TWITTER SENTIMENT ANALYSIS OF POST NATURAL DISASTERS USING COMPARATIVE CLASSIFICATION ALGORITHM SUPPORT VECTOR MACHINE AND NAÏVE BAYES

Ainun Zumarniansyah[1]; Rangga Pebrianto[2]; Normah[3]; Windu Gata [4]

[1,2,4]Computer Science
[3]Informatics Engineering
Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri
www.nusamandiri.ac.id
[1]14002362@nusamandiri.ac.id; [2]14002396@nusamandiri.ac.id; [3]normah.nor@nusamandiri.ac.id;
[4]windu@nusamandiri.ac.id

(*) Corresponding

Abstract— Natural disasters trigger people, especially Twitter users to provide information or opinions in the form of tweets. The Tweet can be an expression of sadness, concern, or complaint. Processing of data from these tweets will create trends that can be used for information needs such as education, economics, and others. Natural disasters are events that threaten human life caused by nature, including in the form of earthquakes. The method used is the Support Vector Machine and Naive Bayes from the tweet. The data collected is filtered from tweets by deleting duplicate data. In calculating the Natural Disaster sentiment analysis using a comparison of the Support Vector Machine and the Naive Bayes algorithm, the difference in accuracy is 3.07% where the results of the Support Vector Machine are greater than Naive Bayes. The purpose of this research is to analyze sentiment for the distribution of disaster aid that does not flow information due to information & coordination in the field. so as to provide information on the location of natural disasters, natural disaster management, and its presentation to victims that can be shared evenly in an efficient time due to information and natural management so that the distribution of aid is hampered

Keywords: Naive Bayes, Natural Disasters, Support Vector Machine, Twitter.

Abstrak— *Bencana alam memicu masyarakat khususnya pengguna Twitter untuk memberikan informasi atau opini dalam bentuk tweet. Tweet tersebut bisa menjadi ekspresi kesedihan, kepedulian, atau keluhan. Pengolahan data dari tweet ini akan menciptakan trend yang dapat digunakan untuk kebutuhan informasi seperti pendidikan, ekonomi, dan lain-lain. Bencana alam adalah peristiwa yang mengancam kehidupan manusia yang disebabkan oleh alam, termasuk berupa gempa bumi. Metode yang digunakan adalah Support Vector Machine dan Naive Bayes dari tweet. Data yang dikumpulkan disaring dari tweet dengan menghapus data ganda. Pada perhitungan analisis sentimen Bencana Alam menggunakan perbandingan Support Vector Machine dan algoritma Naive Bayes didapatkan selisih akurasi sebesar 3,07% dimana hasil Support Vector Machine lebih besar dari Naive Bayes. Tujuan penelitian ini adalah analisis sentimen untuk penyaluran bantuan bencana yang tidak mengalir karena informasi & koordinasi di lapangan. sehingga memberikan informasi lokasi bencana alam, penanggulangan bencana alam, dan penyajiannya kepada korban yang dapat dibagikan secara merata dalam waktu yang efisien karena informasi dan pengelolaan alam sehingga penyaluran bantuan terhambat.*

Kata Kunci*: Bencana Alam, naive Bayes, Support Vector Machine, Twitter.*

INTRODUCTION

In this digital era, humans entered a new lifestyle that is not detached from the electronic devices, very important digital forensure that everyone both in the village and is equipped with digital information and communication skills (Gallardo, 2020). The growing need for Data and information encourages people to develop new technologies so that data processing and information can be done easily and quickly, one of them in terms of using social media applications. The use of social media applications among the community has become an important part of everyday life and as if interactions have been moved into a virtual platform.

