A new approach to achieve the users' habitual opportunities on social media

Arif Ridho Lubis, Mahyuddin K. M. Nasution, Opim Salim Sitompul, Elviawaty Muisa Zamzami
Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia

Article Info

Article history:

Received Oct 26, 2021
Revised Jul 19, 2022
Accepted Aug 17, 2022

Keywords:

Approach
Habitual
Naïve Bayes
Probability

ABSTRACT

The data generated from social media is very large, while the use of data from social media has not been fully utilized to become new knowledge. One of the things that can become new knowledge is user habits on social media. Searching for user habits on Twitter by using user tweets can be done by using modeling, the use of modeling lies when the data has been preprocessed, and the ranking will then be checked in the dictionary, this is where the role of the model is carried out to get a chance that the words that have been ranked will perform check the word in the dictionary. The benefit of the model in general is to get an understanding of the mechanism in the problem so that it can predict events that will arise from a phenomenon which in this case is user habits. So that with the availability of this model, it can be a model in getting opportunities for user habits on Twitter social media.

Corresponding Author:

Arif Ridho Lubis
Faculty of Computer Science and Information Technology, Universitas Sumatera Utara
Padang Bulan 202155 USU, Medan, Indonesia
Email: arifridho.l@students.usu.ac.id

1. INTRODUCTION

The use of data on social media is very diverse and growing rapidly [1], [2]. Currently, social media produces huge amounts of data every day [3], [4]. This very large data can be used for various purposes and depends on what purpose the data is used for [5]. Consciously or unconsciously, Twitter social media users every time they tweet produce a maximum of 140 characters of letters in it, where each character forms a word and then forms a sentence. Sentences formed from words have their own meaning [6], to get the habits of users on Twitter social media, it can be seen from the frequent repeated words used by users. Where as in real life, activities or words that are often done are habits, habits are closely related to words that are verbs. Searching by utilizing available data in the world of social media is growing rapidly, but each method or method used by developers and researchers is different depending on the purpose and each method they do has advantages and disadvantages [7], [8].

The word habit on social media means a lot [9], where the habits of users on social media become new knowledge that can be used for other purposes, such as for the industrial world and other communities. One of the word searches uses the string match [10], [11], by utilizing the match string to get the number of words that are repeated from the initial data where the ranking applies according to the repetition of a word and is matched on a dictionary that has been labelled for look for the top-ranking word according to the labelled word in the dictionary. To facilitate the search, it is necessary to develop a new approach using mathematics where mathematics is the basic science of the computer itself. By utilizing modelling that serves to get an understanding or clarity of the mechanism in the problem so that it can predict events that will arise from a phenomenon.

2. MATERIAL AND METHOD

2.1. Document

Every document on social media can be used for research purposes. A social media document has many interrelated contexts, including user-provided annotations containing information [12]. The annotation itself contains tags [13], time, place, title and others that are closely related to user posts. From the structured context into a social media document, it is used as a research source, especially text data to become a very valuable source of information [14]. Text is also the simplest type of representation of information. Text documents include text document classification, grouping, topic detection, and several other processes [15].

Usually, the document is symbolized by D, so if there are several documents it becomes D_1, D_2, D_3… D_n. The document itself consists of a series of sentences that are interconnected with words, the word is symbolized by w. The number of words that may be obtained from a sentence so that it is symbolized by w_1, w_2, w_3…. w_n.

2.2. String matching

String matching technique is a pattern search in natural language processing, text, image processing, pattern and speech recognition that are commonly used [16]. There will be terms that are often encountered in string matching, namely patterns and text. The string matching algorithm is used to match a text with other text [17], [18]. A simple example of string matching is: i) Pattern: Watch and ii) Text: I watch animated movies on TV.

