International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V69I12P212

Volume 69 Issue 12, 103-113, December, 2021

Original Article

Implementation of a Geocoding In Journalist Social Media Monitoring System

Abba Suganda Girsang[1], Sani Muhamad Isa[2], Raditya Fajar[3]

[1]*BINUS Graduate Program–Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*
[2]*BINUS Graduate Program–Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*
[3]*BINUS Graduate Program–Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480*

[1]agirsang@binus.edu, [2]sani.m.isa@binus.ac.id, [3]raditya.fajar@binus.ac.id

Abstract - *Conversations on Twitter as one of the biggest social media platforms, especially in Indonesia, which can be related to problems or events that occur around them, can easily become viral and spread widely. It is also supported by the fact of its evolution, that a piece of news is published on a television station, print media, or online media; in fact, some of it comes from issues or viral events that thrive in the community. This research is a continuation of previous research in building an information system platform for journalists, which helps to find what events or issues have the potential to become viral or continue to be updated with ongoing issues. Coupled with the application of the geocode method and the proposed conversation clusterization using Lingo Algorithm that's provided by Carrot2 Tools. In this study, authors used this algorithm to help determine which conversations were considered important and which were not. These collected conversations can be mapped based on the description of the location or address discussed in the text of the conversation. This will really help journalists to find news material around them, which has proximity to their location and news sources. The success in the geocode process in this study depends on several parameters such as writing location names greatly affects the effectiveness of location name extraction using the NER model that was created, even though it has been trained with the characteristics of the Indonesian region, but the use of slang in showing names locations can be misinterpreted, this is also influenced by the punctuation included, so the separation of location names greatly affects the effectiveness of geocoding. Then the collection of spatial data used also affects the level of the match in finding the described location, as in the example that has been discussed.*

Keywords — *Data Mining, Geocoding, Lingo Clustering, Naive Bayes, Named Entity Recognition, Twitter*

I. INTRODUCTION

By using social media, Twitter, which is generally widely used by the public in disseminating information and announcing an event in real-time. These posts (tweets) generally contain opinions and expressions from users, such as social issues, political issues, crime incidents, and so on. Twitter has 192 million daily active users as of the fourth quarter of 2020. According to senior analyst Neuberger Berman, Hari Srinivasan, the number of daily active Twitter users increased by 26% compared to the previous year.

In Indonesia, the number of Twitter users reached 14.05 million as of January 2021. Twitter is the fifth most popular social media after YouTube, WhatsApp, Instagram, and Facebook [1]. This is driven by the increasing number of smartphone users whose prices are very affordable so that it can make it easier for people to access the internet and also social media as a place to share various things.

Interactions that occur in this media tend to be easier and can easily reach all levels of society, and it can easily describe issues that are developing and even newly formed from community conversations on this Twitter platform. Twitter is not only used by individuals but also by other institutions such as television mass media, especially journalists, to find the latest topics and issues that are developing in society. In an effort to find, obtain and use information that is developing in society, anticipated by the mass media, which is included in the dynamics of a free press, a journalist can use conversations on this Twitter platform to easily find issues that are currently hot. While in general, the Natural Language Processing research area for analytical texts is one of the most important research areas. Studies that are specific to the discipline of microblogging services, especially Twitter, have shown a lot of improvement [2].

The task of finding the location of the incident originating from a Twitter conversation is a challenge for a journalist. To be able to solve these problems, a method is needed to search for location coordinates from a location description tweeted by Twitter users.

Text mining is a branch of data mining that analyzes data in the form of text documents [3]. According to Han, Kamber, and Pei, text mining is one step of text analysis that is carried out automatically by a computer to extract quality information from a series of texts summarized in a document [4]. The initial idea of making text mining is to find patterns of information that can be extracted from an unstructured text [5]. Text mining combined with spatial data can enrich the quality of information from classified text, such as adding location information in the form of coordinates to regional administration in accordance with the governance of the administrative boundaries of Indonesia.

In the field of Geographic Information Systems (GIS), the process of recognizing geographic context is called geoparsing. This geoparsing can be grouped into the Named Entity feature, but this geoparsing extracts more variations from a geographical entity such as street names, province names, city names, village names, telephone numbers, to postal codes.

The process of assigning longitude and latitude attributes to extract the geographic context is called Geocoding. Extraction process address is the most important thing in geoparsing because by getting an address it can be used to do the geocoding process accurately for performing gazetteer features [27], besides the completeness of the address data reference with location information is also the most important thing so that the Geocoding process can run successfully.

If a location can be mapped to an entity in a knowledge graph, a resolution toponym – a special case of entity resolution – can be used to resolve the reference to the location. To support the location reference process to be carried out, a task is needed to extract location names from a text. The names of these locations will be matched with the administrative structure of the Indonesian territory, which consists of the names of Provinces, Regencies, Districts, Villages, Street Names, and Map Objects.

Some of the software that is currently commonly used to perform the geocoding process include:

- Google Maps API
- Yahoo Maps API
- Bing Maps API
- OpenStreetMap API
- Esri ArcGIS API

These services provide a function to find the coordinates of the location based on the address entered. In choosing the service, it will usually consider several things, such as the price and the number of queries that can be performed. For example, the Yahoo Maps API is free, but its geocoding service is limited to 5,000 queries per IP address per day. This geocoding process requires a large amount of time and resources. To use the Google Maps API, an API Key is needed, and it can be used for free, but they don't get complete query results such as the parcel data (geometry); they only return a pair of location coordinates without any parcel data and structure. Administrative boundaries in accordance with the geographical conditions of Indonesia.

Furthermore, to make it easier to determine the tweets that are important and talked about by users in real-time, a clustering process of Twitter posts is needed to be able to group conversations based on the most discussed issues. Ripati et al. [6] proposed in their research using the clustering technique based on word frequency and topic taxonomy on Wikipedia to find the topics discussed in the tweet. The research said that the proposed algorithm had given better results than algorithms involving only word frequency.

