INTERNATIONAL JOURNAL
ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Social Media Engineering for Issues Feature Extraction using Categorization Knowledge Modelling and Rule-based Sentiment Analysis

M Tafaquh Fiddin Al Islami[a,*], Ali Ridho Barakbah[a], Tri Harsono[a]

[a]*Department of Information and Computer Engineering, Politeknik Elektronika Negeri Surabaya, Jl. Raya ITS,Surabaya, 60111, Indonesia*
Corresponding author: tafaquh.fiddinal@gmail.com

***Abstract*— A company maintains and improves its quality services by paying attention to reviews and complaints from users. The complaints from users are commonly written using human natural language expression so that their messages are computationally difficult to extract and proceed. To overcome this difficulty, in this study, we presented a new system for issues feature extraction from users' reviews and complaints from social media data. This system consists of four main functions: (1) Data Crawling and Preprocessing, (2) Categorization Knowledge Modelling, (3) Rule-based Sentiment Analysis, and (4) Application Environment. Data Crawling and Preprocessing provides data acquisition from users' tweets on social media, crawls the data and applies the data preprocessing. Categorization Knowledge Modelling provides text mining of textual data, vector space transformation to create knowledge metadata, context recognition of keyword queries to the knowledge metadata, and similarity measurement for categorization. In the Rule-based Sentiment Analysis, we developed our own rules of computatioal linguistics to measure polarity of sentiment. Application Environment consists of 3 layers: database management, back-end services and front-end services. For applicability of our proposed system, we conducted two kinds of experimental study: (1) categorization performance, and (2) sentiment analysis performance. For categorization performance, we used 8743 tweet data and performed 82% of accuracy. For categorization performance, we made experiments on 217 tweet data and performed 92% of accuracy.**

***Keywords*— Issues feature extraction; categorization knowledge modelling; context recognition; rule-based sentiment analysis.**

Manuscript received 9 Sep. 2020; revised 31 Dec. 2020; accepted 8 Jan. 2021. Date of publication 31 Mar. 2021.

I. INTRODUCTION

Social media is an online tool used by people to interact with each other, and exchange information and mutual ideas. It is also the third largest type of internet content which have users account for as many as 129.9 million people 97.7% of the total 132.7 million internet users in Indonesia [1]. This huge amount is supported by the advantages of social media using mobility and web-based technology to create an interactive media where individuals and communities can share, create, discuss and modify the content. It introduces and changes the way communication between organizations, communities and individuals. This huge amount and benefit has resulted in the interaction of the Indonesian people becoming very high on social media platforms. The interaction of Indonesian people on social media also attracts many other parties to analyze the inter-community interaction which is one of the considerations for decision making [2], especially for airlines.

In addition to being used as an object of social media analysis by airlines, airline users also use social media to provide feedback and complaints about airline services or products and even everything about the airline. Airlines can know the quality of their products or services from this interaction, making it easier for airlines to evaluate and develop their services or products. Number of responses and complaints from users on social media has increased. It is no longer possible for airlines to analyze responses by looking at social media data one by one, so it takes a lot of time and human resources to analyze thousands of tweets from customers on social media. The length of time makes airlines must have to wait longer to implement the results of response analysis on social media and hamper airline business decision making [3]. In addition, Indonesian language on social media also has obstacles that are increasingly degraded every day [4]. This degradation has an impact on the use of good and right

Indonesian discussion into an informal Indonesian language. Shortening habits when writing notes on daily activities carried over to social media. This causes sound changes and meaning in language.

Hashimoto, Kuboyama and Chakraborty [5] conducted a study on extraction of topics from twitter social media data using Singular Value Decomposition and Feature Selection which claimed their methods were more precise and faster to obtain topics from very much social media data. Tikawa and Nagayoshi [6] presented a study that groups twitter data based on the similarity of the text model in a political context so that political group networks can be known among one another. Jotikabukkana, Sornlertlamvanich , Manabu, and Haruechaiyasak [7] proposed a new method for classifying social media text using an improved trained text model. Purwarianti, Andhika, Wicaksono, Afif, and Ferdian [8] proposed a processing toolkit for natural language in Indonesian that contains many natural language processing modules. A. Erianda and I. Rahmayuni [27] bring off research on how to improve the classification of Twitter and email data using naive bayes. Their research concluded that word noise can reduce the accuracy of the classification system. The noise word consists of the miss spelled word and the white space converted into another symbol [28]. M. Zulqarnain [29] conducted research on text classification based on word embeddings. Khan and Urolagin [9] conducted research with airline objects to research, measure and predict consumer loyalty based on analytical sentiment. Wan and Gao [10] conducted a study by raising airline objects to calculate the sentiment value of airlines using several sentiment classification methods. M. Kamal, A.R. Barakbah, and N.R. Mubtadai [11] applied a new approach to the analysis of opinions on AFTA using opinion mining based on the analysis of temporal sentiments. Putra, Helen, and Barakbah [12] conducted a study to measure sentiment values from Surabaya community comment data on Surabaya City Government services.

Liu [15] proposed a method called Context-Based Term Frequency that will process text based on context with the rules of context correlation positively or negatively. Liu [16] also presented a technique called CRHTC (context recognition for hierarchical text classification) that performs hierarchical text classification by recognizing the context of discussion (COD) of each category. Context recognition in computational language is a method for recognizing languages based on contexts [17] that are the focus of language computing. In order to recognize the language itself there are several kinds of word features to recognize. One of them is syntax which is a grammar or rule that regulates the relationship between words or letters in constructing a sentence or word. In this study, context recognition is used to adjust the relationship between words. Between words between tweets with other tweets must have relevance and can be grouped with one another. Grouping is done so that airlines can more easily understand the specific issues that are being discussed a lot by airline users who express their expressions on social media twitter platforms. For the problems raised there are very few studies that raise the issue of airlines. There is research on airlines by Khan and Urolagin [9] raising the object of tweets of airlines but the purpose of their research is to measure and predict airline user loyalty based on analytical sentiment. There was also research by Wan and Gao [10] about sentiment analysis on airlines using several sentiment classification methods. Tiwari et al [23] conducted research on sentiment analysis on airline twitter datasets using the BIRCH clustering method and association rule mining.

II. DATA

In this section we discuss the data that will be used. Where the data is obtained, how much and what will be used will be discussed in this section.

A. Dataset

Data sources are the data that will be used in this study both as training and testing data. Data sources were obtained from two sources, namely, twitter social media and raw data from the Indonesian NoLimit company.

Data from social media twitter is taken from airline social media accounts and social media user accounts that do mentions to airline accounts. Every data taken from these two accounts is legal. Basically, each account and account posting is legal to see and retrieve data because by default it is public (can be accessed by everyone who has a Twitter account). However, it cannot retrieve or view data on social media user accounts that are private or have been arranged. The amount of data that has been obtained is 14000 twitter data.