2.3. Basic probability

It is a statistical experiment which produces only one of many possible outcomes [19], [20]. The set of the whole possible oucames, the sample space is symbolized withl **Ω**. This sample space is also known as the set of events. Usually, a result can be denoted by ω. For an event probability is defined by P(.). For example, an experiment about choosing a word from a text "The definition of statistical experimentation can be widely stated as a process". So here the sample space is:

$$\Omega = \{a, statistical, experiment, can, be, broadly, defined, as, a, process\} \quad (1)$$

The incident occurred when choosing a word with:

$$\omega_1 = a, \omega_2 = statistical, w_3 = experiment, \dots, \omega_{10} = process \quad (2)$$

The total word count of the text is 10, or it is also known as the cardinality of **Ω**. Then it can be defined that the probability of choosing a word, for example, "experiment", is written $P(experiment) = \frac{1}{10}$ or it can be written as:

$$P = \frac{The\ number\ of\ occurrences\ of\ choosing\ the\ word\ experiment}{The\ total\ number\ of\ events\ that\ can\ occur} \quad (3)$$

While the probability of choosing the word "a", P(a) is $\frac{2}{10}$, because the number of words "a" in the text is 2. The complement of **Ω** is the incident of **Ω** not occured and symbolized by Ω^c with the note that $P(\Omega) = 1 - P(\Omega^c)$. Moreover *P* (*D*|*R*) defines the conditional distribution of *D* where *R* is known. Furthermore, ∅ represents the empty set, that is ∅= {}. Suppose *D* and *R* are two occurrences of the sample space **Ω**, finite with *N* elements, this can be expressed as a Venn diagram as shown in Figure 1. The combination of events between *D* and *R*, can be described by *D* ∪ *R*, where either event *D* or *R* or both occur [21].

In case of *D* ∩ *R*, the intersection of occurrence *D* and *R* [22]. Here, they are considered mutually exclusive when the matter of one event prevents another's event as shown in Figure 2. If the object is in ellipse R, what is the probability that the object is also in D. In order to be in D, the object must also be in slice D ∩ R. Therefore, the probability is equal to the number of components in |D ∩ R|, split by those in R, i.e., |R|. Officially the pattern is as (4):

$$P(D|R) = \frac{|D \cap R|}{|R|} = \frac{|D \cap R|/|\Omega|}{|R|/|\Omega|} = \frac{P(D \cap R)}{P(R)} \quad (4)$$

2.4. Architecture research

To be directed and precise according to the rules, it is limited by general architectural research as shown in Figure 3. So, the discussion will be more focused on what the author wants to get. From the figure, the following steps are obtained:

i) Initial data from Twitter that is ready for use, where the data has been pre-processed previously [19], where pre-processing follows from the purpose of this study.
ii) The preprocessing data is then ranked by word [23], where the assumption is that the word with repeated frequency is a word that refers to the habits of Twitter social media users.
iii) The results of this ranking will be checked by the campus that has been prepared, normally checking is done using string matching.
iv) Here the research environment plays a role between word ranking activities and checking the dictionary, where a new model is made to get opportunities for user habits on Twitter social media.

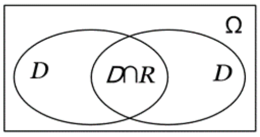

Figure 1. Sum of 2 events

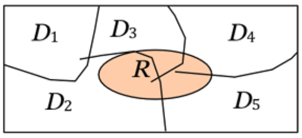

Figure 2. Mutually exclusive

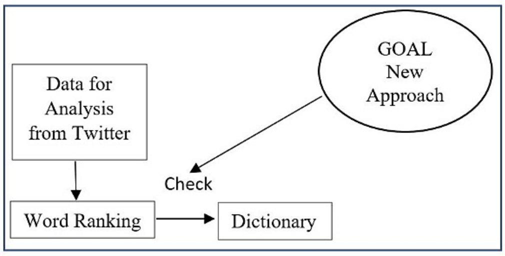

Figure 3. Architecture research

3. RESULTS AND DISCUSSION

3.1. Total probability theorem

The total probability theorem should be first comprehended before starting to discuss about Bayes' theorem, theorem [24]. Starting with the regulation to add two events *A* and *B*, it is clear form Figure 1 and Figure 2 that the (5). Suppose that the sample space is subdivided in n independent occurrence D_i, i=1.n, as seen in Figure 3. In Figure 3, it can be conluded if *DR* is expressed by $R = (R \cap D_1) \cup (R \cap D_2) \cup (R \cap D_3) \cup (R \cap D_4) \cup (R \cap D_5) \cup \cup (R \cap D_n)$ so that the total probability theorem can be obtained as:

