Indonesian Journal of Electrical Engineering and Computer Science
Vol. 22, No. 2, May 2021, pp. 976~984
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v22.i2.pp976-984

The effect of the TF-IDF algorithm in times series in forecasting word on social media

Arif Ridho Lubis, Mahyuddin K M Nasution, Opim Salim Sitompul, Elviawaty Muisa Zamzami
Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Sumatera Utara, Indonesia

Article Info

Article history:

Received Oct 8, 2020
Revised Mar 22, 2021
Accepted Mar 30, 2021

Keywords:

Forecasting
Social media
TF-IDF
Times series
Word

ABSTRACT

Forecasting is one of the main topics in data mining or machine learning in which forecasting, a group of data used, has a label class or target. Thus, many algorithms for solving forecasting problems are categorized as supervised learning with the aim of conducting training. In this case, the things that were supervised were the label or target data playing a role as a 'supervisor' who supervise the training process in achieving a certain level of accuracy or precision. Time series is a method that is generally used to forecast based on time and can forecast words in social media. In this study had conducted the word forecasting on twitter with 1734 tweets which were interpreted as weighted documents using the TF-IDF algorithm with a frequency that often comes out in tweets so the TF-IDF value is getting smaller and vice versa. After getting the word weight value of the tweets, a time series forecast was performed with the test data of 1734 tweets that the results referred to 1203 categories of Slack words and 531 verb tweets as training data resulting in good accuracy. The division of word forecasting was classified into two groups i.e. inactive users and active users. The results obtained were processed with a MAPE calculation process of 50% for inactive users and 0.1980198% for active users.

Corresponding Author:

Arif Ridho Lubis
Fakultas Ilmu Komputer dan Teknologi Informasi
Universitas Sumatera Utara
Padang Bulan 202155 USU, Medan, Indonesia
Email: arifridho.l@students.usu.ac.id

1. INTRODUCTION

TF-IDF is one that only multiplies term frequency and inverse document frequency that TF is the number of appearances of a term in a document and IDF reduces the dominant terms that often appear in various documents or files, by calculating the inverse frequency of them that contain a word and can exclude a collection of words [1]-[3]. TF-IDF is commonly used in the classification process in existing documents and was studied by several classification methods [4]-[6].

Classification, prediction and forecasting are included in techniques that utilize data to use in the hope that finding a new model in the computing of a particular interest [7]-[9]. Both classification and forecasting processes have an accurate level of techniques which the difference from each model produced and a good level of accuracy occurs if it approaches 100%, meaning that the resulting model shows the right results in the development of the model which consists of training and testing data [10].

In contrast, forecasting using TF-IDF for developed a framework for forecasting text using the KNN and TF-IDF methods and the test results showed the advantages and disadvantages of the algorithm [11], providing guidance for further development on the same framework. In addition [12], classification by using naive bayes for classification of text in machine learning that based on the values of conditional probability and had compared

algorithm methode of the multinomial, naive bayes, bernoulli, and gaussian to the SVM algorithm and the results state for representation of statistical text, TF, TF-IDF and it used character level 3 (3-Gram).

In addition to the TF-IDF text classified [13], measured the performance of the DT and SVM in classifying emotions from Malay folk tales with 100 documents which were taken from children's short stories collected and applied as data set from text-based emotion recognition experiments and the TF-IDF process was extracted from text documents and classified using DT and SVM, classifying online news by applying TF-IDF and cosine similarity, requires preprocessing, namely tokenizing, stopword and stemming to reduce terms so as to speed up the process of calculating term weighting using TF-IDF and speeding up the process of cosine similarity [5], [14]. The aim is to facilitate human error and reduce the occurrence of categorization errors. Classification is able to classify news with an accuracy rate of 91.25%. However, with the development of online media, it does not rule out the possibility of classifying online media that is often used, namely social media [15]. The use of social media in the learning process [16], especially in online discussion forums [17], is increasing [18], [19]. However, the widening of the discussion is beyond the scope of the study which should even lead to the habitual level of using social media [20]. Therefore, it is necessary to have a classification of the words that appear on social media. From several previous studies regarding TF-IDF, it was not found in forecasting posted words in social media. Thus, the need of word forecasting on social media is to get the frequency value in the posts on social media. The method used in word forecasting on social media is the time series where collaborations and improved methods are used to get to what extent TF-IDF works in forecasting words using time series.