TABLE I
AIRLINES ACCOUNT

No	Airlines Name	Twitter Account
1	Garuda Indonesia	@IndonesiaGaruda
2	Air Asia	@AirAsia_indo
3	Batik Air	@BatikAirINA
4	Citilink Air	@Citilink
5	Sriwijaya Air	@SriwijayaAir
6	Lion Air	@LionAirID

Table I contains lists of airlines and their social media accounts which are first data source for this study. This data is taken from Twitter. The total data is 136713 tweet data, but only around 9000 tweet data are used in the experiment to develop the system. The data from the NoLimit Indonesia company was obtained by the author during the internship period at the company. The author has also asked permission from the company and is allowed too. The category that authors get when had intership and the company's report to ailines, evolves after author observing more than 80000 tweets data based on the author's assumptions. From these two data, author found many new keywords and categories. These data will also test in the categorizer experiment to see the accuracy of data categorization.

In addition to the two data above, there are special data to assist in the process of sentiment analysis called dictionary data. This data is temporarily obtained from previous research in the Rule-based Sentiment Degree Measurement of Opinion Mining of Community Participatory in the Government of Surabaya [9].

B. Preparing The Data

Data processing is a process of data to be processed to produce research supporting data such as test data, category data and keyword data. Before processing data, you need to understand about airlines. What issues are needed by the

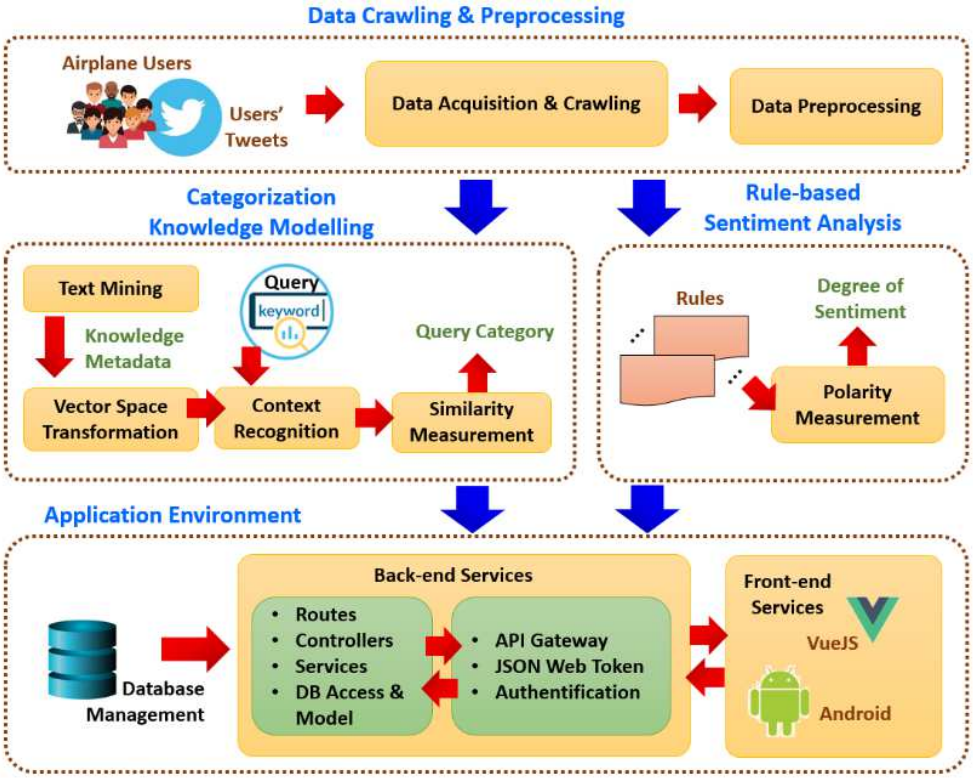

Fig. 1 System Design

airline, so that after knowing what is needed we can make the appropriate category. The appropriate category works so that the data classification does not widen and remains in accordance with the airline environment. To determine the category the author uses two reference sources, namely from the Indonesian NoLimit company and the airline's social media data.

III. Materials and Methods

The system to be built in this study is to make easier for airlines to get information on issues that exist on their social media. The system will automatically process social media data obtained from social media sites using the categorization and analysis sentiment features.

Fig 1 shows the system built in this study. The first part is data crawling and preprocessing. This first sub-system functions to retrieve twitter data periodically every week on airlines' twitter accounts in Table I. These accounts are open so it is legal for the general public including researchers to retrieve their data from twitter. The Twitter platform provides open API (Application Programming Interfaces) to the public without any payment. Because of it, access to the Twitter API is limited. We only can retrive twitter data in recent past week. So we must get twitter data periodically. The second part is categorization of knowledge modelling. This sub-system serves to categorize twitter data. Category labels on twitter data help airlines categorize the issues being discussed on their twitter. The third part is rule-based sentiment analysis. Categories of airline twitter data are processed and analyzed in this sub-system to obtain the sentiment value of each data. The sentiment value consists of three ranges of values, -1, 0, and +1. A value of -1 indicates that the tweet has negative sentiment. A value of 0 indicates that the tweet has neutral sentiment, does not have any emotions. A value of +1 indicates that the tweet has a positive sentiment. The fourth part is application environment. The fourth sub-system is divided into three part, namely front end, back end-services, and database. Front end is a system that can be run by users or people from airlines to see the final results of the application. Front end consists of web based and mobile based. Web based uses the Vue js front end JavaScript framework, while mobile based uses Android technology. Back end is a service API (Application Programming Interface) serves to manage communication from the front end to the database. The database uses the type of NoSQL database based on the JSON format, MongoDB. The database stores the results of crawler operations when retrieving twitter and user data during registration and also changes in user data on the application.

A. Data Crawling and Preprocessing

A crawler is a system for retrieving data from the Twitter Developer API. In this crawler program uses and modifies the twitter API usermentions and usertimeline. First, the crawler uses the twitter API usermention. The Twitter API can retrieve tweets based on twitter account parameter. For example, if you want to retrieve tweets from Garuda Indonesia, the parameters used are "IndonesiaGaruda". As a result, all tweets that contain mentions to @IndonesiaGaruda twitter account are taken. The twitter API system uses a crawlback system. Crawl back is the collection of twitter data in sequence from the present to the past. The twitter account used by researchers can be backed up for the past week. Therefore, the crawler module on the server is scheduled to run once a week.

The usermention API crawler module has a weakness. The weakness is that when a user replies to an airline's tweet and

use does not include mentions to the airline's account, the tweet from that user cannot be retrieved. Actually, this weakness has been overcome by Twitter by including automatic mentions. However, when researchers conducted a crawler in 2018, this feature was still missing. Therefore we use the usertimeline API module. All tweets addressed to airlines will be covered by the usertimeline API module. The results are very satisfying. However, there is a data redundancy problem. We use the filter provided by the mongodb command to filter duplicate data and remove duplicates.

Following is the basic process of the text mining pre-processing sanitizing process, selecting steps according to need [22].