Social Media in question is Twitter. Twitter is one of the social media that allows its users to share information with others at all times. Information shared on Twitter is commonly referred to as a tweet consisting of 140 characters

displayed on the Twitter User profile page having unique writing characteristics and formatting with symbols or special rules (Antinasari, Perdana, & Fauzi, 2017). It only took less than three years, referring to the latest U.S. global firm predictions, We Are Social, last February, Indonesian Internet users in 2020 reached 175.4 million or up 17% from 2019. This amount is 64%, or there are more than half of Indonesians being accessed by cyberspace.

Events such as natural disasters trigger people, especially Twitter users, to provide their information or opinions in the form of Twitter (tweets). These tweets can be expressions of sadness, care, or complaints. The processing of data from this tweet will create a trend that can be used for information needs such as education, economics, and other areas. Natural disasters are events or events that threaten human life caused by nature, such as earthquakes, tsunamis, landslides, floods, drought. Indonesia, including countries with high levels of natural disasters, is caused by the location of the Indonesian state that is among the confluence of three world tectonic plates, namely Eurasia, Indo Australia, and the Pacific (Ramadhan,2017.). The tweet data set can be processed using the text mining concept.

Some opinions are written on social media or many Twitter will influence the Twitter user. Tweet is the user's status text used to provide information on Twitter. Based on excerpt of research results (Nurhuda et al., 2013). However, monitoring the opinions of the community is also not easy. Opinions that have been on social media are too much if they are processed manually. Therefore, a method can categorize these opinions automatically, whether it belongs to the category of positive or negative comments. (Aaputra et al., 2019).

Sentiment analysis is a process of understanding and expressing the opinions, evaluations, assessment attitudes, or views contained in a text and obtain information automatically by processing textual data (Amalia, Bijaksana, & Darmantoro, n.d.,2016), The magnitude and benefit of sentiment analysis cause research and application-based sentiment analysis to develop rapidly (Buntoro, 2017). Sentiment analysis is also referred to as opinion mining. This is a branch of data mining that analyzes, and cultivate in form opinions about objects such as products, services, organizations and specific topics (Ajeng et al., 2019). the process of determining sentiment and classifying the polarity of text in a document or sentence so that it can be categorized as positive, negative or neutral sentiment (Rokhman Fathur, 2020).

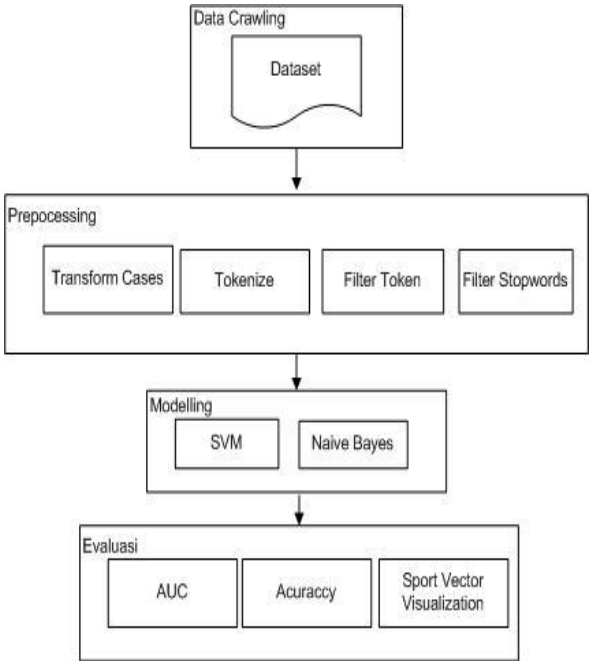
In the study that has been done about sentiment analysis. There is research on the identification of basic needs at the temporary evacuation site after the Merapi eruption. In the study, using the method used is Support Vector Machine which is combined with Lexico ¬ N-Based. These two methods have different characteristics, but still complement each other, the Lexicon method is used to create tweets that are used as training data in SVM, so there will be no manual labeling process, Data used by Research NI as 2200 data divided into 2 data, namely 90% training data comparison and 10% test data. The expected outcome of the study is to know the basic needs of a refugee site. The accuracy was achieved by doing a Cross-Validation of 10 fold from the classification model Support Vector Machine 7.96% and Maximum Entropy 87.45 (Moningka et al., 2018).