II. RELATED WORK

Dinesh, in his research, carried out automatic detection and extraction of locations from accident events sourced from news reports in Norway [7]. The accident data was sourced from the Accident Investigation Board of Norway (AIBN) report. He also said that although reports related to accidents were also reported on the Twitter platform online and in real-time, the data could not be used in his research because the reported incidents were not good enough to describe the location of the incident. Therefore Twitter data was not suitable data for his research.

In his research, he used Named Entity Recognition (NER) using StanfordNLP to extract the locations of reported events, but the model used was not specific to regions in Norway, so errors or omissions could occur during location extraction. The task of extracting label names from an entity is constantly creating new methods, tools (hardware), and publications of articles [8], [9]. In conclusion, this study focuses on three important tasks, namely: news classification, whether it is news about road accidents or not sourced from Google News, extraction of accident locations using Tagged Posts and NER, and mapping the coordinates of the accident location on the map accurately using the Google Maps API. and MySQL Spatial Reference (Dinesh, 2016).

Then Basuki et al. conducted research on the classification of conventional crimes such as immorality, psychotropic narcotics, non-crime, arson, murder, kidnapping, theft, persecution, embezzlement, hit-and-run, and anarchic demonstrations, using data from social media Twitter [10]. From there, data was obtained in the form of tweets from Twitter users in which there were sentences containing elements of the crime. The method used for classification uses the Naive Bayes Classifier, with 2 different datasets, namely a dataset containing lexical features or bag of words and a dataset containing syntactic features. With a model accuracy of 88.1398% for datasets with syntactic features and an accuracy of 79.25% for datasets with lexical or bag of words features. In addition, this study also extracted the location of the crime using the Named Entity Recognition (NER) method, with an accuracy of 65%. Although in other studies, SVM has shown significant results for the NER extraction task [11], Dakka and Cucerzan demonstrated that SVM achieved a level of f1-score of 0.954 for location entities using a Wikipedia article, and a measure of f1-score of 0.884 in all NER classes. At the same time, it uses a conditional random field (CRF) for named entity recognition (NER) which achieved an f-measure of over 85% for all named entities when tested on the CoNLL 2003 test data [12].

Kubler and Robert conducted a comparative study of some of the latest NER software, namely StanfordNLP, NLTK, OpenNLP, SpaCy, and Gate. Tests were carried out using the CoNLL2003 corpus and GMB. The results showed that StanfordNLP performed better than others (~30% better) on the CoNLL2003 dataset and 15% NLTK worse for tagging “Location” [13].

Drost et al. conducted a study to conduct crisis-related geocoding on the social media platform Facebook [14]. By

extracting and filtering from posts that appear on Facebook using Supervised Text Classification, and also extracting incident location information when there is no geotag feature on the post. This research is classified using the Decision Trees, Support Vector Machines, Naive Bayes, Multinomial Naive Bayes, and K-Nearest Neighbors methods, with the classification of "relief" and "non-relief". The geocoding process uses the free GeoNames database to match the extracted location coordinates using NLP.

The end result is a social media monitoring platform to respond to emergencies in times of crisis by helping to track the location of filtered Facebook posts that can help coordinate the process of volunteers assisting the rescue process.

Goldberg et al., explain in their article evaluated how county-based cancer rates are affected by the assignment of ZIP-code level incidence counts [15], [16], coming from the California Cancer Registry, tet a single accuracy value for a dataset throughout the entire area of coverage is a fundamental assumption on which most spatial health analyses are founded [17], but the measurement proved difficult and time consume [18].

This research is still continuing previous research [19], [20] to take advantage of conversations on the Twitter social media platform that can be used by journalists so that they can continue to update their information with issues that are developing in the community. A geocoding method was developed to map the location of the collected conversations so that journalists can more closely find out what news or events, or issues are developing around them. This requires a Spatial approach to be able to find and display the locations of these conversations.

Then the data collection approach proposed in this study, by using more monitored Twitter accounts, also by using a combination of keywords to stream Twitter conversations. With the dominance of these two methods, it is hoped that the data collected can be much more, and the crawling data system built can last a long time.

This study explores the LINGO Algorithm to classify text documents using latent semantic indexing. The topic cluster subfields include text mining, where large volumes of text are analyzed to find patterns between documents [21]. [22] conducted research on the classification of tweets into several categories. Several research approaches were inspired by Google News provided by Google. They used LDA and K-means to group tweet data.

III. PROPOSED METHOD

Determining the location of Twitter posts has its own challenges and things that require a lot of time and effort if done manually. By utilizing the power possessed by computer machines, it is not possible for the work to be carried out in real-time and produce broader insights. The next is how to determine in the topic what tweets are posted. This is needed so that understanding the current situation can be easier and faster because the tweets have been classified based on the topic.

In the process of collecting twitter post data, authors use the streaming data method using the Streaming API that has been provided by Twitter. It can easily collect Twitter posts by using keywords or with targeted Twitter, usernames to monitor and always get the latest tweets regularly, in near real-time.

In this study, an analysis of the classification of types of crime that occurs using the Naive Bayes method that has been used will be carried out using a collection of Twitter post data that has been collected for each topic. Topics include politics, law, economics, social & culture, health, natural disasters, sports, technology, defense & security, entertainment, automotive, culinary.

The NER method is used to extract location names in the form of names of provinces, districts, sub-districts/villages, street names, and other map objects so that the NER model is used specifically to adjust the administrative structure and map objects of Indonesia. For the geocoding process, a spatial reference query is made to the index engine, which contains spatial reference data for locations in Indonesia sourced from the Badan InformasiGeospasial (BIG) and OpenStreetMap (OSM).

Fig. 1 shows an overview of the research methods carried out in this study. A detailed explanation of each item is described in this subsection.