2. MATERIAL AND METHOD

2.1. TF-IDF

TF-IDF is an algorithmic method useful to compute the weight of each commonly used word [11]. This method is also known to be efficient, easy and has accurate results [21]. The TF and IDF values for each token (word) in each document in the corpus would be computed by applying this method [22]. In simple terms, the TF-IDF method is used to find out how often a word appears in a document.

In this this, we conducted the TF-IDF algorithm method which was then be combined into NBC [23]. Consequently, the final result of this study was to create a word classification-based program on social media data obtained from twitter [24], [25] then the performance of the document was made based on the tweets that appear. The first step was to determine how often the word appears in a document. Thus, the more frequency of occurrence of the word, the greater its value will be.

Related to TF, there would be some of patterns that can be used [26]:

a) Binary TF.
b) Pure TF.
c) TF logarithmic, have high frequency.

$$TF = \begin{cases} 1 + log_{10}(f_{t,d}), & f_{t,d} > 0 \\ 0, & f_{t,d} = 0 \end{cases} \quad (1)$$

In where, the $f_{t,d}$ value, d is the frequency term (t) in document (d). Therefore, if a word or term is contained in a document 5 times, the weight = 1 + log (5) = 1,699 is obtained. However, if the term is not included in the document, the weight is zero [27].

Then, the next or the second one is the IDF which is a calculation of how widely distributed terms are in the collection of documents concerned. In contrast to TF, the more frequent words appear, the greater the value. In IDF, the less frequent words appear in the document, the greater the value. To determine the amount of the IDF value, we use the formula [28]:

$$IDF_j = \log(\frac{N}{df_j}) \quad (2)$$

Where N is the number of whole documents in the collections while df_j is that of documents containing term (t_j). The type of TF formula commonly used for calculations is pure TF. Thus, the general formula for Term Weighting TF-IDF is a combination of the raw TF calculation formula with the IDF formula by multiplying the TF value by the IDF value:

$$w_{ij} = tf_{ij}\ x\ idf_j \quad (3)$$

$$w_{ij} = tf_{ij}\ x \log(\frac{N}{df_j}) \quad (4)$$

Where W_{ij} is the weight of the term (t_j) to the documents (d_i). While tf_{ij} is the number of occurrences of term (t_j) in the documents (d_i). N is the number of all documents in the database and df_j is the number of documents that contain term (t_j) (at least one word is term (t_j)). Regardless of the value of tf_{ij}, if N = df_j, then the result would be 0 (zero), because the result is log 1, for the IDF calculation. For this reason, a value of 1 can be added on the IDF side, so that the weight calculation becomes as follows:

$$w_{ij} = tf_{ij}\, x \log\left(\frac{N}{df_j}\right) + 1 \quad (5)$$

2.2. Times series

There are various methodologies in time series such as ESM, SDM, VAM, SAM using variable time series, ARMAX models [29], [30]. The use of data in research is also carried out for analysis of forecasting patterns of rainfall distribution in various regions of the world [31]. Various time series methods for various purposes have been applied to investigate rainfall information in various literatures.

Time series is the management of data collected and observed over a certain time span [32]. There are four elements in time series data such as seasonal data, cycles, trend data, and random components. Trend patterns are usually seen from charts that go up or down over a long period of about 10 to 20. Meanwhile, seasonal data usually goes up and down in the short term, for example one year. This is what distinguishes the cycle, the cycle also shows an up and down pattern, but over a long period of time. The last component is random, that is, other variables that cannot be explained by the previous three components are random data [33].

Time series technique is historical data that is used to predict the next data. Almost simila to regression, Y is historical data and X is the period or time data itself, it can be 1 for the earliest data, and 2 for the next data and so on. The resulting model will be used to predict the next Y value. Then whether to use r-squared? The answer is yes, although time series in measuring accuracy does not use R-squared, because time series is also an equation model, R-squared should also be used to assess whether the resulting equation is good or not [34].