$$P(D \cup R) = P(D) + P(R) - P(D \cap R) \tag{5}$$

$$=\sum_{i=1}^{n} P(R|D_i)P(D_i) \tag{6}$$

becomes

$$\mathrm{P(R) = P(R|D)\ P(D) + P(R|Dc)P(Dc)} \tag{7}$$

as D2 ∪ D3 ∪ ... ∪ Dn is the complement of D1. Bayes' theorem [25] can be described:
Suppose $|D| \neq 0$ dan $|R| \neq 0$ The following conditions can be stated:

$$P(D|R) = \frac{|D \cap R|}{|R|} = \frac{P(D \cap R)}{P(R)} \tag{8}$$

$$P(R|D) = \frac{|R \cap D|}{|D|} = \frac{P(R \cap D)}{P(D)} \tag{9}$$

where the press (8) and (9) is clear as:

$$P(D \cap R) = P(D|R)P(R) = P(R|D)P(D) \tag{10}$$

as the result:

$$P(D|R) = \frac{P(R|D)P(D)}{P(R)} \tag{11}$$

When the sample Ω space can be split into a finite number of independent occurrences D_1, D_2, D_3, D_4, D_5 ...D_n, and when R constitute an occurrence with $P(R)>0$, that is the combined subset of all D_i, then for each D_i, then for each Di, Bayes formula which can be generalized as:

$$P(D_i|R) = \frac{P(R|D_i)P(D_i)}{\sum_{j=1}^{n} P(R|D_j)P(D_j)} \tag{12}$$

In (12) results from (11) due to the total probability theorem (6) and (7). With the recognized observational data, Bayes' theorem may be used to calculate the posterior probability of a hypothesis.

3.2. Naïve Bayes classification

Naive Bayesian learning adverts to the formation of a Bayesian probability model, which applies the posterior class of probability to a case: $P(Y=y_j|X = \mathbf{x}_i)$. Naive Bayes classifier [26] applies this probability to give a case to a class. Implement Bayes' Theorem (11) and simplify the notation, it would be obtained:

$$P(y_j|x_i) = \frac{P(x_i|y_j)P(y_j)}{P(x_i)} \tag{13}$$

Where the numerator in the (13) is the combined probability of $\mathbf{x}_i$ and y_j (10). As a result, the denominator can be changed into: only use $\mathbf{x}$, omit the index i for simplify:

$$P(\mathbf{x}|yj)P\ (yj) = P(\mathbf{x}, yj) = P(x1|x2, x3, \dots, xp, yj)\ P(x2|x3, x4, \dots, xp, yj)\ P(xp|yj)\ P(yj)$$

Suppose the individual xi is not dependent one another. This is a strong hypothesis that clearly is against practical application, and thus Naive-as the suggested name. This supposition leads to $P(x_1|x_2, x_3, ..., x_p, y_j) = P\ (x_1|y_j)$. So, the combined probability of $\mathbf{x}$ and y_j is:

$$P(x|y_j) = \prod_{k=1}^{p} P(x_k|y_j)P(y_j) \tag{14}$$

which can be entered into press (13), so that there are:

$$P(y_j|x) = \frac{\prod_{k=1}^{p} P(x_k|y_j)P(y_j)}{P(x)} \tag{15}$$

It should be noted that the denominator, $P(\mathbf{x})$, has nothing to do with the category, for example for the categories y_j and y_l are the same. $P(\mathbf{x})$ works to be a scale factor and convinces that the posterior probability $P(y_j|\mathbf{x})$ is appropriately scaled. If we focus in clear classification rules, that is, to accurately assign each case to a class, we only need to calculate the numerator of each class and choose the maximum value of this value. The regulation is called the posterior maximum rule (16). The result class is also called the posterior maximum class (MAP), for the case of $\mathbf{x}$ it is calculated as $y\hat{}$:

$$y = argmax_{y_j} \widehat{\prod_{k=1}^{p}} P(x_k|y_j)P(y_j) \tag{16}$$

The maximum likelihood probability of a word belonging to a certain category is set by (17):

$$P(x_i|c) = \frac{Number\ of\ words\ x_i\ in\ class\ c\ document}{Total\ number\ of\ words\ in\ class\ c\ document} \tag{17}$$

According to Bayes' rule, the probability which a particular document corresponds to class c_i is given by:

$$P(c_i|d) = \frac{P(d|c_i)P(c_i)}{P(d)} \tag{18}$$

If we use the assumption of conditional free simple form, that we know a class, the words are conditionally independent of each other. Because of this assumption, the model is called Naïve.

$$P(c_i|d) = \frac{\prod P(x_i|c_j)P(c_j)}{P(d)} \tag{19}$$

Here xi is a word from the document. Classifier returns the class with the maximum posterior probability.