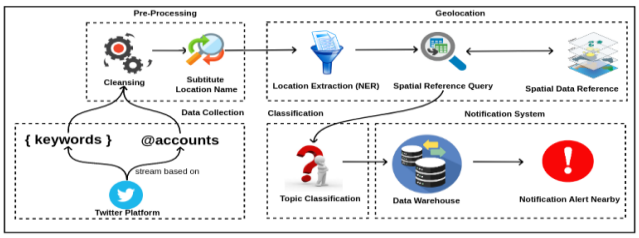

Fig 1: Methodology

A. Data Collection

This study uses data sourced from Twitter posts obtained using the API on the Twitter platform. The data collected is obtained by streaming on one of the Twitter APIs; the streaming parameters used are keywords and Twitter usernames. The collected Twitter posts are then used for various needs such as:

- As a training set for criminality classification
- As articles to be classified by topic
- As data for location analysis (geocoding) to find the location of Twitter posts around the user by using the search distance radius

As for the location data used for searching the location of the incident, open data provided by the Geospatial Information Agency (BIG) and OpenStreetMap (OSM) is used, from administrative area data (province, district/city, sub-district, kelurahan/village) to street names and map object.

Authors use some relevant keywords to get posts that may contain news or information related to an issue that is being raised as many as 307 keywords, including political parties, government, opposition, coalitions, conflicts, local elections, presidential elections, cases, corruption, criminal

acts, disputes, law enforcement, kpk, corruption, embezzlement, money laundering, bribery, and so on.

As for streaming using monitored Twitter usernames, there are at least 96 local Indonesian Twitter accounts that authors used for this research, including @kompascom, @detikcom,@beritasatu,@mediaindonesia, @tempodotco, @breakingnews, @tmcpoldametro, @mncnewschannel, @tribunindonesia, @cnnindonesia, and so on. By using mainstream media, Twitter accounts with the assumption that there will be many people who complain about an incident for the first time, one of which is to these mainstream media channels on the Twitter platform.

Besides that, using the accounts of influencers or social media practitioners who use the Twitter platform to convey their aspirations. This is also done with the assumption that an issue will become viral and increase in popularity if an influencer or other social media practitioner popularizes or deliberately raises the issue.

B. Pre-Processing

The pre-processing process is carried out to select the important features of a Twitter post that are taken, including removing unnecessary taglines such as the content of the Twitter post link or the attached news link, to adapt to the conversational style of the Twitter post, which is usually only as short sentences and briefly, sometimes abbreviations are also used and the use of location names that are commonly used in social language, it needs to substitute the name of the location used in a Twitter post, such as the use of the word West Java which refers to the West Java area, South Jakarta to refer to the South Jakarta area. Etc.

a) Cleansing

Before analyzing the tweet conversation, it is necessary to carry out several cleaning processes. This cleaning process is intended so that the resulting text is minimally noisy and so that the text is more uniform, making it easier to analyze for the engine by removing the symbol character (regex) or punctuation and link links that exist in the tweet text. Tweet text looks like below will be cleaned to make it more uniform.

"RT @GoRiauCom: DorongPercepatan Herd Immunity, Polda Riau VaksinRibuanMahasiswa dan Dosen di Unrihttps://t.co/GOs2ywbLby".

After the cleaning process, the tweet conversation will look like this.

"RT GoRiauComDorongPercepatan Herd Immunity, Polda Riau VaksinRibuanMahasiswa dan Dosen diUnri".

b) Substitute Location Name

Sometimes the post writer uses a figurative name or an abbreviation of the name of a location that he writes, such as the Kota Hujan, which refers to the Kota Bogor or writes Jabar for Jawa Barat. So a pre-processing process is needed to substitute the location name so that it becomes a location name that is understood by the model created. Besides that, it can also save data stored for spatial reference, for example, the name of the location of the Kota Hujan and Kota Bogor as different records, but this does not need to be done because it has made substitutions at this preprocessing stage.

"Dinilai Banyak Sumbangsih Di dunia Sepak Bola, KapolresTasikDidorong Jadi KandidatPengurus PSSI Jabar https://t.co/SCMdnc264S | Tasik Zone https://t.co/Iv50kdpkHa"

After the cleaning process and substitution process, the text will look like this.

"Dinilai Banyak Sumbangsih Di dunia Sepak Bola, KapolresTasikDidorong Jadi KandidatPengurus PSSI Jawa Barat Tasik Zone"

C. Geocoding

Work to link geocodes with geographic features apart from using addresses is also most often associated with services provided by gazetteers [23]. The problem is, the gazetteer does not have a function to generate geocode results. Instead, it acts as a storage mechanism after the geocode is determined using another method. As such, geocoders are commonly used to generate geocodes for features in address-based gazetteers, emphasizing the important relationship between the two components as part of a large spatial query and analysis framework where the geocoder can consist of many data sources for the gazetteer, which consists of several data sources.

Gazetteer is a dictionary or geographical directory as well as a reference data set to search for information about places and place names (toponyms) accompanied by a complete map or atlas. This Gazetteer contains a geographical description of a country, region, or continent to social statistics or landscapes, such as mountains, waters, or roads.

The output of this geographic reference is determined by the algorithm process to represent the input. In many situations, the output generated is in the form of simple geographic data such as points but is not limited to other types of valid geographic objects. The processing algorithm determines the geographic data to return for a given input based on its attribute values and attribute values in the reference data set. Which is by far the most complex of the geocoding processes that most other research is currently working on. The key to this geocoding problem topic consists of standardizing and normalizing input into a format and syntax that is compatible with the collected geographic reference data, matching algorithms that select the best query results from the reference data, and the final mechanism of geocoding generation that determines what will be returned based on features. Selected as the most suitable result. Deterministic algorithm processes can run using standardization, normalization, and attribute relaxation. In general, the key to these processes is to define each part of the input that comes in and convert it to a version of the data that is consistent with the reference data set [24].

After the incoming input is made compatible with the reference data set, the matching process will select the best candidate to use to determine the final output.