Trend technique is a technique commonly used in forecasting quantitative data analysis. Because basically in looking for patterns in trend data such as linear, quadratic, S curve or exponential, the model is then used to estimate the next data. The formula for forecasting with time series contained in the following formula.

$$\text{Model linear: } Y_{pred} = a + bT + e \quad (6)$$

$$\text{Model quadratic: } Y_{pred} = a + bT^2 + cT + e \quad (7)$$

$$\text{Model S curve: } Y_{pred} = L/(1 + \exp(a + b(T) + e)) \quad (8)$$

$$\text{Model exponential: } Y_{pred} = a + e^{b.T} \quad (9)$$

2.3. Mean absolute percentage error (MAPE)

Mean absolute percentage error (MAPE) is a formula to calculate the accuracy with a relatively small error size developed by applying the Mean absolute error (MAE) formula [35]. In the case of calculating the difference between the estimated and actual value, MAPE is usually used more frequently than similar formulas such as MSE and MAD because MAPE states in percentage the result of the error in predicting or forecasting the actual results during the certain period and will provide information in percentage of high or low error. In other words, MAPE is the absolute average error during a certain period which is then multiplied by 100% to get the percentage result.

Measuring the relative precision by applying MAPE is intended to determine the percentage of deviation of the estimation results [36]. This approach is useful when the size of the forecast variable is important in evaluating the accuracy of the forecast. MAPE indicates how many errors occur in estimating compared to the real values. The MAPE equation is as follow:

$$MAPE = \frac{\sum_{t=1}^{n} \left|\frac{y-\hat{y}}{\hat{y}}\right|}{n} \times 100\% \quad (10)$$

Where,
n = the amount of data in the measurement error
y = actual yield value
ŷ = the estimated result value

2.4. General architecture

The general architecture of this study can clearly be seen in Figure 1. Explanation in Figure 1 are as follows:

a) Dataset is obtained from twitter, and each tweets will be counted as a document on the IDF.
b) Perform frequency calculations based on the TF-IDF method.
c) Forecasting words whose frequency has been calculated using TF / IDF using a time series which will interpret the words whose frequency often appears can represent the documents that will appear in the future.
d) Get results from classification using timeseries in the form of comparison of forecasting results and reality on word posts on social media (MAPE).

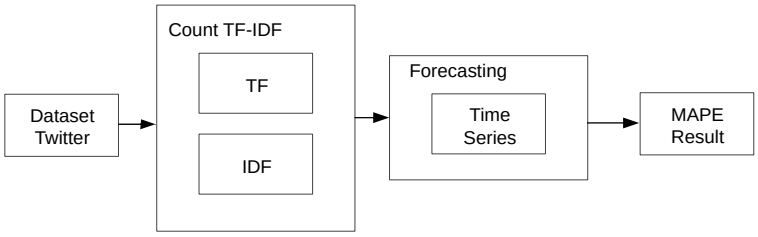

Figure 1. General architecture

3. RESULTS AND ANALYSIS

From the research methodology, a general architecture study was designed using the twitter dataset to classify a person's tweets where the tweets were converted into documents in order to get the frequency of each word. In calculating the weight using TF-IDF, the word TF value was calculated first with each word's weight is 1. Whereas the IDF value was formulated in (2). Where IDF (tweets) is the value of IDF of each word to search for, td was the number all existing documents, df the number of appearances in all documents. After getting TF and IDF values, to get the final weight of TF-IDF it was formulated in (1) where w (tweets) is the weight value of each word, TF ($word_i$) is the calculation result of TF. IDF_i is the result of the IDF calculation.

The thing that needs to be considered in seeking information from a heterogeneous collection of documents / tweets was weighting terms. Term could be a word, phrase or other indexed unit in a document which could be applied to find out the document context. Because every word possesses importance in a different level in the document, an indicator for each word is given, namely the term weight. To calculate the weight value using TF-IDF, some of the steps taken are as follows:

Table 1 shows that the TF-IDF calculation had been achieved with the frequency that often appeared, namely the word wkwk which was the slack language of LOL then followed by the word happy. However, in the TF-IDF calculation the more words appeared, the smaller the frequency that appeared. So that the frequency numbers were weighted for the use of number processing. In accordance with the purpose of this study, this figure was done by calculating time series to forecast words on social media data obtained from twitter, then the performance of the document was made based on the tweets that appear. The first step was determining how often the word appeared in a document. Thus, the more frequency of occurrence of the word, the greater its value woull be.