In this study, sentiment analysis was carried out to help the right needs after a disaster so that the information generated could help many parties to make decisions or choices such as with various community needs, namely health, food, clean water and electricity, by utilizing public comments about natural disasters that have been this is only used as a general discussion only, not used in research. The lesson for sentiment analysis is the Support Vector Machine and Naïve Bayes is an accurate method in terms of accuracy.

The aim of this research, sentiment analysis for distribution of disaster assistance which was not distributed due to information & coordination in the field. so that providing information on the location of natural disasters, natural disaster management, and providing funds for victims can be distributed evenly in an efficient time due to Lack of information and management of natural disasters so that the distribution of aid becomes constrained.

MATERIALS AND METHODS

In this study, data processing was performed against Twitter user tweet data taken from the Twitter API. The data then enters the pre-processing stage to avoid less-than-perfect data, data disruption, and inconsistent data so that the output of the classification has high accuracy. Model research methods can be seen in the following Figure 1:



Source: (Zumarniansyah et al., 2020)

Figure 1. Research methods

A. Data Collection Techniques

In the study, there are 2 data collection techniques namely primary data and secondary data. Primary data is Data obtained through oral inquiries using observation methods, interviews, questionnaires, etc. Secondary data is data obtained from a study that is documentation (Erni Ernawati, 2019). In this research, the data collection techniques that are taken are using secondary data that is obtained from the Twitter documentation by retrieving the RapidMiner 9.6 application into the Excel application hereinafter referred to as the Dataset.

B. Data Processing

Once the data is successfully collected, the next step is to process the data obtained from Twitter by querying "natural disasters" with the automatic Crawling method in RapidMiner.

Then at the Pre-processing stage is an early stage for processing text data into sentiment analysis, used to convert all the words into lowercase by using the Transform Cases operator, it is used to remove links/links and remove the "@" sign using the Tokenize operator, for token selection with a minimum length of 3 letters use the Filter Token operator, while to remove the words that have no meaning and meaning using the Stopwords operator.

After performing the processing stage then the weighted process is done based on the number of occurrences of the words in a document so that the document can be represented in the vector. The weight of the feature used is the Term frequency-inverse Frequency document (TF-IDF).

C. Model and Data Testing

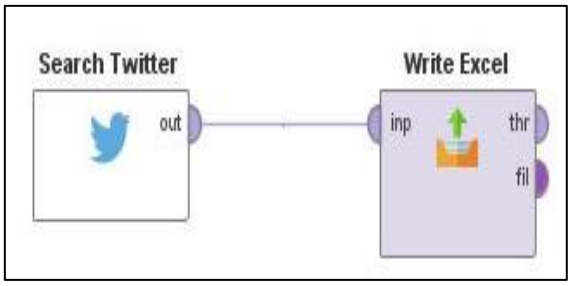
Then the data will be done the modelling process by using the SVM algorithm. Then from the data, the modelling process will be carried out using the SVM algorithm. SVM format requirements, the data collected is modified The value of the features used equals 'Input variable' in regression analysis (Son et al., 2010). I was able to work on a high-dimensional dataset using kernel tricks and Support Vector Machine entered the supervised learning class, wherein its implementation it is necessary to use the sequential training SVM and the testing phase (Rofiqoh et al., 2017). As well as comparing with the testing phase using the Naïve Bayes method is a statistical analysis algorithm, which performs the processing of numerical data using the Bayesian probability. Classification – The Bayes classification is a statistical classification that can predict the class of a probability member. The sentiment analysis of the word appears to have a weighted individual which is then calculated in total weight whether the sentence includes positive, neutral, or negative (Saputra et al., 2015). Also, Naive Bayes is one machine learning is an algorithm for classifying data (Pamungkas, Setiyanto, & Dolphina, 2015). Then the data will be evaluated in the test parameters used to evaluate is the accuracy or correctness of the classification process level which is calculated from the matrix table using AUC, accuracy, and Support Vector Visualization.