To get the coordinates of the location of the crime incident that occurred, there are several steps that must be passed in order to produce good results with a high level of accuracy. The location extraction from a news article is carried out; the locations to be extracted are the names of the provinces, districts/ cities, sub-districts, villages, street names, and map objects. From the extracted location, to perform a spatial reference query by utilizing the information structure that has been extracted, the query results are in the form of the location coordinates of the detected location names.

a) Location Extraction

Locations using Stanford NER using news articles. However, in this training process, the method used is a curative (traditional) method, which is labeling one by one word according to the entity classification, so that the machine can later extract the location entity from a tweet post.

StanfordNER is a licensed open-source software tool developed with the Java programming language and developed with feature extraction to detect entity names [25]. StanfordNER was trained using the English language CoNLL2003 training data. This tool also provides a common implementation of the Conditional Random Field (CRF) model, also known as the CRFClassifier.

For example, give a sentence like an example below.

"Bandung – PoldaJabarUngkapSindikatRelawanPembuatSertifikatVaksinPalsu https://t.co/9XQxwyxPyF"

Before labeling each word of the sentence, it is necessary to tokenize each word so that it becomes as shown in Table 1.

Table 1. Tokenization Label

Bandung	KAB
Polda	POI
Jawa	POI
Barat	POI
Ungkap	O
Sindikat	O
Relawan	O
Pembuat	O
Sertifikat	O
Vaksin	O
Palsu	O

b) Spatial Data Reference

The results of the NER extraction of the location obtained will be queried using the spatial data reference that has been collected (data source: BIG and OSM catalogs). So at the end of the geocoding process, it is to get the coordinates and the administrative boundary relation of the location of the crime in the news article.

c) Spatial Data Query

To search for location addresses (province, district, sub-district, kelurahan, street names, and map objects) spatially based on NER Location extracted from a news article. So it takes a reference location for searching (searching).

Then, after all the data has been collected, which consists of data from provinces, districts/cities, sub-districts, sub-districts/villages, street names, and map objects. So it is necessary to index these data into a data store that is stored on Apache Solr as a search engine location, which can be seen as Fig. 2.


```
{
  "code":"11",
  "latitude":3.909987,
  "kode_prop":"11",
  "type":"PROVINSI",
  "propinsi":"NANGROE ACEH DARUSSALAM",
  "name":"NANGROE ACEH DARUSSALAM",
  "longitude":96.73982,
  "id":"87434071-6d9b-46e2-b49d-94f2db6f96e5",
  "_version_":1601687603099205632},
{
```


Fig 2: Spatial Data Reference

D. Classification

To be able to label a Twitter conversation text with its topic classification, it is necessary to first build a classification model. This study uses the Naive Bayes method to build a topic classification model. The classification to be formed consists of 11 classes, which can be seen in Table 2.

Table 2. Topic Classification

No	Classification	Keywords
1	Politik	partaipolitik, pemerintahan, oposisi, koalisi, konflik, pilkada, pilpres
2	Hukum	kasuskorupsi, tindakkriminal, sengketa, penegakanhukum, kpk, kepolisian
3	Ekonomi	inflasi, resesi, hutang, pajak
4	Sosial&Budaya	intoleransi, penggusuran, pelestarianbudaya, kebhinekaan, kepercayaanmasyarakat, sosialisasi, penyuluhan, tradisi

5	Kesehatan	vaksin, virus, mutasi, covid, imunisasi, pandemi
6	BencanaAlam	banjir, gempabumi, tsunami, tanahlongsor, banjirbandang
7	Olahraga	olimpiade, pekan olahraga, sea games, sepak bola, bulutangkis
8	Teknologi	sains, penjelajahan, roket, bulan, mars, nasa, spacex, ponselpintar, chip, luarangkasa
9	Pertahanan&Keamanan	terorisme, radikalisme, separatisme, kkbpapua, lautcinaselatan, kapalasingmasuk
10	Hiburan	artis, gosip, film, drama, sinetron, box office, k-pop
11	Otomotif	esemka, toyota, honda, tesla, yamaha, suzuki, motogp, f1, moto2
12	Kuliner	makanan, minuman, resep, kuliner

Naive Bayes Classifier uses probability theory, and this method is used to classify text. There are two stages that occur in the text classification process; and the first stage is conducting training on a collection of documents (data set), the second is the process of classifying documents that have not been labeled with categories [26].

The method proposed by British scientist Thomas Bayes uses probability and statistics to predict future opportunities based on past experience, known as Bayes' theorem. By using this method, let's assume that the independence of each condition or event is very strong. Therefore this method is said to be naive.

E. Notification System

The notification system works at certain time intervals. This can help a journalist to stay updated on current issues around his location. Its location position is obtained with the help of GPS, which is integrated with the web application used. The web application will record the location of the last logged-in journalist and periodically record its location as long as the application is in use.

a) Clustering Tweets Topics

Tweets search results are basically only presented in the form of a long list sorted by their relevance to a given query operation. The search results, of course, ignore the proximity between tweets. Similarities between tweets can be presented through a clustering algorithm that groups tweets based on certain topics.

The function of clustering tweets is very important so that the system can determine what conversations are considered important, which have the potential to develop into viral issues. In this study, the authors use the number of documents collected in each topic cluster obtained, which is obtained after clustering selected conversational documents using a spatial query with a certain time span, according to the configuration of the notification scheduler that has been initialized.

The minimum number of documents (threshold) used in this study is 30, meaning that in a cluster group consisting of a minimum of 30 documents. All conversations contained in the topic cluster will be considered as important documents, while cluster groups that do not meet the threshold are not considered unimportant documents.

This algorithm consists of several stages. The first is preprocessing, which involves language identification, tokenization, and stemming. The second stage is feature extraction to determine the label candidate for the cluster to be formed. Next is the cluster label induction, where the label candidates are selected using dimension reduction. After the label candidates are selected, it is continued with the cluster content discovery stage to determine the group membership of each tweet.

IV. ANALYSIS RESULT

In this study, there are at least several important variables shown in Table 3.