In conducting forecasting based on time series, there were several steps taken, namely the trainning stage and the tetsing stage which produced forecasting. At the trainning stage, the analysis process was carried out on a sample of documents in this study in the form of social media data such as tweets from twitter, i.e. the words that may appear in the sample document collection and determined the habbit of social media users who represented documents as much as possible, at the training stage there are training documents which were a reference for the testing process. Training documents, serves for class formation and as a reference for how documents will be classified, in this study the authors used data sources that had been classified into documents from twitter, the intended reference was document labeling based on the expert domain.

Testing documents, in the research carried out, the type of document used in this study was social media documents, namely twitter which contained information and was obtained with unstructured content because there were html tags and messages which made them meaningless, while for classification accuracy, structured documents were needed and should be understandable. The tweets used was the original document listed on the @a*fr***o account. The experimental document was a document totaling 1734 tweets which in

this case were categorized as an inactive twitter account and the @y*l**a***r**i account as active for analyzing the time series. The stage taken before the forecasting process was preprocessing to find meaning in training and testing documents and to support the forecasting process, this process must be done because the document test data were in the form of paragraphs and tags that eliminated the meaning of the document. This paper had difficulty in understanding the contents of the test document before the preprocessing process was carried out. Preprocessing could also affect the identification of text aimed at determining features.

The first thing in document processing was breaking down character sets into words or tokens, often referred to as tokenization. Tokenization is complex for computer programs because some characters can be found as token delimiters. Delimiters are space, tab, and line characters, while @ () <>! ? " are often used as a delimiter, but depending on the environment. Then carried out the text identification process because it is very important to recognize the text patterns that will be forecasted and to recognize the types of text that will be used for training. The problem that arised during identification was the irregularity of the text pattern that was obtained even though it had been processed using stopwords in the previous step, this caused the writer to have a little confusion in identifying the text and requiring accuracy in observation. In the identification process, the writer needed to open the documents one by one to understand the existing patterns in the text, for the patterns themselves were found irregular in the placement of content.

The process of determining the label on the training document was done manually based on the expert domain taken from www.twitter.com based on the predetermined category in the domain. Label determination was used to provide a reference in the document classification process or classify according to predetermined label categories. Based on the results of document identification that referred to the content contained in the document, the data were classified into two categories, that is inactive users and active users.

For inactive users, word forecasting was carried out on social media using time series based on frequency using TF-IDF, the word "happy" was obtained so that the results of TF-IDF were used as weight in forecasting with time series. The word "happy" was made as a label and be forecasted. The dataset for fore casting was as follows. From Table 2, forecasting was conducted into the system by applying times series method then the result could be seen in Table 3.

Table 1. Term dataset

Term (t)	D1	D2	..	..	Dn	DF	IDF
happy	0	0	..	..	0	39	=log(112/39)= 0.4578818967
makasih	0	0	..	..	0	17	=log(112/17)= 0.8188854146
Anniversary	0	0	..	..	0	4	=log(112/4)= 1.447158031
wkwkw	0	0	..	..	0	112	=log(112/112)=0
Love	0	0	..	..	0	9	=log(112/9)= 1.09482038
Selamat pagi	0	0	..	..	0	22	=log(112/22)= 0.7067177823

Table 2. User dataset is inactive

Date	Dataset	Date	Dataset	Date	Dataset
05/2010	0	12/2011	0	07/2013	0
06/2010	0	01/2012	0	08/2013	0
07/2010	5	02/2012	0	09/2013	0
08/2010	6	03/2012	0	10/2013	0
09/2010	0	04/2012	0	11/2013	0
10/2010	0	05/2012	5	12/2013	0
11/2010	0	06/2012	6	01/2014	0
12/2010	0	07/2012	0	02/2014	0
01/2011	0	08/2012	1	03/2014	4
02/2011	0	09/2012	0	04/2014	0
03/2011	0	10/2012	0	05/2014	0
04/2011	0	11/2012	0	06/2014	0
05/2011	0	12/2012	1	07/2014	0
06/2011	0	01/2013	0	08/2014	2
07/2011	0	02/2013	0	09/2014	4
08/2011	0	03/2013	0	10/2014	0
09/2011	18	04/2013	0	11/2014	2
10/2011	0	05/2013	0	12/2014	0
11/2011	0	06/2013	0	01/2015	0