RESULTS AND DISCUSSION

This step is necessary to understand the object of research by extracting information from social media i.e. tweet data, sourced from Twitter on the sentiment of natural disaster analysis, from tweet posts can know the public comment, the tweet classification is done to see tweets that are positive and negative value. Based on this, the approach of the tweet classification model will be used for the Support Vector Machine and Naïve Bayes algorithms.

A. Dataset Retrieval

The Data used in this study is public opinion on natural disasters taken from social media Twitter, with crawling using RapidMiner with the Query "natural disasters" and taken as much as 798 data.



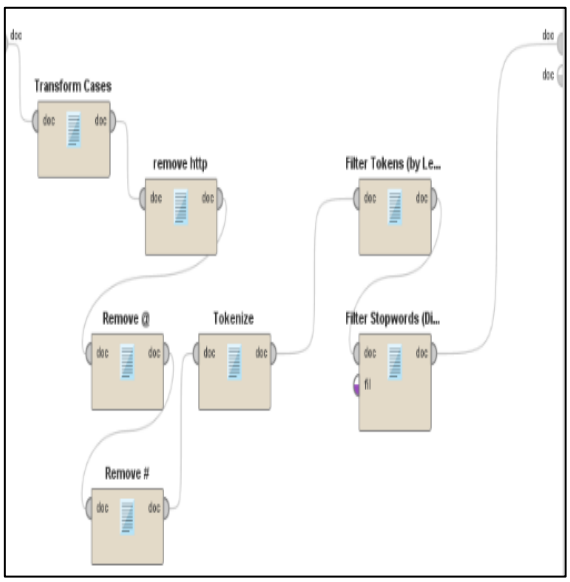
Source: (Zumarniansyah et al., 2020)

Figure 2. Crawling process for the sentiment analysis process

B. Preprocessing

This stage is an early stage for the processing of text data into sentiment analysis. After the data is successfully saved into the Excel format, then the word cleanup is done eliminating the word "RT" in the text. Pre-processing to process incoming data and be processed in a Classification to eliminate noisy data (Adiwijaya, 2006). to prepare the text to be more structured so that it is ready to be obtained by certain algorithms. The classification process is carried out to find a model or function (Prasetyowati, 2017). Pre-processing is the initial stage of text mining to convert data according to the required format. This process is carried out to process and organize information and to analyze the relationship between structured data and unstructured data (Luqyana et al., 2018).

After that, a duplicate word deletion is done to avoid word equations or loops. Here are the Preprocessing stages used:



Source: (Zumarniansyah et al., 2020)

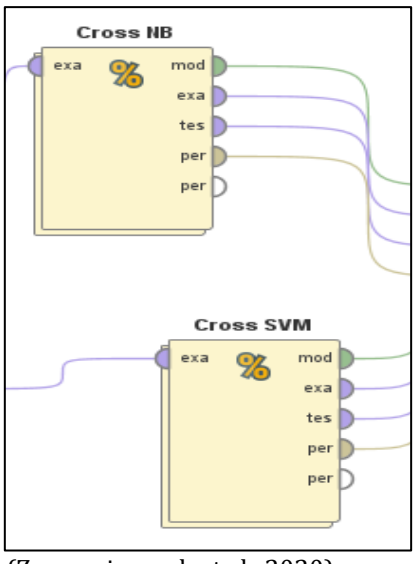
Figure 3. Process Documents from Data, Tokenizer, Transform Case, Filter Tokens, Filter Stop Words.