3.3. User

To determine what is the probability that the selected word is a verb. So, to determine the probability of the verb obtained in the dictionary, it is necessary to first describe the following things. For users, do the following steps to get a new model, as: i) number of users: U; ii) with U: $\{b_1, b_2, \dots, b_u\}$; iii) average number of texts user: T; iv) with $T = t^{b_1},, t^{b_2}, \dots, t^{b_u}$; v) the meaning is $t^{b_1}$is the text from user b_1 , $t^{b_2}$is the text from user b_2; vi) average number of words per user: W; vii) with $W = w^{b_1}, w^{b_2}, \dots, w^{b_u}$; viii) average number of verbs from the dictionary: V; ix) so, for the whole community; x) the number of texts from all users is T.U; xi) the number of words for all users is W.U; xii) the order of occurrence of the word R; xiii) choose a word from the set W taken from the set of texts T from community U; xiv) the problem is to determine what is the probability that this selected word is a verb that belongs to the R occurrence ranking; and xv) the probability of choosing a word from the set W from the entire text is:

$$P(w^{b_i}|T) = \frac{WU}{TU} \tag{20}$$

Take the class t^{b_j} where j=1,2,…, u is the text class that comes from user $j \in U$. The probability that the class t^{b_j} given that the word w^{b_i} \ is in that class can be revised in conditional probability as:

$$P(t^{b_j}|w^{b_i}) = \frac{P(w^{b_i}|t^{b_j})P(t^{b_j})}{P(w^{b_i})} \tag{21}$$

In this equation $P(w^{b_i}|t^{b_j}) = \frac{WU}{TU}$ (from the Press. 20). The probability of choosing class t^{b_j} is $\frac{1}{T}$. While the probability of choosing the word w^{b_i} is $\frac{1}{W}$.

$$P(t^{b_j}|w^{b_i}) = \frac{({}^{WU}/_{TU})\,{}^{1}/_{T}}{{}^{1}/_{W}}$$

$$\text{or} = \frac{({}^{WU}/_{TU})\,W}{T}$$

WU is the total number of words from all users, TU total number of texts from all users, W total number (words)/user, T total number (texts)/user. Now we want to determine that the selected word is a verb. The number of verbs in the dictionary is V. The probability of choosing a word w^{b_i} knowing that it is a verb, is written as:

$$P(w^{b_i}|V) = \frac{WU}{V} \tag{22}$$

and the probability that the word w^{b_i} is contained in the V verb dictionary.

$$P(V|w^{b_i}) = \frac{P(w^{b_i}|V)P(V)}{P(w^{b_i})}$$

$$= \frac{({}^{WU}/_{V})\,{}^{1}/_{V}}{{}^{1}/_{W}}$$

$$= \frac{(WU)W}{V^2}$$

So now what we want to determine is the probability that the word w^{b_i} \ is included in the text of T and is included in the verb dictionary V. In other words, the word w^{b_i} is a verb.

$$P(w^{b_i}|V,T) = P(V|w^{b_i})P(T|w^{b_i})$$

$$P(V|w^{b_i})$$

Is the probability of selecting a verb from the user b_i. That is $\frac{w^{b_i U}}{v^2}$ W, $P(T|w^{b_i})$. It is the probability that the selected word from user b_i comes from the text. That is $\frac{w^{b_i}}{T}\frac{w^{b_i}}{TU}$. With this new model, we can get a new approach to find opportunities for user habits on Twitter social media, making it easier for us to analyze user habits.