Table 3. forecasting the word “happy” in inactive users

Date	Forecasting	Date	Forecasting
02/2015	0	02/2016	0
03/2015	2	03/2016	0
04/2015	2	04/2016	-1
05/2015	1	05/2016	2
06/2015	0	06/2016	1
07/2015	0	07/2016	1
08/2015	0	08/2016	0
09/2015	0	09/2016	0
10/2015	2	10/2016	-1
11/2015	1	11/2016	-1
12/2015	1	12/2016	2
01/2016	0	01/2017	1

From Table 3 it can be seen that the forecasting process could be calculated. However, it could be seen that the forecasting results were minus (-) or below zero. The forecasting process was carried out using

the happy weights achieved by using the TF-IDF of 0.4578818967. So that the forecasting results for inactive users could be seen through the graph in Figure 2.

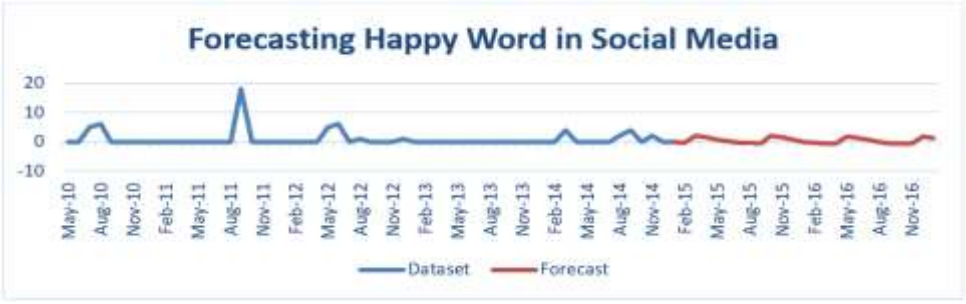

Figure 2. Forecasting happy word

In Figure 2 the blue graph is the trainning dataset and the orange graph is the forecasting result. In the test, it could be seen that in the February 2015 period the system received no forecasting value for the appearance of the word "happy", but in April 2015 the system received forecasting results 2 times while the dataset or anything actually did not exist at all. So that the MAPE results achieved by inactive social media users were 50%.

However, for active social media users, data scrolling was carried out with the activeness of each making it on twitter social media which consisted of 573 tweets which were then carried out by calculating the TF-IDF on words that often appeared. There were 4 words taken as words that often appeared. They are shown in Table 4.

From Table 4, it showed that the TF-IDF calculations could be obtained with the frequency that often appeared, i.e. the word bisa which is a statement word. However, in the TF-IDF calculation the more words appeared, the smaller the frequency that appears. So that the frequency numbers were weighted for the use of number processing. In accordance with the purpose of this study, this figure was carried out by calculating time series to forecast words on social media data obtained from twitter with a weight of 1.214843848. Then the performance of the document was made based on the tweets that appear.

After obtaining the weight value on the TF-IDF, then forecasting with the time series was done before forecasting the dataset used from October 2017 to August 2020. Where the training data started from October 2017 to March 2020 and data testing was used in April 2020 and August 2020. The dataset for active social media users was shown in Table 5. From Table 5 was a dataset of active social media twitter users and the data were those that had been obtained from the TF-IDF calculation, namely the word "bisa" and the word was carried out by training on the time series method. The results of forecasting the word "bisa" could be seen in Figure 3.

Table 4. Term dataset active social media users

Term (t)	D1	D2	..	..	Dn	DF	IDF
bisa	0	0	..	..	0	35	=log(573/35)= 1.214843848
Mau	0	0	..	..	0	25	=log(573/25)= 1.359835482
Kalau	0	0	..	..	0	23	=log(573/23)= 1.396199347
Kenapa	0	0	..	..	0	11	=log(573/11)= 1.716837723