1) *Select document scope*
   The document scope in this study is one tweet. Each tweet will be considered one document. Each document will be processed separately from other documents.
2) *Case folding*
   Equalize letters so that when processing text, there is no difference between letters in the same word (case sensitive).
3) *Tokenization*
   Break up text into collections of words or tokens. This process can have many types in doing the process, depending on the language to be analyzed. For this research we break up a collection of words with blank space, dot, and commas.
4) *Stop words*
   Words that are considered not important. Stopwords deletion serves to clear text of words that are not important. It would be very useful in text mining to get rid of some words such as "which" and "with" which appear in almost every document with a large number.
5) *Punctuation*
   Punctuation is punctuation such as commas, periods, exclamation points, question marks. Not all punctuation is deleted because it will affect the processing of the sentiment analysis system that requires punctuation to break sentences and or phrases.

TABLE II
SPECIAL PUNCTUATION LIST

Name	Symbol
At sign	@
Hastag	#
Comma	,
Dot	.

Table II shows the punctuation which received special treatment. For punctuation at sign and hashtag will be deleted not only punctuation, but along with the words that follow. For example @garudaindonesia and #garuda, these two words will be completely deleted and will not be used in the next system process. The dot will not be erased. The dot will be used to break up tweets containing multiple sentences such as paragraphs. The comma will not deleted due to 8th rule of sentiment analysis. Punctuation other than table II, will delete the symbol only.

6) *Stemming tagging*
   Stemming is the process for mapping and deciphering the form of a word into its basic word form or in other words the process of changing the word affixes into basic words.

B. Categorization of Knowledge Modelling

After all processing on the preprocessing data is complete, the tweet is clean and ready to be processed. The tweet will then be referred to as a tweet query. In the categorization of knowledge modeling sub-system, query tweets are converted into vectors. Each word of the tweet query is matched against every keyword in the category keyword list. The list of keywords and categories are already defined before. The list of keywords and categories are define manually by the author after observing all airlines tweet and suggestion from NoLimit Indonesia company.

After that, the frequency for each matched word is calculated. Then a vector is created from the query tweet with the keyword feature that matches the word in the tweet query. Features are chosen from only suitable words because they apply context recognition. So that words that do not match the category keyword list are ignored. This makes vector calculations in programs faster.

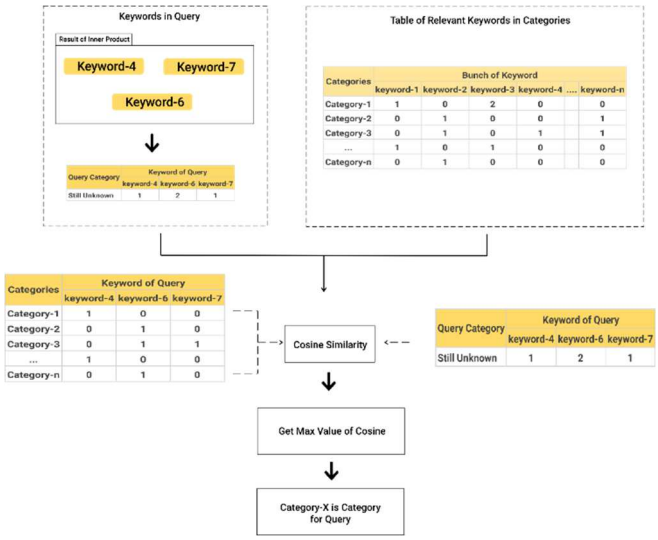

Fig. 2 Categorization System

Fig 2 shows the list of keywords per category is changed to a vector matrix [17]. The number of word features in the category vector matrix will be adjusted to the tweet query vector matrix. The number of word features in the category vector matrix will be adjusted to the tweet query vector matrix. For example, the tweet query matrix vector have features of keyword-4, keyword-6, and keyword-7. The category vector matrix will also have features of keyword-4, keyword-6, and keyword-7. The number of word features in the category vector matrix will be adjusted to the tweet query vector matrix. For example, the tweet query matrix vector features keyword-4, keyword-6, and keyword-7. The category vector matrix will also feature keyword-4, keyword-6, and keyword-7. So, the program will always change the feature in the vector matrix category to match the vector matrix matrix of the tweet

being processed. Both matrices will be operated using cosine similarity to calculate the proximity of the two matrices with equation 1 [14].

$$similarity = \cos(\theta) = \frac{A.B}{||A||.||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

After the query tweet matrix is compared to all categories of matrices, the largest value is taken, which means that the largest value is the most category matrix representing the tweet query. If there is no match or there are no keywords in the tweet, it will be categorized as a "nocategory" category which means it does not have a category.

C. Rule-based Sentiment Analysis

Sentiment analysis used in this study is the analysis of rule-based sentiment. The researcher provides several rules for analysing word type combinations in social media content. Each word will be checked part-of-speech or the type of word to find out whether there is a series of words that are in accordance with the rule. A series of words that are in accordance with the rule will be aggregated first and calculated the value of sentiment in accordance with the rule. After there are no words in accordance with the rule, then each value of the word in the content will be added up and the result is the sentiment of the content. A value greater than zero is a positive sentiment, a zero value is a neutral sentiment, and a value less than zero is a negative sentiment. The rule can define by sum of value form each word sentiment value and compare the result [18][19][20][21]. Previous research also consider the part-of-speech divide into adjective[18][19][20][21], verb[18][19][20][21], adverb[18], intensifier[18], preposition[19][20], symbol[19][20], modal operators[22] and modifier[22]. The sentiment value of sentence also consider the word location on the sentence[18][19][20][21], and it uses the certain rule-based. The following part-of-speech rules are found in the sentiment analysis in this study [12].

1) Single Adjective

If a tweet query contains adjectives without being followed by verbs and prepositions, then the tweet query sentiment value is the same as the adjective sentiment value. For example, the phrase "*saya marah*" means “I am angry”. *Saya* (I) word is not a rule because *saya* word class is a noun. Then the word *marah* (angry) will be proceed because *marah* (angry) word is a adjective. So, the example sentence has one adjective word. The word angry has the value of sentiment -1. Value -1 means the example sentence is a negative sentence.

2) Single Verb

If a tweet query contains verbs only, then the tweet query sentiment value is the same as the verb sentiment value. For example, the phrase "*maskapai merugikan saya*" means “airlines adverse me”. The *maskapai* (airline) word has no meaning and its word class is a noun. The *merugikan* (adverse) words is a verb. The *saya* (me) word is a noun. The *merugikan* word has sentiment value of -1. While the *maskapai* word is not in the dictionary of sentiments and the *saya* word have a class of nouns, so the representation of sentiments of the sentence is -1.

3) Adverb and Adjective

Table III shows and operator table for adverb and adjective sentiment value calculation. If there are adjectives word after the adverb word, then the sentence value is calculated by rule AND operator.