Table 3. Variables

Variable	Value	Description
Number of Twitter Account	96	@kompascom, @detikcom,@beritasatu, @mediaindonesia, @tempodotco, @breakingnews, @tmcpoldametro, @mncnewschannel, @tribunindonesia, @cnnindonesia, etc
Number of Keywords	307	political parties, government, opposition, coalitions, conflicts, local elections, presidential elections, cases, corruption, criminal acts, disputes, law enforcement, kpk, corruption, embezzlement, money laundering, bribery, etc
Notification Interval	3 Hours	Issue detection process time span
Notification Time Range of Conversation	24 Hours	Conversation query time span between interval ranges

Time Period	Realtime	Twitter conversations are collected continuously, with streaming api
Threshold Number of Document in Cluster	30	The minimum number of documents in one cluster can trigger notifications.
Search Radius	50 km	Coverage of the search area for conversations from the user's position.

Table 4 shows the conversation gains that can be collected starting from 2021-05-14 to 2021-09-15; when compared to the previous study [19][20], it can be seen by using more Twitter accounts and merging with streaming using keywords. The data collected increases significantly and can take place non-stop.

Table 4. Tweets Distribution

No	Topic	Number of Tweets
1	hiburan	2.793.703
2	kesehatan	2.397.141
3	teknologi	2.097.294
4	kuliner	1.621.330
5	hukum	1.366.152
6	otomotif	1.147.346
7	politik	775.600
8	hankam	711.878
9	bencanaalam	584.372
10	sosialbudaya	492.195
11	olahraga	418.751
12	ekonomi	400.838
	Total	**14.806.600**

While for the acquisition of data collected on a daily basis, which can be collected on 2021-09-15, it can be seen as the acquisition of existing data, as shown in Table 5.

TABLE 5. Tweets Distribution At Certain Time

No	Topic	Number of Tweets
1	hiburan	75.306
2	Teknologi	55.565
3	kesehatan	52.554
4	kuliner	42.175
5	hukum	32.113
6	otomotif	29.891
7	politik	22.661
8	hankam	15.781
9	ekonomi	12.328
10	sosialbudaya	11.552
11	olahraga	10.721
12	bencanaalam	7.809
	Total	**368.456**

As an interface for the identification process of developing issues and the position of the location of the conversation that is spread within a radius of the user's position, all conversations that meet the predetermined variable criteria can be presented on a map display. As the first interface that is first seen, the user will be presented with all notification history data related to his GPS location history. This is done so that users can review their historical data.

Table 6. Geocoded Conversation Comparison

Status	Number of Tweets	Percentage
Total	14.806.600	
Geocoded	718.645	4,85%
Not-Geocoded	14.087.955	95,15%

From conversations that have been collected and geocoded analyzed. Several analysis results were obtained, as shown in Table 6; it can be seen that the percentage of conversation data that were successfully geocoded, with a comparison of the total population of the data.

And in Table 7, it can be seen the distribution of geocoding detection results at each regional level in the location extraction classes used.

Table 7. Geocoded Match Type

Level	Number of Tweets	Percentage
PROP	244.996	34%
KAB	342.517	47.7%
KEC	20.167	2.8%
KEL	17.672	2.4%
STREET	11.572	1.6%
POI	83.252	11.6%
	Total	**718.645**

In Fig. 3, it can be seen that one of the Twitter conversation cluster analyzes was obtained from 1.806 conversations. Seen in each topic cluster has a significant number where they have passed the minimum threshold for us to consider that the topic is considered important.

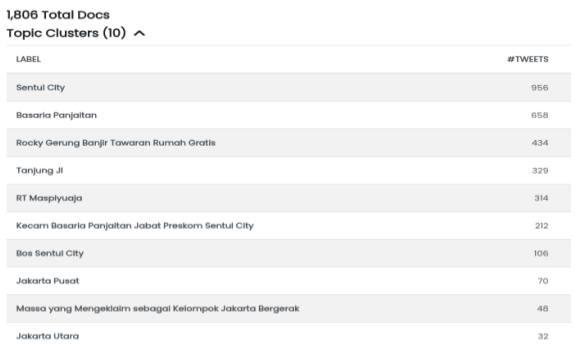

Fig 3: Topic Cluster

Include also a collection of conversations that are considered important, as shown in Fig. 4. This important criterion is based on the conversations that appear in each cluster. In order to avoid repeated information such as re-tweets, it can be solved by doing deduplication of conversations until there are only a few conversations that have been properly filtered.

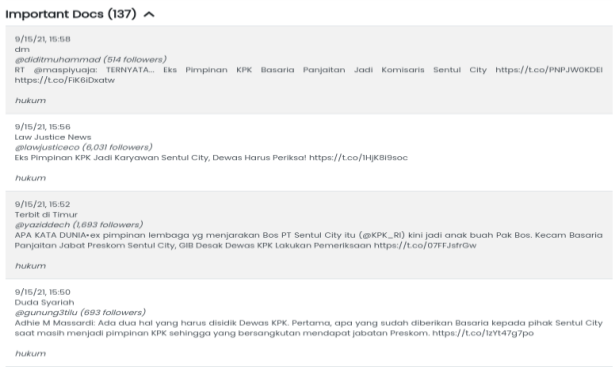

Fig 4. Important docs

It can be seen in Fig. 5, the distribution of conversation locations that were successfully mapped using the geocoding method. The search radius is represented as a red circle above the map.

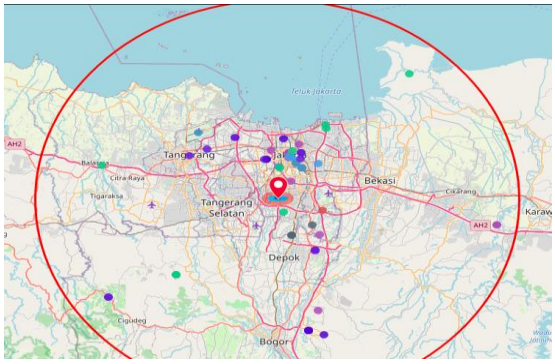

Fig 5: Conversation spread

The distribution of topics is represented by circles scattered on the map with different colors. It can be seen that the location of the conversation that is obtained can be properly queried based on the radius of the distance and the time interval that has been determined, as conversation detail can be shown as in Fig. 6.