4. CONCLUSION

User habits on Twitter social media can be used for new knowledge. The same as getting a model by getting a mathematical model to look for word opportunities in the text that are included in the verb dictionary to be able to get users' habits on Twitter social media. Where is very helpful in terms of finding user habits. The design of the new model $\frac{w^{b_i}}{T}\frac{w^{b_i}}{TU}$ can be used to find user habits. In the future, with this model, users can get habits, so that it is useful for other things such as in the field of product promotion, communities that have the same habits and others.

REFERENCES

[1] E. Amereihn *et al.*, "The social habit 2019." Edison Research, pp. 1–42, 2019.
[2] A. RA, R. S. Hegadi, and M. TN, "A big data security using data masking methods," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 7, no. 2, pp. 449–456, 2017, doi: 10.11591/ijeecs.
[3] A. R. Lubis, M. K. M. Nasution, O. S. Sitompul, and E. M. Zamzami, "The effect of the TF-IDF algorithm in times series in forecasting word on social media," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 2, p. 976, 2021, doi: 10.11591/ijeecs.v22.i2.pp976-984.
[4] E. B. Setiawan, D. H. Widyantoro, and K. Surendro, "Measuring information credibility in social media using combination of user profile and message content dimensions," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 4, pp. 3537–3549, 2020, doi: 10.11591/ijece.v10i4.pp3537-3549.
[5] A. R. Lubis, M. K. M. Nasution, O. S. Sitompul, and E. M. Zamzami, "Obtaining value from the constraints in finding user habitual words," in *2020 International Conference on Advancement in Data Science, E-learning and Information Systems (ICADEIS)*, Oct. 2020, pp. 1–4, doi: 10.1109/ICADEIS49811.2020.9277443.
[6] M. AbdullahAl-Hagery, M. AbdullahAl-Assaf, and F. MohammadAl-Kharboush, "Exploration of the best performance method of emotions classification for arabic tweets," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 2, pp. 1010–1020, 2020, doi: 10.11591/ijeecs.v19.i2.pp1010-1020.
[7] G. U. Vasanthakumar, D. R. Shashikumar, and L. Suresh, "Profiling social media users, a content-based data mining technique for Twitter users," in *2019 1st International Conference on Advances in Information Technology (ICAIT)*, Jul. 2019, pp. 33–38, doi: 10.1109/ICAIT47043.2019.8987304.
[8] A. Mittal and S. Patidar, "Sentiment analysis on twitter data: a survey," in *ACM International Conference Proceeding Series*, 2019, pp. 91–95, doi: 10.1145/3348445.3348466.
[9] B. Gardner, A. L. Rebar, B. Gardner, and A. L. Rebar, "Habit formation and behavior change," *Psychology*, no. December, 2019, doi: 10.1093/obo/9780199828340-0232.
[10] D. E. Knuth, J. H. Morris, Jr., and V. R. Pratt, "Fast pattern matching in strings," *SIAM Journal on Computing*, vol. 6, no. 2, pp. 323–350, 1977, doi: 10.1137/0206024.
[11] W. Xuan, C. Ziyang, and W. Ding, "Uncertain string matching based on bitmap indexing," in *ACM International Conference Proceeding Series*, 2020, pp. 384–389, doi: 10.1145/3383972.3384007.
[12] H. Becker, M. Naaman, and L. Gravano, "Learning similarity metrics for event identification in social media," in *Proceedings of the third ACM international conference on Web search and data mining - WSDM '10*, 2010, p. 291, doi: 10.1145/1718487.1718524.
[13] H. Lian, Z. Qin, T. He, and B. Luo, "Knowledge graph construction based on judicial data with social media," in *2017 14th Web Information Systems and Applications Conference (WISA)*, Nov. 2017, pp. 225–227, doi: 10.1109/WISA.2017.46.
[14] Z. Wang, J. Liu, G. Sun, J. Zhao, Z. Ding, and X. Guan, "An ensemble classification algorithm for text data stream based on feature selection and topic model," in *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, Jun. 2020, pp. 1377–1380, doi: 10.1109/ICAICA50127.2020.9181903.
[15] W. H.Gomaa and A. A. Fahmy, "A survey of text similarity approaches," *International Journal of Computer Applications*, vol. 68, no. 13, pp. 13–18, 2013, doi: 10.5120/11638-7118.
[16] S. I. Hakak, A. Kamsin, P. Shivakumara, G. A. Gilkar, W. Z. Khan, and M. Imran, "Exact string matching algorithms: survey, issues, and future research directions," *IEEE Access*, vol. 7, pp. 69614–69637, 2019, doi: 10.1109/ACCESS.2019.2914071.
[17] I. Markic, M. Stula, and M. Zoric, "String pattern searching algorithm based on characters indices," 2019, doi: 10.23919/SpliTech.2019.8783109.
[18] M. AbuSafiya, "String matching algorithm based on letters' frequencies of occurrence," in *2018 8th International Conference on Computer Science and Information Technology (CSIT)*, Jul. 2018, pp. 186–188, doi: 10.1109/CSIT.2018.8486384.