TABLE III
AND OPERATOR

Adverb	Adjective	Value
1	1	1
1	-1	-1
-1	1	-1
-1	-1	-1

For example, the sentence "*Ngurus refund kok sangat lama ya min? bisa di jelaskan?*" means “Admin, why take care of refund took a very long time? Can you explain?”

The *sangat* (very) word is a adverb word. Its followed by the *lama* (long time) word which is a an adjective word. While the other words are not meet the sentiment rule, we will ignore the others word. The *sangat* word has sentiment value of +1. The *lama* word has sentiment value of -1. The calculation method is simple, we just need to input the two sentiment values in the and operator. Sentiment values +1 and -1 if entered in the and operator according to table III, then the sentiment value of *sangat lama* (very long time) phrases is -1. Because of the others word have no sentiment value, the sentiment value of the example sentence is -1.

4) Adverb and Verb

If there is a verb word after the adverb word, the sentence value is calculated by the AND operator. For example the sentence "*Delay lagi, delay lagi. Kenapa maskapai tidak mengerti kebutuhan orang ya*" means “Delay again, delay again. Why don't airlines understand people's needs? “. The tidak (*not*) word is a adverb word that has a sentiment value of -1, whereas the *mengerti* (understand) word is a verb word that has a sentiment value of +1. While the other words are not meet the sentiment rule, we will ignore the others word. The calculation method is simple, we just need to input the two sentiment values in the and operator. Sentiment values -1 and +1 if entered in the and operator according to table III, then the sentiment value of *tidak mengert* (not understand) phrases is -1. Because of the others word have no sentiment value, the sentiment value of the example sentence is -1.

5) Verb and Adjective

Table IV shows XOR operator table for adverb and adjective sentiment value calculation. If there are adjectives word before the verb word, then the sentence value is calculated by rule XOR operator.

TABLE IV
XOR OPERATOR

Verb	Adjective	Value
1	1	1
1	-1	-1
-1	1	-1
-1	-1	1

For example, the sentence "*Maskapai ini terbang dengan buruk. Badan pesawat seolah getar mau jatuh*" means “This airline flies badly. The fuselage seemed to be shaking to fall”.

The *terbang* (fly) is a verb word that has a sentiment value of +1, while the *buruk* (bad) word is an adjective word that has a sentiment value of -1.

The *jatuh* (fall) word is a verb word that has sentiment value of -1. The phrase *terbang dengan buruk* (flies badly) is the fifth rule. The *dengan* word is stop word, it will delete in pre-processing. The calculation method for this phrase is simple, we just need to input the two sentiment values in the XOR operator. Sentiment values +1 and -1 if entered in the XOR operator according to table IV, then the sentiment value of *terbang dengan buruk* (flies badly) phrases is -1. Also, we need to add the *jauh* word sentiment value. So, (-1) + (-1) = -2. The sentiment value of the example sentence is -2. It indicates that the example sentence is a negative sentence.

6) Adverb, Verb, and Adjective

Sometimes adverb, verb and adjectives appear together in a sentence. To calculate sentiment, first the system calculates the verb and adjectives using XOR operators. Second, the results of the first calculation and the adverb sentiment value are calculated using AND operators.

For example, the phrase "*Maskapai tidak paham dengan baik, apa sebenarnya kebutuhan pennumpang*" means "Airlines do not understand well what the passengers need".

The *tidak* (not) word in the example sentence above is a adverb word with sentiment value of -1. The *paham* (understand) word is a verb word with sentiment value of +1. The *baik* (good) wordis an adjective word with sentiment value of +1. Because the three words fit into the sixth rule, program will calculate verb and adjective word first in the XOR operator. The phrase *paham dengan baik* (understand well) will have sentiment value of +1. Then we calculate the phrase sentiment value with adverb sentiment value. So, *tidak paham dengan baik* (not understand well) sentiment value is -1. Because of the others word have no sentiment value, the sentiment value of the example sentence is -1.

7) All symbols on tweets are deleted except for dot, comma, exclamation mark and question mark

Dot, comma, exclamation mark and question marks are not deleted. Dot, exclamation mark and question mark will be changed to comma. So that, in the 8th rule the sentence is more easily separated only by the comma parameter.

8) Dot and comma as parameters for separating sentences

Every tweet query usually contains several sentences separated by commas and dots. For example "*delay lagi nih, apa gaada kemajuan?*" The sentence will be separated into two sentences. First is "*delay lagi nih*". Second is "*apa gaada kemajuan?*" The sentiment value of the first sentence is -1 while the second sentence is -1. To calculate the sentiment value of the two sentences, then add the value of the second sentence (-1) + (-1) = -2 (negative).

9) Word phrases consist of two words that have one meaning that cannot be separated

10) The value of the "thank you" or praise sentiment depends on the position of the word in the sentence.

If '*terima kasih*' is in front of the sentence, then the sentiment value of the word '*terima kasih*' is 1, but if the word 'terima kasih' is in the middle or at the end of the sentence, the sentiment value of '*terima kasih*' is 0. Example "*terima kasih garuda sudah tidak delay lagi*".

D. Application Environment

The application environment is divided into three part, namely front end, back end-services, and database. Front end is a system that can be run by users or people from airlines to see the final results of the application. Front end consists of web based and mobile based. Web based uses the Vue js front end JavaScript framework, while mobile based uses Android technology. Back end is a service API (Application Programming Interface) serves to manage communication from the front end to the database. The database uses the type of NoSQL database based on the JSON format, MongoDB.

1) Front End

The front end is created using javascript, html and css. In this study, we use javascript library called Vue js. Vue js was chosen because it is very light and very easy to develop. Many display and chart libraries that support the Vue js framework. Vue js are also very modular with component based components.

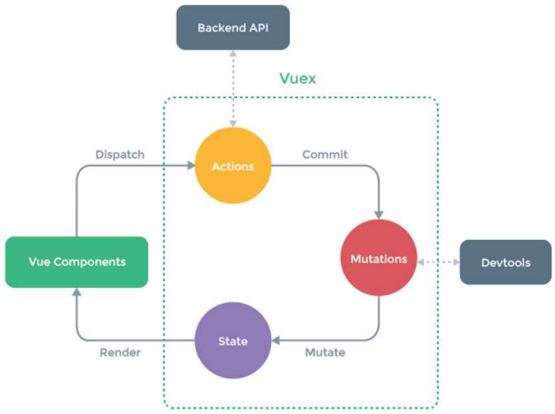

Fig. 3 Vue js lifecycle

Fig 3 [30] shows the life cycle of Vuex with back end and vue js [24]. When the Vue JS component, simply a small part of the Vue JS web page, requires data to be visualized, the component can retrieve directly to the state. State is data storage model on Vuex. According to best practice in vuex development, it is better to use the getters method provided by Vuex to retrieve data from state. State will call back end API service to retrieve data from the database.

2) Back End

Back end is a system for serving user interaction on the front end with the database using the express js library. For authentication, back end services use a JSON Web Token so that communication with the database is safe because the database cannot be accessed by users who are not registered to the application.