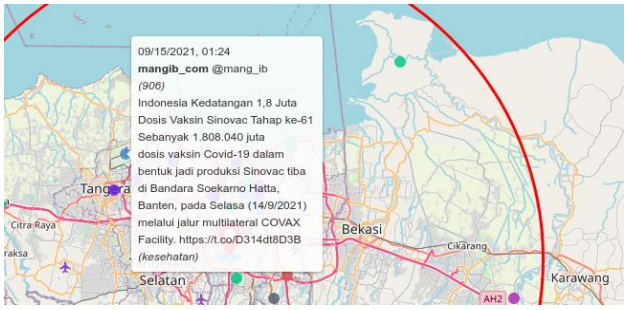

Fig 6: Conversation shown on maps

In Fig. 7, you can see the number of clusters on the map. In the Bogor area, more precisely in the Sentul City area, there are lots of conversations gathered in one location. This is because, at the time this manuscript was written, there was the issue of the eviction of an activist in Indonesia whose house is located in the Sentul City area. Lots of conversations refer to the location of Sentul City or mention the area as part of the conversation.

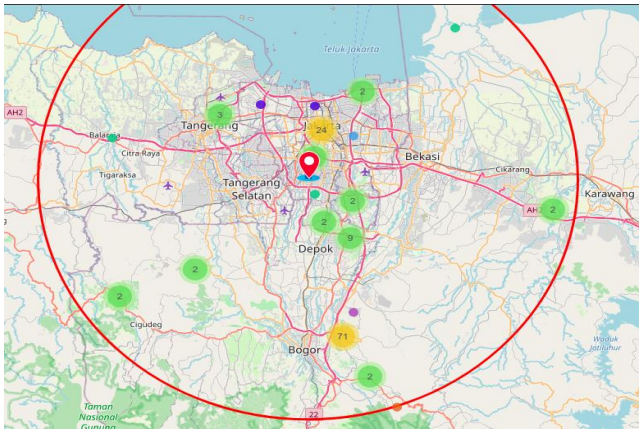

Fig 7: Clustered location on maps

In Bogor City (where Sentul City is located), it can be seen at one point in the same location, and there are a lot of conversations gathered. This is because all these conversations refer to the same location or use the same location name.

This also has a relationship with the formed topic cluster, where there are topics that have a number of conversations that have the same topic. If viewed from the point of view of location geocoding results, conversations that refer to a particular topic can also show adjacent locations and even refer to the same location.

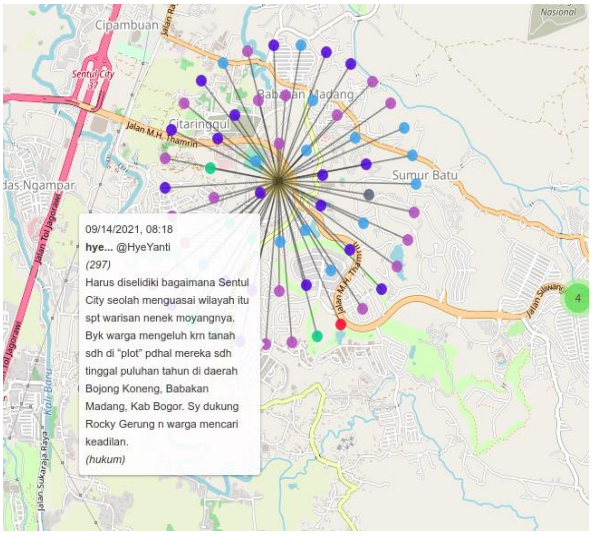

Fig 8: Focused clustered location

There is also a travel route facility that can be taken if the user wants to cover the location, which can be seen in Fig. 9.

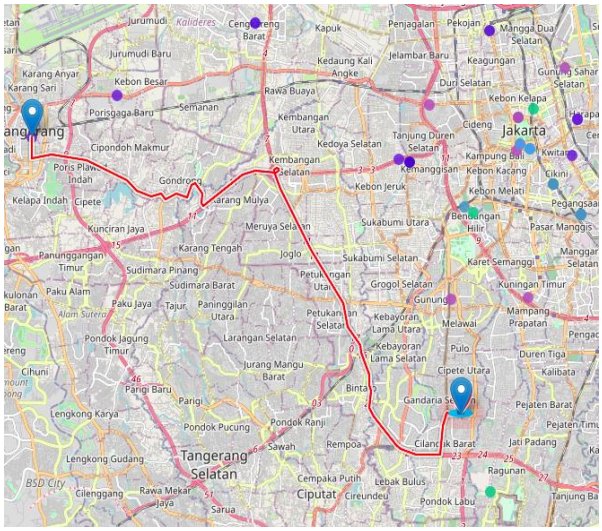

Fig 9: Route to tweets location

In Fig. 10, you can see an example of a conversation sample that was successfully geocoding to get the coordinates of the location of the conversation. There is the phrase 'Pakar Hukum Universitas Al Azhar Indonesia ...'.

Fig 10: Sample conversation location

The system can pinpoint the location or coordinates based on the university name information from the conversation. Referring to the name of the university and searching for it using the Google Maps search engine, it returns the location where the University is located, as shown in Fig. 11.