Table 5. Dataset of active user

Date	Dataset	Date	Dataset	Date	Dataset
10/2017	0	10/2018	0	10/2019	0
11/2017	0	11/2018	3	11/2019	2
12/2017	0	12/2018	1	12/2019	3
01/2018	0	01/2019	0	01/2020	0
02/2018	0	02/2019	2	02/2020	2
03/2018	0	03/2019	0	03/2020	7
04/2018	0	04/2019	1	04/2020	0
05/2018	1	05/2019	0	05/2020	2
06/2018	2	06/2019	0	06/2020	2
07/2018	0	07/2019	0	07/2020	0
08/2018	0	08/2019	1	08/2020	0
09/2018	1	09/2019	4		

Figure 3. Forecasting “bisa” word

From Figure 3, it could be seen that the blue one is the training data and the orange one is the testing data. The word that was forecasted was the data "bisa" because it was taken from the TF-IDF calculation. From the testing data we could compare or measure the accuracy using the MAPE formula. However, before doing the MAPE calculation, you could see the forecasting results with the actual data in Table 6.

From Table 6, it could be seen that the results of forecasting with 5 testing data showed 4 data the results were the same and in August 2020 1 forecasting the word "bisa" is not there. From this data, the calculation of accuracy could be calculated with MAPE:

$$MAPE = \frac{0{,}990099}{5} \times 100\% = 0{,}1980198\ \%$$

So that a MAPE of 0.1980198% has been obtained, which was the result of very small MAPE so that word patterns or forecasting in social media could be done.

Table 6. Error and MAPE

Date	Actual	Forecasting
04/2020	0	0
05/2020	2	2
06/2020	2	2
07/2020	0	0
08/2020	0	1

4. CONCLUSION

Finally, this study drew the conclusion that twitter data could be forecasted and could find out that every content of twitter was in the form of a good activity that could reveal characteristics from users, which in the future can be used for the benefit of the industrial world. In this study, a word forecasting process in social media was carried out starting with the TF-IDF calculation where it could be concluded that the TF-IDF values with frequencies that often appeared get a smaller frequency value and vice versa words with less frequency TF-IDF values will be greater. After TF-IDF calculations were carried out, forecasting was carried out using the time series method using a method where the division of word forecasting was divided into two categories that is the category of inactive users and active users. The test data on inactive users is 1734 tweets, the results referred to 1203 categories of slack words and 531 tweets and the test data on active users is 573 tweets with 60613 words. The results obtained were done with a MAPE calculation process of 50% for inactive users and 0.1980198% for active users.

ACKNOWLEDGEMENTS

Thanks to Universitas Sumatera Utara and Politeknik Negeri Medan.

REFERENCES

[1] A. Rahmah, H. B. Santoso, and Z. A. Hasibuan, “Exploring Technology-Enhanced Learning Key Terms using TF-IDF Weighting,” *Fourth International Conference on Informatics and Computing (ICIC)*, 2019, doi: 10.1109/ICIC47613.2019.8985776.