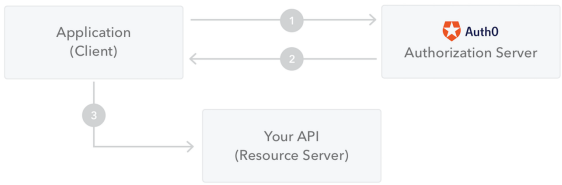

Fig. 4 JSON Web Token Diagram

Fig 4 [31] explains that the client or application first makes an authentication request to the authentication server. When

the authentication is given by the server, the server will return the access token to the application. The application will use this access token to access resources that are protected by JWT. If you do not have this access token, the application cannot use the protected resource.

After authentication is successful, the request will be forwarded to the back end. Request data will be directed by the route module to find out which controller to use. The controller contains logic modules to process data before it is returned to the requesting application. The controller will call the services module to retrieve and process data according to predetermined logic. After the data is successfully fetched and processed, the data will be returned to the requesting application.

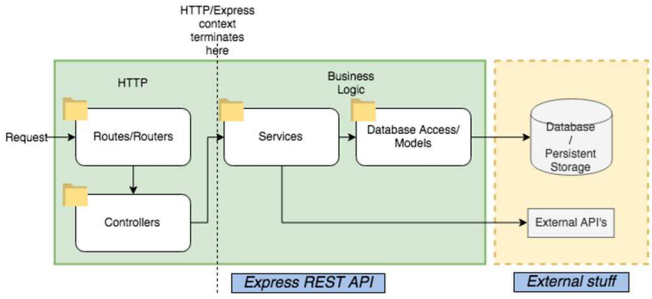

Fig. 5 Diagram Express JS

Fig 5 [32] shows an overview of the communication flow diagram between requests from outside applications, internal express js, and the database. For this research, external APIs as shown in the picture are not used or not applied. External APIs use the API from twitter. The API will be processed using a crawler program that is separate from the back end services.

3) Database

The database in this study uses Mongo DB. Mongo DB is a document-based NoSQL database using the JSON format. In the SQL database, the data is stored in the form of tables. In the NoSQL-based Mongo DB, the data is stored in a document using the JSON format. Mongo DB was chosen because of its reliability in storing data without having a relationship with one data collection to another. Mongo DB is also very appropriate to use as a medium-scale storage. Complete documentation and queries that can help in analyzing social media in this study are another advantage of Mongo DB. In addition, the data processing in Mongo DB is easy to use, implement and synchronize with back end services.

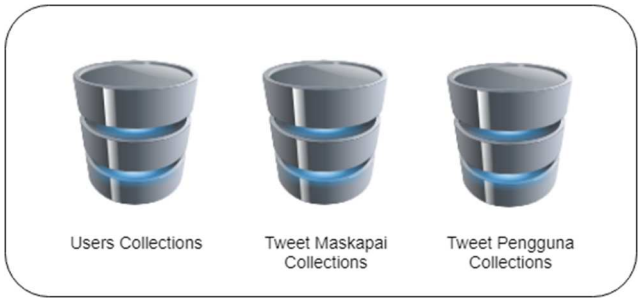

Fig. 6 Mongo DB Data Collection

Fig 6 shows that there are three collections used, namely the user collection, the airline collection, and the airline user collection. User collection contains user data. The intended user is an application user from the airline. Only one airline user per airline. However, one user can register more than one twitter social media account.

Fig 7 is an example of a user object in the JSON format. The object contains six fields. The first is the object to identity field. The first field is twitter_object, which contains the airline's social media accounts. The third and fourth fields are username and password. The fifth and sixth fields are the password and API auth version.

Fig. 7 User Object Example

The airline collection is a collection of tweet objects in JSON format from an airline account. Collection of airline users is a collection of tweet objects in JSON format from airline user accounts that interact with airline accounts either through direct mentions or replies to airline account tweets.

IV. Result And Discussion

We developed the data crawling, the categorization of knowledge modeling method and rule-based sentiment analysis with experimental research methods. The trial starts from the initial hypothesis of the study until the last hypothesis is carried out.

A. Data Crawling and Preprocessing

The database used in this study uses a NoSQL-based database, namely MongoDB. MongoDB was chosen because it has a document storage format in the form of JSON. JSON stands for JavaScipt Object Notation. This JSON format is the default format for return data from Twitter APIs retrieved by a crawler program.

Fig. 8 Document structure in JSON format

Fig 8 shows the JSON-shaped document structure on mongodb. The document structure in mongodb is very simple, it only consists of the document object id that is accommodated in the "_id" and sub-document attributes. Sub-documents can contain single values or contain objects. Sub-documents containing objects can freely contain any object, from simple objects, nested objects, to objects that refer to

objects from other documents. Objects that refer to other documents usually have another document's object id attribute.

We will do a test in database. The test is try to find and remove redundant data from the crawler program. This test is conducted because when checking the data, there is a tweet that has the same id_tweet with the same timestamp, and contains the same object. Therefore, it is necessary to filter redundancies twitter data based on id_tweet which is owned by the tweet object. There are two collections that store tweet objects, namely airline collection tweets and collection user tweets. Collection airline tweets contain tweets from the airlines' own accounts, while collection user tweets contain tweets from airline user accounts.

For the experiment, the aggregation method provided by Mongodb will be used. The method name is aggregate. This aggregate method has several stages which have several functions to carry out the data modeling logic. These stages act as parameters to be passed to the aggregate method. The contents of the stage are objects. The stages used in this experiment are stage group, sort, replaceRoot, and out [25]. We did two experiments with the same stage but with differences in the second experiment using the allowDiskTrue command field.

TABLE V
TESTING TO SOLVE REDUNDAN DATA PROBLEM

No	Code
1	db.tweet_pengguna.aggregate([{"$sort": {"_id": 1}}, {"$group": {"_id":"$id_str","doc": {"$first":"$$ROOT"} } }, {"$replaceRoot": {"newRoot": "$doc"} }, {"$out": "tweet_pengguna"}])
2	db.tweet_pengguna.aggregate([{"$sort": {"_id": 1}}, {"$group": {"_id":"$id_str","doc": {"$first":"$$ROOT"} } }, {"$replaceRoot": {"newRoot": "$doc"} }, {"$out": "tweet_pengguna"}], { allowDiskUse: true})

Table V shows code used to aggregate twitter data. There are two code. First code is the first attempt. The first attempt failed because the data in the tweet_user collection was too large. User data tweet has reached tens of thousands of data objects around 136731 data. The aggregate method cannot process if the data is too large because there is a limitation of 100 MB RAM memory usage [26]. Second code is the second attempt. Second attempt has been successfully solve redundan data with large memory usage problem. To solve the problem of limited memory used by the Mongodb database, it is necessary to add one command from Mongodb, namely allowDiskTrue. This allowDiskTrue command allows stage aggregate to write data to temporary files. So that memory usage is more efficient and the data processing process is not directly processed as a whole but is processed separately. The purpose of being processed separately is that the data will be divided into several parts. The parts of the data will be processed one by one and the results will be stored in temporary files. Then temporary files will be further aggregated if needed. And so on so that the RAM memory problem can be resolved.