In the data preprocessing process in Figure 3, it can be seen that there are transform case operators, tokenize, filter tokens, and Stopwords filters :

1. *Transform Cases*
   The Operator in this stage is used to convert capital letters to lowercase letters. This is done to avoid errors in the tokenizing process.
2. *Tokenize*
   This Operator has 2 parameters to be used. In a regular expression, the mode will remove punctuation, symbols, and so on. While the latter sentence mode will be broken down into a word of words to know the weight.
3. *Filter Token*
   This Operator serves to remove any number of words after the tokenizing process with a specific character length. In this study, the minimum and maximum length of the character used is 4 characters and the maximum length is 25 characters. Meaning that the word whose character is less than 4 characters and more than 25 characters, the word will be eliminated.
4. *Filter Stopwords*
   Operator Filter Stopwords. This Operator serves to eliminate the word in Bahasa Indonesia that has no meaning and no relation to the text content such as the link "and", "," will "and others.

C. Algorithm Testing Process

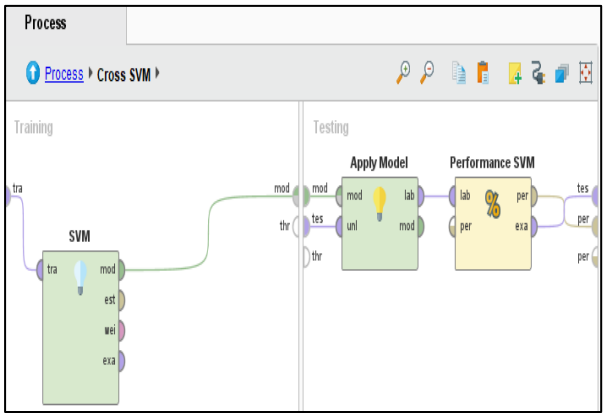
In this process using 2 algorithms, namely SVM and NB will be tested with K-fold cross-validation in Figure 4 with a value of k = 10. The following operators are used:



Source: (Zumarniansyah et al., 2020)

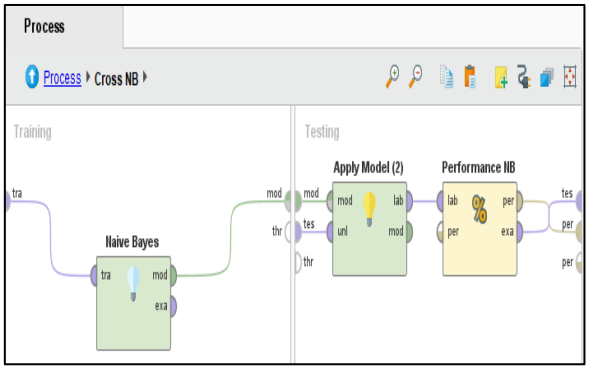
Figure 4. *Operator K-Fold Cross-Validation*

Also, there are several operators and the process can be seen in Figure 5 Process K-Fold Cross Validation Algorithms SVM :



Source: (Zumarniansyah et al., 2020)

Figure 5. Process *K-Fold Cross Validation* Algorithms SVM



Source: (Zumarniansyah et al., 2020)

Figure 6. Process K-Fold *Cross Validation* Algorithm NB

D. Accuracy And Evaluation Results

At the next stage of comparison using 2 algorithms, SVM and Naïve Bayes, each through the process of training and testing. The next step is to do in the modeling process and will get value accuracy, precision, recall in Figures 7 and 8, and the AUC graph located in Figures 9 and 10, all of these results are obtained from the SVM and Naïve Bayes algorithm processes.