[19] G. Sarna and M. P. S. Bhatia, "A probalistic approach to automatically extract new words from social media," in *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016*, 2016, pp. 719–725, doi: 10.1109/ASONAM.2016.7752316.

[20] L. V. Williams, *Probability, choice, and reason*. CRC Press, 2021.

[21] D. Berrar, "Bayes' theorem and Naive Bayes classifier," *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, vol. 1–3, pp. 403–412, 2018, doi: 10.1016/B978-0-12-809633-8.20473-1.

[22] V. B. Nevzorov, M. Ahsanullah, and S. Annanjevskiy, *Probability theorry*. 2014.

[23] N. Ragavan, "Efficient key hash indexing scheme with page rank for category based search engine big data," *Proceedings of the 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing, INCOS 2017*, vol. 2018-Febru, pp. 1–6, 2017, doi: 10.1109/ITCOSP.2017.8303118.

[24] A. S. Wagaman and R. P. Dobrow, *Probability: with applications and R*. Wiley, 2021.

[25] G. Michael and W. David, *Bayes linear statistics theory and methods*. Wiley, 2007.

[26] A. S. Chan, F. Fachrizal, and A. R. Lubis, "Outcome prediction using Naïve Bayes algorithm in the selection of role hero mobile legend," *Journal of Physics: Conference Series*, vol. 1566, no. 1, 2020, doi: 10.1088/1742-6596/1566/1/012041.

BIOGRAPHIES OF AUTHORS

Arif Ridho Lubis got master from Universiti Utara Malaysia in 2012 and graduate from Universiti Utara Malaysia in 2011, both information technology. He is a lecturer in Department of Computer Engineering and Informatics, Medan State Polytechnic in 2015. His research interest includes computer science, network, science and project management. He can be contacted at email: arifridho.l@students.usu.ac.id.

Mahyuddin K. M. Nasution is Professor from Sumatera Utara University, Medan Indonesia. Mahyuddin K. M. Nasution was born in the village of Teluk Pulai Dalam, Labuhan Batu Regency, North Sumatera Province. Worked as a Lecturer at the Sumatera Utara University, fields: mathematics, computer and information technology. Education: Drs. Mathematics (USU Medan, 1992); MIT, Computers and Information Technology (UKM Malaysia, 2003); Ph.D. in Information Science (Malaysian UKM). He can be contacted at email: mahyuddin@usu.ac.id.

Opim Salim Sitompul received the Ph.D. degree in information science from Universiti Kebangsaan Malaysia, Selangor, in 2005. He is currently a Professor with the Department of Information Technology, Sumatera Utara University, Medan, Indonesia. His skill and expertise are in AI, data warehousing, and data science. His recent projects are in natural language generation (NLP) and AIoT. The result of his work can be seen in his most recent publication is ''Template-Based Natural Language Generation in Interpreting Laboratory Blood Test.' He can be contacted at email: opim@usu.ac.id.

Elviawaty Muisa Zamzami graduated from Bandung Institute of Technology (Indonesia), magister of informatics, 2000 and awarded Doctoral in Computer Science from the University of Indonesia in 2013. She is a lecturer at Department of Computer Science, Sumatera Utara University, Indonesia. Currently her research interests are reverse engineering, requirements recovery, software engineering, requirements engineering and ontology. She can be contacted at email: elvi_zamzami@usu.ac.id.