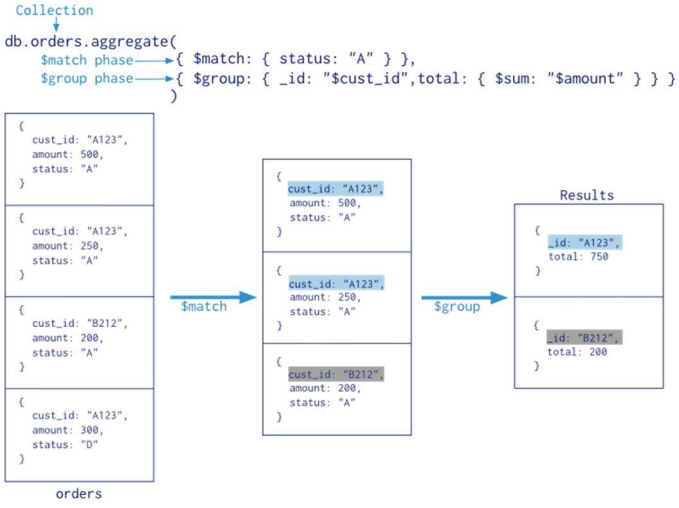

Fig. 9 Aggregate data simulation on mongodb

Fig 9 [33] shows an example of data aggregation simulation in Mongodb. All data will be processed according to the defined stages. The stages will be processed in sequence. Researchers still cannot understand the sequence of working on this stage, which stage will be processed first in the Mongodb documentation.

B. Categorization of Knowledge Modelling

Table V shows the results of categorization trials on 8743 social media data. Data of 8743 tweets were obtained from the results of crawling programs on the server. From the results of the trials in Table V obtained an accuracy of 82.64898%.

TABLE VI
FIRST EXPERIMENT RESULT OF CATEGORIZATION

Category Label	Fail	Success	Grand Total
Application	58	711	769
Bagage	114	2023	2137
Passangers Data	48	1060	1108
Delay	20	715	735
Facilities	13	472	485
Price	1244	627	1871
Staff	14	1515	1529
Refund	6	103	109
Grand Total	**1517**	**7226**	**8743**
Accuracy		**82,64898**	

Table VI shows the results of categorization trials on 8743 social media data. Data of 8743 tweets were obtained from the results of crawling programs on the server. From the results of the trials in Table V obtained an accuracy of 82.64898%.

Of the 8743 data entered in the category, there are a number of typos that were intentional or not by the user but were not detected by the program. The word intentional typos are like replacing letters with (*) stars, with the intention of censoring the real word, while unintentional typos are purely typographical errors from airline users.

TABLE VII
MISSTYPED FOUND

No	Keyword	Misstyped Keyword
1	*Aplikasi* (application)	applikasi
2	*Traveloka* (travel apps)	travel*ka
3	*Bawaan* (luggage)	bawa
4	*Bagasi* (bagage)	begasi
5	Booking	bookinh, boking, boking
6	Reschedule	re-schedule, rescheduling, schedule
7	*Tambah* (add)	nambah
8	*Cepat* (fast)	cepet, zepat
9	Delay	dwlay, dellay
10	*Lama* (long time)	lamma
11	*Harga* (Price)	hrga
12	*Makanan* (food)	makan
13	*Murah* (cheap)	mura
14	*Mahal* (expensive)	muahaaal
15	Refund	redund, fefund
16	*Kru* (crew)	krunya

Table VII shows the results of finding typos. Discovery of typos that were not detected by the program.

C. Rule-based Sentiment Analysis

Table VIII shows the experimental results of sentiment analysis from real twitter data. The sentiment analysis is still using the dictionary as a basis sentiment value of a word. This dictionary will be a reference to rule-based system. Program also have been applied all rules. The analysis sentiment has been able to detect phrase words like "*terimakasih*" means "thank you" as mention in section III on 10th rule of sentiment analysis. The phrase "*terimakasih*" should be considered as one word but we also discover that phrase "*terimakasih*" also typed in form "*terima kasih*". The "*terima kasih*" phrase will detected that phrase as two words. In the first hand, program cannot detect this case. The program has been updated to detect the word thank you as one word or two words. The program can also detect whether the word thank you is in front, in the middle, or at the end of a sentence. The only thing that sentiment analysis programs haven't solved yet is detecting non-indonesian-language words. However, this is ignored because the sentiment analysis program has produced very satisfying accuracy. Rule-based sentiment analysis experiment are using 217 sample data tweets taken randomly. Retrieval of this tweet is done randomly both for sentiments worth -1, 0, or 1. From the experiment, the accuracy is 92%. The next experiment will compare the accuracy of the rule-based sentiment method with sentiment classification.

TABLE VIII
RESULT OF SENTIMENT EXPERIMENT

No	Tweet	Sentiment Value
1	*@IndonesiaGaruda harga tiket garuda thn 2019 kok **mahal** 2 kali lipat ya min*	-1
2	*@IndonesiaGaruda Untuk bagasi bisa check in through?*	0
3	*Min, pada rute penerbangan cgk hkg connect dps itu kan conect timenya 9 jam. Saya kalo keluar bandara untuk menginap dihotel boleh tidak?*	0
4	*@IndonesiaGaruda Harga tiket walaupun promo sama aja, **mahal** min.*	-1
5	*@IndonesiaGaruda **Baik**, terima kasih atas informasinya min*	1
6	*@IndonesiaGaruda Alhamdulillah saya november kemarin msh pake rute solo-aceh-jeddah, pilihan rute yg **bagus** jika memungkinkan pesawat memakai 777 dengan konfigurasi duduk 3-4-3 sehingga seat per mile nya bisa lbh **murah** dan jamah lbh banyak.....*	1
7	*@spectatorindex Menjadi yang **terbaik** kedua amongst the other airlines is something to be proud of! Well done @IndonesiaGaruda*	1

Table VII shows the experiment on sentiment classification. Sentiment classification is an analysis sentiment based on the pattern of words in the tweet sentence or in other words the words are the features of the tweet. First the tweets are preprocessed to clean, the stopword removal process, the removal of non-alphabet characters, stemming and lemma are removed. After the data is clean from the preprocessing process, the data is converted as a vector with the label. Vector is processed using the KNN (K-Nearest Neighbor) algorithm model. Calculation of accuracy to measure the accuracy of the model used two cross-validation methods namely K-Folds and Algoithm LOO (Leave One Out).

TABLE IX
RESULT OF SENTIMENT CLASSIFICATION

No	Evaluation Method	Accuracy
1	K-Folds	66.67%
2	LOO Algorithm	70.83%

The difference in accuracy between rule-based sentiment analysis that has been developed with sentiment classification is clear. The rule-based sentiment analysis program succeeded in producing an accuracy of 92%.