[2] G. Li and J. Li, "Research on Sentiment Classification for Tang Poetry based on TF-IDF and FP-Growth," in *Proceedings of 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference, IAEAC 2018*, 2018, pp. 630-634, doi: 10.1109/IAEAC.2018.8577715.
[3] I. Arroyo-Fernández, C. F. Méndez-Cruz, G. Sierra, J. M. Torres-Moreno, and G. Sidorov, "Unsupervised sentence representations as word information series: Revisiting TF–IDF," *Comput. Speech Lang.*, vol. 56, pp. 107-129, 2019, doi: 10.1016/j.csl.2019.01.005.
[4] K. Chen, Z. Zhang, J. Long, and H. Zhang, "Turning from TF-IDF to TF-IGM for term weighting in text classification," *Expert Syst. Appl.*, vol. 66, pp. 1339-1351, 2016, doi: 10.1016/j.eswa.2016.09.009.
[5] T. Shouzhong and H. Minlie, "Mining microblog user interests based on TextRank with TF-IDF factor," *J. China Univ. Posts Telecommun.*, vol. 23, no. 5, pp. 40-46, 2016, doi: 10.1016/S1005-8885(16)60056-0.
[6] S. A. Mohamed, M. Othman, and M. Hafizul Afifi, "A review on data clustering using spiking neural network (SNN) models," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 15, no. 3, p. 1392, 2019, doi: 10.11591/ijeecs.v15.i3.pp1392-1400.
[7] A. Khowarizmi, Akhm, M. Lubis, and A. R. Lubis, "Classification of Tajweed Al-Qur'an on Images Applied Varying Normalized Distance Formulas," *ACM Int. Conf. Proceeding Ser.*, no. 3, pp. 21-25, 2020, doi: 10.1145/3396730.3396739.
[8] A. R. Lubis, M. Lubis, and Al-Khowarizmi, "Optimization of distance formula in k-nearest neighbor method," *Bull. Electr. Eng. Informatics*, vol. 9, no. 1, pp. 326-338, 2020, doi: 10.11591/eei.v9i1.1464.
[9] A. R. Lubis, M. K. M. Nasution, O. S. Sitompul, and E. M. Zamzami, "A Framework of Utilizing Big Data of Social Media to Find Out the Habits of Users Using Keyword," 2020, pp. 140-144.
[10] A. R. Lubis, M. Lubis, Al-Khowarizmi, and D. Listriani, "Big Data Forecasting Applied Nearest Neighbor Method," in *ICSECC 2019 - International Conference on Sustainable Engineering and Creative Computing: New Idea, New Innovation, Proceedings*, 2019, pp. 116-120, doi: 10.1109/ICSECC.2019.8907010.
[11] B. Trstenjak, S. Mikac, and D. Donko, "KNN with TF-IDF based framework for text categorization," *Procedia Eng.*, vol. 69, no. November 2013, pp. 1356-1364, 2014, doi: 10.1016/j.proeng.2014.03.129.
[12] N. Rezaeian and G. Novikova, "Persian text classification using naive bayes algorithms and support vector machine algorithm," *Indones. J. Electr. Eng. Informatics*, vol. 8, no. 1, pp. 178-188, 2020, doi: 10.11591/ijeei.v8i1.1696.
[13] M. M. Saad, N. Jamil, and R. Hamzah, "Evaluation of support vector machine and decision tree for emotion recognition of malay folklores," *Bull. Electr. Eng. Informatics*, vol. 7, no. 3, pp. 479-486, 2018, doi: 10.11591/eei.v7i3.1279.
[14] M. Fikri and R. Sarno, "A comparative study of sentiment analysis using SVM and Senti Word Net," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 13, no. 3, pp. 902-909, 2019, doi: 10.11591/ijeecs.v13.i3.pp902-909.
[15] D. B. Seo and S. Ray, "Habit and addiction in the use of social networking sites: Their nature, antecedents, and consequences," *Comput. Human Behav.*, vol. 99, no. May, pp. 109-125, 2019, doi: 10.1016/j.chb.2019.05.018.
[16] H. Pan, E. Hou, and N. Ansari, "RE-NOTE: An E-voting scheme based on ring signature and clash attack protection," in *GLOBECOM - IEEE Global Telecommunications Conference*, 2013, pp. 867-871, doi: 10.1109/GLOCOM.2013.6831182.
[17] H. Pan, E. Hou, and N. Ansari, "M-NOTE: A Multi-part ballot based E-voting system with clash attack protection," *IEEE Int. Conf. Commun.*, vol. 2015-Septe, no. February 2017, pp. 7433-7437, 2015, doi: 10.1109/ICC.2015.7249514.
[18] P. S. Dandannavar, S. R. Mangalwede, and P. M. Kulkarni, "Social Media Text - A Source for Personality Prediction," in *Proceedings of the International Conference on Computational Techniques, Electronics and Mechanical Systems, CTEMS 2018*, 2018, pp. 62-65, doi: 10.1109/CTEMS.2018.8769304.
[19] K. S. Devi, E. Gouthami, and V. V. Lakshmi, "Role of Social Media in Teaching – Learning Process," *J. Emerg. Technol. Innov. Res.*, vol. 6, no. 1, pp. 96–103, 2019, [Online]. Available: https://www.researchgate.net/publication/330497773_Role_of_Social_Media_in_Teaching-Learning_Process.
[20] Q. Liu, Z. Shao, and W. Fan, "The impact of users' sense of belonging on social media habit formation: Empirical evidence from social networking and microblogging websites in China," *Int. J. Inf. Manage.*, vol. 43, no. 13, pp. 209-223, 2018, doi: 10.1016/j.ijinfomgt.2018.08.005.
[21] M. Yuan and C. Zou, "Text Keyword Extraction Based on Meta-Learning Strategy," in *International Conference on Big Data and Artificial Intelligence, BDAI 2018*, 2018, pp. 78-81, doi: 10.1109/BDAI.2018.8546672.
[22] M. A. Rahmat, Indrabayu, and I. S. Areni, "Hoax web detection for news in bahasa using support vector machine," in *2019 International Conference on Information and Communications Technology, ICOIACT 2019*, 2019, pp. 332-336, doi: 10.1109/ICOIACT46704.2019.8938425.
[23] H. Langseth and T. D. Nielsen, "Classification using Hierarchical Naïve Bayes models," *Mach. Learn.*, vol. 63, no. 2, pp. 135-159, 2006, doi: 10.1007/s10994-006-6136-2.
[24] A. R. Lubis, F. Fachrizal, and M. Lubis, "The Effect of Social Media to Cultural Homecoming Tradition of Computer Students in Medan," *Procedia Comput. Sci.*, vol. 124, pp. 423-428, 2017, doi: 10.1016/j.procs.2017.12.173.
[25] A. R. Lubis, M. Lubis, and C. D. Azhar, "The Effect of Social Media to the Sustainability of Short Message Service (SMS) and Phone Call," *Procedia Comput. Sci.*, vol. 161, pp. 687-695, 2019, doi: 10.1016/j.procs.2019.11.172.
[26] M. Mohammed and N. Omar, "Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec," *PLoS One*, vol. 15, no. 3, pp. 1-21, 2020, doi: 10.1371/journal.pone.0230442.
[27] A. Jalilifard, V. Caridá, A. Mansano, and R. Cristo, "Semantic Sensitive TF-IDF to Determine Word Relevance in Documents," 2020, [Online]. Available: http://arxiv.org/abs/2001.09896.