This classification is done before combining the two accuracies, wherein Figure 7 about the accuracy of the Support Vector Machine algorithm which has an accuracy value of 51.18%.

accuracy: 51.17% +/- 4.05% (micro average: 51.18%)

	true Positif	true Negatif	class precision
pred. Positif	189	162	53.85%
pred. Negatif	45	28	38.36%
class recall	80.77%	14.74%	

Source: (Zumarniansyah et al., 2020)

Figure 7. The accuracy value in the SVM algorithm

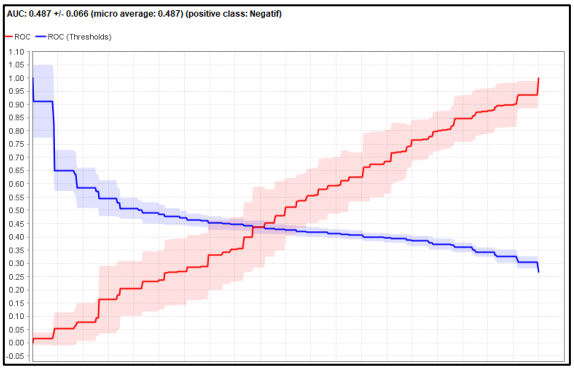
accuracy: 48.13% +/- 7.90% (micro average: 48.11%)

	true Positif	true Negatif	class precision
pred. Positif	97	83	53.89%
pred. Negatif	137	107	43.85%
class recall	41.45%	56.32%	

Source: (Zumarniansyah et al., 2020)

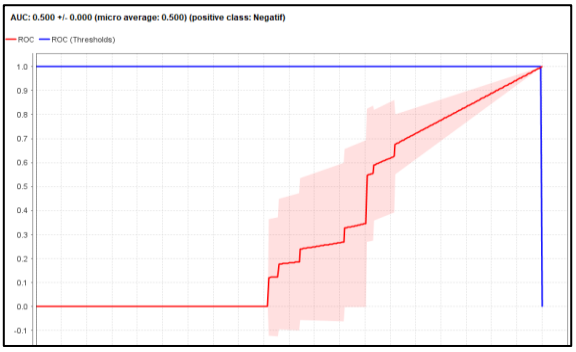
Figure 8. The accuracy value in the NB algorithm

The following are some descriptions of the test data from Figure 9 in the form of an AUC test graph on the Support Vector Machine algorithm and Figure 10 is a graph of AUC testing on the Naive Bayes Algorithm.



Source: (Zumarniansyah et al., 2020)

Figure 9. Grafik AUC AUC Graph Algorithm SVM



Source: (Zumarniansyah et al., 2020)

Figure 10. Grafik AUC AUC Graph Algorithm NB

From the comparison of the algorithm of Support Vector Machine and Naïve Bayes that get the result of the accuracy comparison of Support Vector Machine 51.18% while Naive Bayes 48.11, with the difference between the algorithm, is 3.07%.

CONCLUSION

From the ongoing data processing, the combination of the Support Vector Machine

algorithm and Naive Bayes as a good feature selection method has been shown to improve classification results. Tweet Data from Twitter can be classified as both positive and negative. The accuracy of Support Vector Machine and Naïve Bayes has a different accuracy value, where Support Vector Machine has an accuracy value of 51.18% while Naïve Bayes has an accuracy value of 48.11% in the two methods has a difference. The model that has been built can be implemented in all texts so that we can see the results directly in positive and negative forms. In subsequent research, this model can be applied to other domains and immediately add new classes that have neutral express.

REFERENCE

Aaputra, S. A., Didi Rosiyadi, Windu Gata, & Syepry Maulana Husain. (2019). Sentiment Analysis Analisis Sentimen E-Wallet Pada Google Play Menggunakan Algoritma Naive Bayes Berbasis Particle Swarm Optimization. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *3*(3), 377–382. https://doi.org/10.29207/resti.v3i3.1118

Adiwijaya, I. (2006). *Text Mining dan Knowledge Discovery*. http://web.ipb.ac.id/~ir-lab/pdf/tm (text summarization).pdf

Ajeng, K. D., Normah, N., & Ahmad, H. (2019). Prediction of Indonesia Presidential Election Results for the 2019-2024 Period Using Twitter Sentiment Analysis. *2019 5th International Conference on New Media Studies (CONMEDIA)*, 36–42. https://doi.org/10.1109/CONMEDIA46929.2019.8981823