D. Application Environmen

Applications are divided into two types, backend-services and front-end. Backend-services is a program to manage the traffic between the front-end communication with the database and also to query data such as, aggregate, find, and count the data. Backend-services use the nodejs library named expressJS and Mongoose. Front-end is a program to visualize data from backend-services and databases. Front-end is built using the VueJS framework.

Fig 10 is a part of the page to see details of positive sentiment for each category of airlines. Users simply click on the category section on the pie chart and a sample tweet will

come out automatically right below the graph. The graph also displays information on the number of positive tweets in each airline category. Airlines get a broader picture of categories that have positive sentiment, such as what airline users respond to airline issues which are divided into eight categories. The airline can also see the username of the airline user who is tweeting. Sometimes there are some famous Figs who tweet about airline issues.

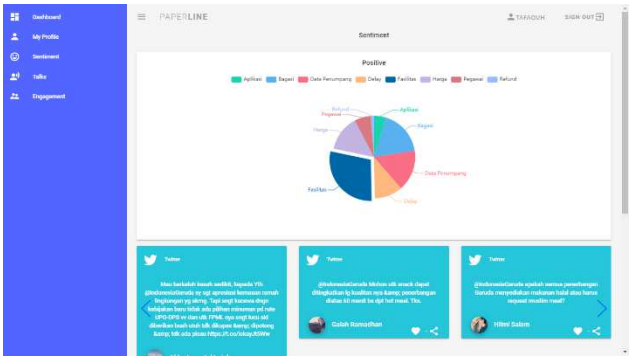

Fig. 10 Positive sentiment page

Fig 11 is the page section to see details of negative sentiments of each category on airlines. Airlines can see examples of tweets for each category that has negative sentiment. The function is the same as in Fig 10.

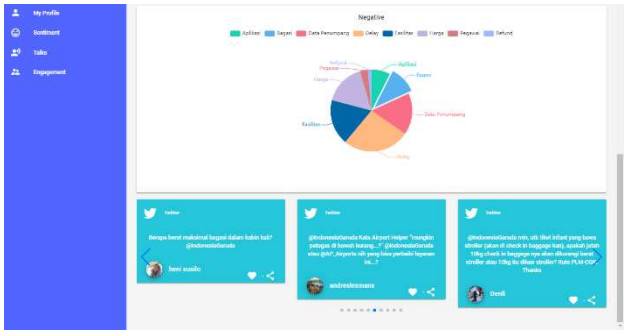

Fig. 11 Negative sentiment page

Fig 12 is part of the talks page. This page contains talk information from airline users. There is a graph of the number of talks with the airline from January to July 2019.

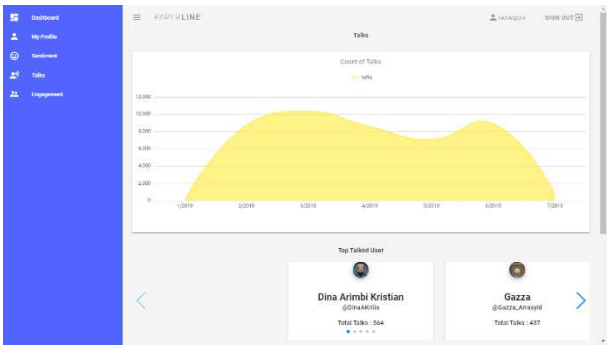

Fig. 12 Talk page

Fig 13 is also part of the talks page. The section in this picture shows airline users who have a lot of talks on social media about the Garuda Indonesia airline. This part of the picture also shows the most popular tweets. The tweet parameter is considered popular is to have a high engagement value, namely the most likes and retweets. There are different colors to indicate the sentiment value of the popular tweet. Red for negative, dark for neutral, and green for positive.

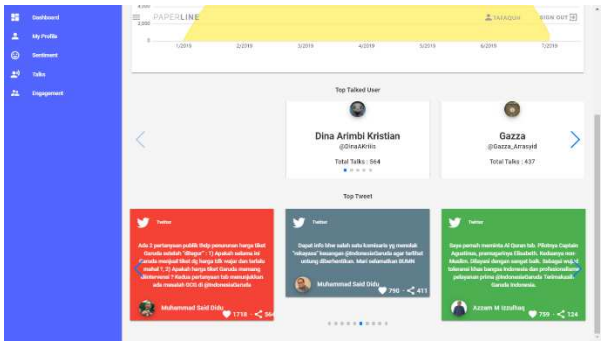

Fig. 13 Talk page 2

Fig 14 is an airline user engagement information page. The information is in the form of devices or devices used, what application languages are used, and from which locations users tweet their tweets. Information in the form of any device used can be used by airlines to find out what platform variations are needed by airlines to serve airline users. For example, as shown in Fig 4.11, dominant airline users use Android devices more than 50%, then airlines must prioritize their service applications on the Android platform and so on for other platforms. This relates to the application category. With this application platform priority the opinion of airline users about airline applications is getting better.

Fig. 14 Engagement page

V. CONCLUSIONS

In this study, the author presents a system of extraction of airline issues by grouping and calculating the value of sentiment on airline social media. This system consists of 4 main functions: (1) data crawling and preprocessing, (2) categorization knowledge modelling, (3) rule-based sentiment analysis, and (4) application environment. data crawling and preprocessing provides data acquisition from users' tweets on social media, crawls the data and applies the data preprocessing. Categorization Knowledge Modelling provides text mining of textual data, vector space transformation to create knowledge metadata, context recognition of keyword queries to the knowledge metadata, and similarity measurement for categorization. In the Rule-based Sentiment Analysis, we developed our own rules of computatioal linguistics to measure polarity of sentiment. Application Environment consists of 3 layers: database management, back-end services and front-end services. For applicability of our proposed system, we conducted two kinds

of experimental study: (1) categorization performance, and (2) sentiment analysis performance.

Data crawling and preprocessing run very well on the server. On the future, we need to code some bash programming to automate the redundant filter. For the categorization of social media twitter data, the accuracy is quite high. However, there is a case when a tweet has the same value for each category in the categorization process. The computer will select the categories that were detected earlier than other categories when the comments have the same value in more than one category. There are also obstacles in detecting spelling errors and miss typed words. Social media users use more devices than laptops or computers [1]. Do not close the possibility of using a laptop or device can still unconsciously do a typo, especially the device with a smaller keyboard size. Spelling and miss typed correction are our future task to do to improve our categorization knowledge modeling.

Sentiment of analysis needs to be done in-depth observation of sentence patterns on social media. The addition of new rules is also important because the language of social media is always developing and some are not in accordance with the rules of Indonesian grammar.

Application environment run excellent with some additional social media analytics. This application still cannot manage multi twitter account in one user. So in the future, user and twitter account management must conduct to improve and satisfy company satisfaction.

