the tweets that gives an example is "I'm out of my room and I'm swimming. Because flood has entered my hous". Further researched tweets containing the word swimming only amounted to 100 tweets among the rest 4000 tested tweets..

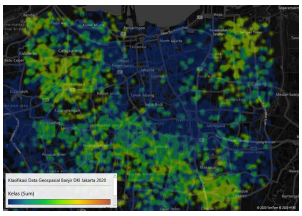

Fig. 6. Experiment result

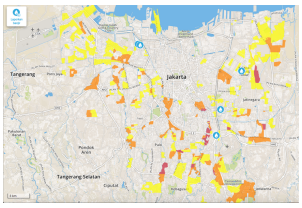

Fig. 7. petabencana map, source : petabencana.id

4) Data Visualization: Folium provides an interactive map on python. We open the test results with an excel 3d map. When we compare the results of the visualization with the flood mapping of the petabencana.id [4], at a glance it has the same pattern. We analyze this mapping will not be accurate with the exact location, but the cluster of affected areas is formed and according to what was reported. For example, the central Jakarta area which has a low frequency of "flooding" and the worst flood area is in the position of south Jakarta [4].

V. Conclusion

Based on our experiments, it can be concluded that the BERT-MLP model with preprocessing without stemming and using stopword removal has better accuracy results than any other scenario. The problem that cause misclassified are including the stemming process cut off affixes to words, tweets that has negative context, and misclassification of impacted user keywords although it doesn't give the exact location of the disaster, the model can predict tweets talking about floods and the geospatial data set creates an affected region from it. For further disaster tweets research it could be better by adding a negative class and adding more unique word data for the people affected by the disaster. This research is limited to case studies of flood disasters, therefore it can be carried out on other types of natural disasters.

[4]BMKG disaster mapping

References

[1] L. King, "Social media use during natural disasters: An analysis of social media usage during hurricanes harvey and irma," vol. 1, 03 2018, pp. 20–23.

[2] Q. Huang and Y. Xiao, "Geographic situational awareness: Mining tweets for disaster preparedness, emergency response, impact, and recovery," *International Journal of Geo-Information*, vol. 4, pp. 1549–1568, 08 2015.

[3] A. Agarwal, R. Singh, and D. Toshniwal, "Geospatial sentiment analysis using twitter data for uk-eu referendum," *Journal of Information and Optimization Sciences*, vol. 39, pp. 1–15, 11 2017.

[4] Yuslianson, *Hujan Deras Landa Jakarta, Tagar Banjir Trending Topic di Twitter*, 2020 (accessed February 25, 2020), https://www.liputan6.com/tekno/read/4185485/hujan-deras-landa-jakarta-tagar-banjir-trending-topic-di-twitter.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[6] Y. Hao, L. Dong, F. Wei, and K. Xu, "Visualizing and understanding the effectiveness of bert," 01 2019, pp. 4134–4143.

[7] M. Häberle, M. Werner, and X. X. Zhu, "Geo-spatial text-mining from twitter–a feature space analysis with a view toward building classification in urban regions," *European journal of remote sensing*, vol. 52, no. sup2, pp. 2–11, 2019.

[8] M. Kibanov, M. Becker, J. Mueller, M. Atzmueller, A. Hotho, and G. Stumme, "Adaptive knn using expected accuracy for classification of geo-spatial data," in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, 2018, pp. 857–865.

[9] H. Zhang, T.-W. Weng, P.-Y. Chen, C.-J. Hsieh, and L. Daniel, "Efficient neural network robustness certification with general activation functions," in *Advances in neural information processing systems*, 2018, pp. 4939–4948.

[10] Z. Zhang, "Improved adam optimizer for deep neural networks," in *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*. IEEE, 2018, pp. 1–2.

[11] A. F. Hidayatullah, C. I. Ratnasari, and S. Wisnugroho, "Analysis of stemming influence on indonesian tweet classification," *Telkomnika*, vol. 14, no. 2, p. 665, 2016.

[12] Yuilyana, *Tagar Anies Gak Becus Kerja Trending di Twitter*, 2020 (accessed February 25, 2020), https://www.kompas.tv/article/60919/tagar-anies-gak-becus-kerja-trending-di-twitter.

[13] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar, and A. Purwarianti, "Indonlu: Benchmark and resources for evaluating indonesian natural language understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020.