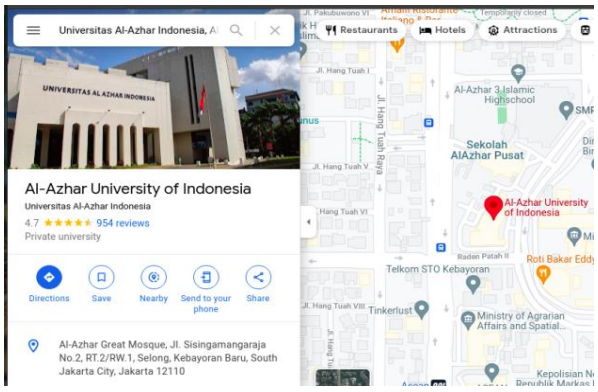

Fig 11: Google maps result for University Al-Azhar Indonesia

When using on the base map provided by OpenStreetMap on Fig. 12, which provide an additional overlay of the conversation locations, can be seen the location of Al-Azhar Indonesia University when compared to Google Maps search results, and the locations obtained by the geocoding method in this study, the level of accuracy of the designated location is not so far away.

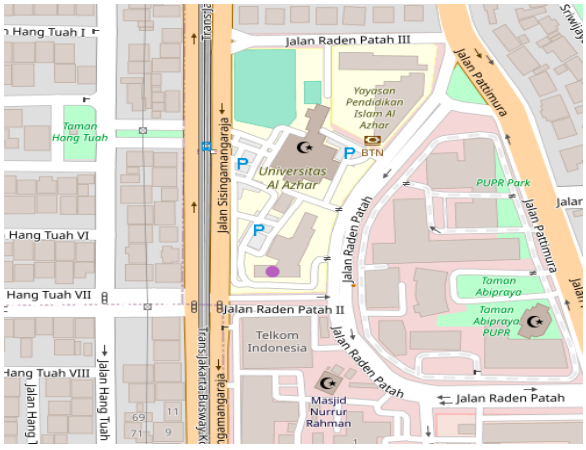

Fig 12: Maps view from OpenStreetMaps

At least it's still in the same neighborhood, on the same nearest protocol road, just different building blocks.

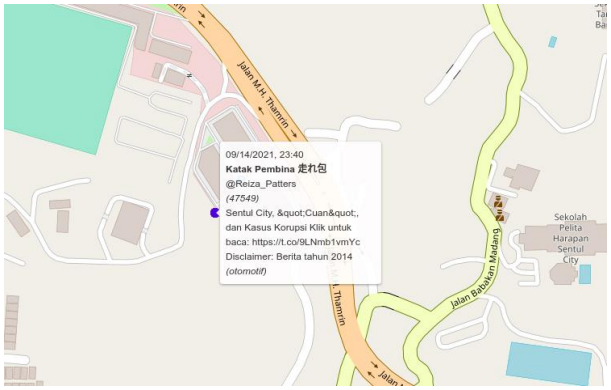

Fig 13: Geocoding result for referencing Sentul City

Then move to another data sample, namely the issue of evictions in Sentul City, which was currently hot when this manuscript was written. Fig. 13 shows the location designated by the geocoding system that was built. The location refers directly to a location called 'Pasar Sentul City', which seems inaccurate because it should refer to the residential area of Sentul City.

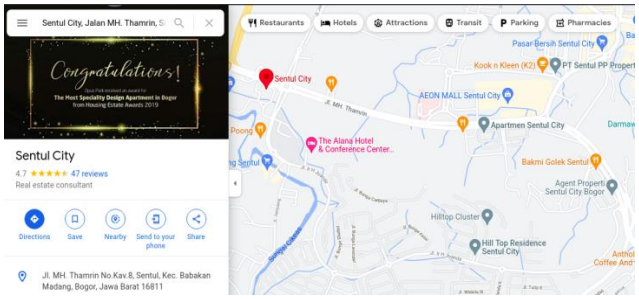

Fig 14: Location of Sentul City

The location referred to by geocoding is 'Pasar Sentul City', which is located not far from the actual area or location of Sentul City. The difference in distance is close to a distance of 2 kilometers.

This can be caused by the reference data set used for the spatial data query process. This means that there is no actual location with the name Sentul City, which can be seen in Fig. 15.


```
{
  "kabupaten":"KAB BOGOR",
  "name":"PASAR TRADISIONAL SENTUL CITY"},
{
  "kabupaten":"KAB BOGOR",
  "name":"RS PERTAMEDIKA SENTUL CITY-ER"},
{
  "kabupaten":"KAB BOGOR",
  "name":"RS PERTAMEDIKA SENTUL CITY"},
{
  "kabupaten":"KAB BOGOR",
  "name":"HARRIS HOTEL SENTUL CITY BOGOR"},
{
  "kabupaten":"KAB BOGOR",
  "name":"KAWASAN ARGENIA SENTUL CITY"},
{
  "kabupaten":"KAB BOGOR",
  "name":"KAWASAN ARGENIA SENTUL CITY"},
{
  "kabupaten":"KAB BOGOR",
  "name":"MASJID JAMI AL MUNAWAROH SENTUL CITY"}]
```


Fig 15: Query for Sentul City

As previously stated, the notification system will run every 3-hour interval. To inform users about the hottest and most discussed issues. The notification is obtained by the user in the form of an email, which will tell what conversations are being discussed a lot so that he can get an update as soon as possible and can consider whether the notification sent is only a false alarm.

Fig 16: Sample of the notification email

An example of an e-mail notification that is sent looks like Fig. 16.

V. CONCLUSIONS

With Twitter conversations that are streamed any time in real-time, a journalist can continuously find out what issues are developing around him with the help of the geocoding method. This also indicates that the data collected from the social media platform Twitter can be good enough for this geocoding process to be carried out.

The success in the geocoding process in this study at least depends on several parameters that are ideal for the system, such as: writing location names greatly affects the effectiveness of location name extraction using the NER model that was created, even though it has been trained with the characteristics of the Indonesian region, but the use of slang in showing names locations can be misinterpreted, this is also influenced by the punctuation included, so the separation of location names greatly affects the effectiveness of geocoding. Then the collection of spatial data used also affects the level of the match in finding the described location, as in the example that has been discussed.