ACKNOWLEDGMENT

We would like to thank NoLimit Indonesia, a social media monitoring company for assistance in assisting data and processing data to determine categories.

REFERENCES

[1] S. M. Metev Tim APJII, “Buletin APJII: Saatnya Jadi Pokok Perhatian Pemerintah dan Industri”, Asosiasi Pengguna Jasa Internet Indonesia, Jakarta, 2016.

[2] K.C.B. Wicaksono, "Mengukur Efektivitas Social Media Bagi Perusahaan", Binus University, Jakarta, 2013.

[3] S. Vinerean, I. Cetina, L. Dumitrescu, and M. Tichindelean, “The Effects of Social Media Marketing on Online Consumer Behavior,” International Journal of Business and Management, vol. 8, no. 14, Jun. 2013.

[4] Suyatno, “Bahasa Indonesia sebagai Sarana Pengembangan Guru Profesional”, Orasi Ilmiah Ilmu Pendidikan Bahasa, Universitas Muhammadiyah Prof Dr Hamka , 2009.

[5] T. Hashimoto, T. Kuboyama, and B. Chakraborty, “Topic Extraction from Millions of Tweets using Singular Value Decomposition and Feature Selection, Proceedings of APSIPA Annual Summit and Conference 2015”, Hong Kong, 2015.

[6] H. Takikawa and K. Nagayoshi, “Political Polarization in Social Media: Analysis of the Twitter Political Field in Japan”, 2017 IEEE International Conference on Big Data (BIGDATA), USA, 2017.

[7] P. Jotikabukkana, V. Sornlertlamvanich, O. Manabu, and C. Haruechaiyasak, “Social Media Text Classification by Enhancing Well-Formed Text Trained Model”, ITB Journal Publisher, Indonesia, 2016.

[8] A. Purwarianti, A. Andhika, A.F. Wicaksono, I. Afif, and F. Ferdian, “InaNLP: Indonesia Natural Language Processing Toolkit Case Study Complaint Tweet Classification”, Institute of Electrical and Electronics Engineering, 2016.

[9] R. Khan and S. Urolagin, “Airline Sentiment Visualization, Consumer Loyalty Measurement and Prediction using Twitter Data”, International Journal of Advanced Computer Science and Applications, 2018.

[10] Y. Wan and Q. Gao, “An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis”, 15th International Conference on Data Mining Workshops: Institute of Electrical and Electronics Engineers, 2015.

[11] M. Kamal, A.R. Barakbah, and N.R. Mubtadai, “Temporal Sentiment Analysis for Opinion Mining of ASEAN Free Trade Area on Social Media”, IES: Knowledge Creation and Intelligent Computing (KCIC), 2016.

[12] B.J.M. Putra, A. Helen, and A.R. Barakbah, “Rule-based Sentiment Degree Measurement of Opinion Mining of Community Participatory in the Government of Surabaya”, EMITTER International Journal of Engineering Technology, Indonesia, 2018.

[13] G. Miner, J. Elder, A. Fast, T. Hill, R. Nisbet, and D. Delen, “Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications”, Academic Press, United States of America, 2012.

[14] W. Ford, Numerical Linear Algebra with Applications using Matlab. Elsevier Inc. First Edition, 2014

[15] R.L. Liu, “Context-Based Term Frequency Assessment for Text Classification” in PRICAI 2008: Trends in Artificial Intelligence, Springer Berlin Heidelberg, 2008, pp. 1004–1009.

[16] R.L. Liu, “Context recognition for hierarchical text classification” Journal of the American Society for Information Science and Technology, vol. 60, no. 4, pp. 803–813, Apr. 2009.

[17] G. Katz, B. Shapira, N. Ofek, Y. Bar-Zev, and I. Negev, “CoBAn: A Context Based Approach for Text Classification”, J. Inf. Sci.: Int. J. Arch., 262(March), pp.137-158, 2014.

[18] K.Z. Aung and N.N. Myo, “Sentiment Analysis of Students’ Comment Using Lexicon Based Approach”, International Conference on Computer and Information Science, pp. 149-154, Wuhan, 2017.

[19] R. Asmara, A. Basuki, and M.H.U. AlRasyid, “Gender Based Temporal Sentiment Analysis in Indonesian on Culinary Places in Surabaya City”, International Journal of Engineering and Technology Innovation, Vol. 7, No. 4, 2017.

[20] A.R. Naradhipa and A. Purwarianti, “Sentiment classification for Indonesian message in social media”, International Conference on Electrical Engineering and Informatics, pp. 1-5, Bandung, 2011.

[21] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, “Sentiful: A Lexicon for Sentiment Analysis”, IEEE Transactions on Affective Computing, Vol. 2, No.1, pp. 22-36, 2011.

[22] Delen, Dursun et al, Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications, Academic Press, United States of America, 2012.

[23] P. Tiwari et al., “Sentiment Analysis for Airlines Services Based on Twitter Dataset,” in Social Network Analytics, Elsevier, 2019, pp. 149–162.

[24] Vue JS, Y. Evan, *What is vuex?*, Feb. 2016. Accessed on: Jun. 23, 2020. [Online]. Available: https://vuex.vuejs.org/.

[25] MongoDB Inc, *Aggregation Pipeline Limits*. 2008. Accessed on: March 23, 2020. Available: https://docs.mongodb.com/manual/core/aggregation-pipeline-limits/#agg-memory-restrictions.

[26] MongoDB Inc, Database Collection Aggregate. 2008. Accessed on: March 23, 2020. Available: https://docs.mongodb.com/manual/reference/method/db.collection.aggregate.

[27] A. Erianda, & I. Rahmayuni "Improvement of Email And Twitter Classification Accuracy Based On Preprocessing Bayes Naive Classifier Optimization In Integrated Digital Assistant," JOIV : International Journal on Informatics Visualization, vol. 1, no. 2, , pp. 53-56, May. 2017

[28] Clark, Alexander & Tim, Issco. (2003). Pre-Processing Very Noisy Text.

[29] M. Zulqarnain, R. Ghazali, M. G. Ghouse, dan M. F. Mushtaq, “Efficient processing of GRU based on word embedding for text classification,” JOIV, vol. 3, no. 4, Nov 2019

[30] Evan You, Vue JS [Online]. Available: https://vuex.vuejs.org/, Accessed on: August 8, 2020.

[31] JSON Web Token [Online]. Available: https://jwt.io/introduction/, Accessed on: August 8,2020.

[32] Corey Cleary, Project structure for an Express REST API when there is no "standard way" [Online]. Available: https://www.coreycleary.me/project-structure-for-an-express-rest-api-when-there-is-no-standard-way/, Accessedon August 8, 2020.

[33] Vaibhav Sharma, Aggregation on MongoDB, Medium.com [Online]. Available: https://vsvaibhav2016.medium.com/aggregation-in-mongodb-4f638df0add0, Accessed on: August 9, 2020