Amalia, R., Bijaksana, M. A., & Darmantoro, D. (n.d.). *A Framework for Sentiment Analysis Implementation of Indonesian Language Tweet on Twitter A Framework for Sentiment Analysis Implementation of Indonesian Language Tweet on Twitter*. https://doi.org/10.1088/1742-6596/755/1/011001

Buntoro, G. A. (2017). Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter. *INTEGER: Journal of Information Technology*, *2*(1), 32–41. https://ejurnal.itats.ac.id/integer/article/view/95

Erni Ernawati. (2019). Ermawati, Algoritma Klasifikasi C4.5 Berbasis Particle Swarm Optimization Untuk Prediksi Penerima Bantuan Pangan Non Tunai. *Sistemasi: Jurnal Sistem Informasi*, *8*(3), 513–528. http://sistemasi.ftik.unisi.ac.id/index.php/stmsi/article/view/576

Gallardo, R. (2020). Bringing Communities into the Digital Age. *State and Local Government Review (SLGR)*, *51*(4), 233–241. https://doi.org/10.1177/0160323X20926696

Kurikulum, K. K., Pamungkas, D. S., Setiyanto, N. A., & Dolphina, E. (2015). *ANALISIS SENTIMENT PADA SOSIAL MEDIA TWITTER MENGGUNAKAN NAIVE BAYES CLASSIFIER TERHADAP*. *14*(4), 299–314.

Luqyana, W. A., Cholissodin, I., & Perdana, R. S. (2018). *Analisis Sentimen Cyberbullying pada Komentar Instagram dengan Metode Klasifikasi Support Vector Machine*. *2*(11), 4704–4713.

Nurhuda, F., Sihwi, S. W., & Doewes, A. (2013). Analisis Sentimen Masyarakat terhadap Calon Presiden Indonesia 2014 berdasarkan Opini dari Twitter Menggunakan Metode Naive Bayes Classifier. *ITSMART: Jurnal Teknologi Dan Informasi*, *2*(2), 35–42. https://jurnal.uns.ac.id/itsmart/article/view/630

Prasetyowati, E. (2017). *DATA MINING* (Moh.Afandi (ed.)). Duta Media.

Ramadhan, M. I., & Prihandoko, P. (2017). PENERAPAN DATA MINING UNTUK ANALISIS DATA BENCANA MILIK BNPB MENGGUNAKAN ALGORITMA K-MEANS DAN LINEAR REGRESSION. *JURNAL ILMIAH INFORMATIKA KOMPUTER*, *22*(1), 57–65. https://ejournal.gunadarma.ac.id/index.php/infokom/article/view/1535

Rofiqoh, U., Perdana, R. S., & Fauzi, M. A. (2017). *Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Twitter Dengan Metode Support Vector Machine dan Lexicon Based Features*. *1*(12), 1725–1732.

Rokhman Fathur, S. (2020). *LINGUISTIK DISRUPTIF : Pendekatan Kekinian Memahami Perkembangan Bahasa* (F. Azzahrah (ed.); 1st ed.). PT. Bumi Aksara.

Saputra, N., Bharata, T., & Erna, A. (2015). *Jurnal Dinamika Informatika Volume 5, Nomor 1, November 2015*. *5*(November).

Son, Y., Kim, H., Kim, E., Choi, S., & Candidate, D. (2010). *Application of Support Vector Machine for Predic- tion of Medication Adherence in Heart Failure Pa- tients*. *16*(4), 253–259. https://doi.org/10.4258/hir.2010.16.4.253

Zumarniansyah, A., Febrianto, R., Normah, N., & Gata, W. (2020). *Laporan Akhir Penelitian Mandiri: Twitter Sentiment Analysis Of Post Natural Disasters Using Comparative Classification Algorithm Support Vector Machine And Naïve Bayes*.