Although the comparison of conversations that were successfully geocoded from the entire conversation population collected, which was only 4,85%, and the match rate for each regional level is occupied by the PROP (34%), KAB (47,7%), and POI (11,6%) levels in the top 3 ranks. This can help journalists in carrying out their duties and responsibilities to be at the forefront of reporting an event that is developing in the community.

ACKNOWLEDGEMENT

This work is supported by the Directorate General of Strengthening for Research and Development, Ministry of Research, Technology, and Higher Education, Republic of Indonesia, as a part of Penelitian DasarUnggulanPerguruan Tinggi Research Grant to Binus University titled “SistemPenugasanJurnalisBerdasarkanTrafik Media Sosial” with contract number: 064/E4.1/AK.04.PT/2021, 3530/LL3/KR/2021, 039VR.RTT/IV/2019 and contract date: July 12, 2021.

REFERENCES

[1] Rizaty. M. A, SiapaTokohTerpopuler di Twitter pada 2021?". [Online]. Available: https://databoks.katadata.co.id/datapublish/2021/07/09/siapa-tokoh-erpopuler-di-twitter-pada-2021. [Accessed: 07-Sep-2021].

[2] Sheela, L. A Review of Sentiment Analysis in Twitter Data Using Hadoop. International Journal of Database Theory And Application, 9 (2016) 77-86.

[3] Abbot, D. Introduction to Text Mining: Virtual Data Intensive Summer School. Abbot Analytics, Inc (2013).

[4] Han, Jiawei &Kamber, M. Data mining: concepts and techniques morgankaufmann. 54 (2006).

[5] Prilianti, K. R. & Wijaya, H. Aplikasi Text Mining untukAutomasiPenentuanTrenTopikSkripsidenganMetode K-Means Clustering. J. Cybermatika, 2(1) (2014) 1–6.

[6] R. M. Tripathy, S. Sharma, S. Joshi, S. Mehta, and A. Bagchi, Theme Based Clustering of Tweets, in Proceedings of the 1st IKDD Conference on Data Sciences, (2014) 1–5.

[7] Dinesh, S. Automatic Detection and Extraction of Event Locations in News Report to locate in Map. Master Thesis. (2016).

[8] Scharl, Arno and Klaus Tochtermann. The Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society. The Geospatial Web (2007): n. Pag.

[9] Y.-F. R. Chen, G. Di Fabbrizio, D. Gibbon, S. Jora, B. Renger, and B. Wei, Geotracker: geospatial and temporal rss navigation, in WWW '07: Proceedings of the 16th international conference on World Wide Web. New York, NY, USA: ACM, (2007) 41–50.

[10] Maghfiroh, Siti & Basuki, Setio&Azhar, Yufis. Klasifikasi Tweets TindakKejahatanBerbahasa Indonesia Menggunakan Naive Bayes. Jurnal Repositor. 2. 10.22219/repositor.v2i7.67. (2020).

[11] W. Dakka and S. Cucerzan, Augmenting Wikipedia with named entity tags, IJCNLP, (2008).

[12] J. R. Finkel, T. Grenager, and C. Manning, Incorporating non-local information into information extraction systems by Gibbs sampling, in ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Morristown, NJ, USA: Association for Computational Linguistics, (2005) 363–370.

[13] X. Schmitt, S. Kubler, J. Robert, M. Papadakis, and Y. LeTraon, A Replicable Comparison Study of NER Software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate, 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), (2019) 338-343, doi: 10.1109/SNAMS.2019.8931850.

[14] Drost, S., Wytzisk, A., & Remke, A. Geocoding of Crisis Related Social Media Messages for Assessing Voluntary Help Efforts as a Contribution to Situational Awareness. (2018).

[15] Goldberg D. A geocoding best practices guide. Springfield, IL: North American Association of Central Cancer Registries; (2008).

[16] Goldberg, D. The effect of administrative boundaries and geocoding error on cancer rates. Spat Spattemporal Epidemiol. (2012).

[17] Goldberg DW, Wilson JP, et al. An effective and efficient approach for manually improving geocoded data. Int J Health Geogr 7(60) (2008).

[18] Goldberg DW, Wilson JP, et al. From text to geographic coordinates: the current state of geocoding. URISA J , 19(1) (2007) 33–46.

[19] Girsang, Ganda & Isa, Sani &Harvy, Ikrar. Recommendation System Journalist For Getting Top News Based On Twitter Data. Journal of Physics: Conference Series. 1807. 012006. 10.1088/1742-6596/1807/1/012006. (2021).

[20] A.S. Girsang, S.M. Isa, Natasya, M.E.C. Ginzel ,Implementation of a Journalist Business Intelligence in Social Media Monitoring System, Advances in Science, Technology and Engineering Systems Journal, 5(6) (2020) 1517-1528.

[21] Godfrey, D., Johns, C., Meyer, C., Race, S. &Sadek, C. A Case Study in Text Mining: Interpreting Twitter Data from World Cup Tweets. Arxiv Preprint Arxiv:1408.5427 (2014).

[22] K. Dela Rosa, R. Shah, B. Lin, A. Gershman, and R. Frederking, Topical clustering of tweets, Proc. ACM SIGIR SWSM, (2011).

[23] Hill, Linda. Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. 10.1007/3-540-45268-0_26. (2000) 280-290.

[24] Christen P, Churches T, Willmore A. A probabilistic geocoding system based on a national address file. Proceedings of the Australasian Data Mining Conference: Cairns, AU. (2004).

[25] Finkel, J., Grenager, T., & Manning, C. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference. 10.3115/1219840.1219885. (2005).

[26] Rish, Irina. An Empirical Study of the Naïve Bayes Classifier. IJCAI 2001 Work Empir Methods ArtifIntell. 3. (2001).

[27] Takalikar, Mukta &M.Kshirsagar, Manali & Singh, Kavita. Pattern-based Named Entity Recognition using context features. International Journal of Computer Sciences and Engineering. 6. 365-368. 10.26438/ijcse/v6i4.365368. (2018).