[28] H. J. Kim, J. W. Baek, and K. Chung, "Optimization of associative knowledge graph using TF-IDF based ranking score," *Appl. Sci.*, vol. 10, no. 13, 2020, doi: 10.3390/app10134590.
[29] P. Esling and C. Agon, "Time-series data mining," *BodyNets Int. Conf. Body Area Networks*, vol. 45, no. 1, 2012, doi: 10.1145/0000000.0000000.
[30] N. Mishra, H. K. Soni, S. Sharma, and A. K. Upadhyay, "A comprehensive survey of data mining techniques on time series data for rainfall prediction," *J. ICT Res. Appl.*, vol. 11, no. 2, pp. 167-183, 2017, doi: 10.5614/itbj.ict.res.appl.2017.11.2.4.
[31] D. Van Dao *et al.*, "A spatially explicit deep learning neural network model for the prediction of landslide susceptibility," *Catena*, vol. 188, no. December 2019, p. 104451, 2020, doi: 10.1016/j.catena.2019.104451.
[32] Z. Zhu *et al.*, "Continuous monitoring of land disturbance based on Landsat time series," *Remote Sens. Environ.*, vol. 238, no. November 2018, p. 111116, 2020, doi: 10.1016/j.rse.2019.03.009.
[33] Y. Wang *et al.*, "Time series analysis of temporal trends in hemorrhagic fever with renal syndrome morbidity rate in China from 2005 to 2019," *Sci. Rep.*, vol. 10, no. 1, p. 9609, 2020, doi: 10.1038/s41598-020-66758-4.
[34] N. Golyandina, "Particularities and commonalities of singular spectrum analysis as a method of time series analysis and signal processing," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 12, no. 4, pp. 1-46, 2020, doi: 10.1002/wics.1487.
[35] Al-Khowarizmi, I. R. Nasution, M. Lubis, and A. R. Lubis, "The effect of a secos in crude palm oil forecasting to improve business intelligence," *Bull. Electr. Eng. Informatics*, vol. 9, no. 4, pp. 1604-1611, 2020, doi: 10.11591/eei.v9i4.2388.
[36] A. Al-Khowarizmi, O. S. Sitompul, S. Suherman, and E. B. Nababan, "Measuring the Accuracy of Simple Evolving Connectionist System with Varying Distance Formulas," *J. Phys. Conf. Ser.*, vol. 930, no. 1, 2017, doi: 10.1088/1742-6596/930/1/012004.