TELKOMNIKA Telecommunication Computing Electronics and Control
Vol. 21, No. 4, August 2023, pp. 815~824
ISSN: 1693-6930, DOI: 10.12928/TELKOMNIKA.v21i4.24180

Experimental of vectorizer and classifier for scrapped social media data

Setiawan Assegaff[1], Errissya Rasywir[2], Yovi Pratama[2]
[1]Information System Department, Computer Science Faculty, Universitas Dinamika Bangsa, Jambi, Indonesia
[2]Informatics Engineering Department, Computer Science Faculty, Universitas Dinamika Bangsa, Jambi, Indonesia

Article Info

Article history:

Received Jun 16, 2022
Revised Dec 06, 2022
Accepted Feb 16, 2023

Keywords:

Classifier
Experiment
Social media
Text processing
Vectorizer

ABSTRACT

In this study, we used several classifiers and vectorizers to see their effect on processing social media data. In this study, the classifiers used were random forest, logistic regression, Bernoulli Naive Bayes (NB), and support vector clustering (SVC). Random forests are used to reduce spatial complexity, and also to minimize errors. Logistic regression is a method with a statistical model whose basic form uses a logistic function to represent the binary dependent variable. Then, the Naive Bayes function uses binary elements and SVC which has so far given good results rivals other guided learning. Our tests use social media data. Based on the tests that have been carried out on classifier variations and vectorizer variations, it was found that the best classifier is a linear regression algorithm based on predictive adaptive compared to the random forest method based on decision trees, probability-based Bernoulli NB and SVC which work by clustering. Meanwhile, from the test results on the count vectorizer, term frequency-inverse document frequency (TFIDF), and hashing, the best accuracy is achieved on the TFIDF vectorizer. In this case, it means that the TFIDF vectorizer has a better value in presenting word feature dimensions.

Corresponding Author:

Errissya Rasywir
Informatics Engineering Department, Computer Science Faculty, Universitas Dinamika Bangsa Jambi
Jln. Jendral Sudirman, Thehok, South Jambi, Jambi City, Indonesia
Email: errissya.rasywir@gmail.com

1. INTRODUCTION

Affordable digital devices and internet access allow people to obtain information quickly and easily. The dissemination of information at this time cannot be separated from the increasingly existing use of social media. The Ministry of Communication and Information stated that Indonesian people use the internet more to access social media [1]. Social media is an online application that allows its users to interact, participate, collaborate, and share information [2], [3]. The process of collecting data from the social media is known as crawling or text mining in natural language processing. Social media currently has an important role in various aspects of human life [4], [5]. Social media becomes a means to share information for its users [6].

Data stored on social media is very useful if it can be processed into information. Before being processed into information, the data contained in social media must be collected first, for further analysis and extraction of information is carried out [7]. To collect information, a data collection technique called the crawling technique can be used [8]. In this research, crawling techniques will be implemented to collect data originating from the social media Twitter [9]. Where the results of this study can be used for analysis and information extraction.

Data from the social media will be tested using several types of classifiers and vectorizers. There are many types of classifiers commonly used to classify sentiment analysis. In doing classifications, classifier algorithms used both machine learning and non-machines learning methods. Some of the algorithms and classification methods most often used are machines learning-based methods such as support vector machines (SVM), neural network (NN), Naive Bayes, decision tree, k-nearest neighborhood, and many other methods [10]–[13]. In this research, the type of classifier used is the random forest classifier. The random forest classification is done to reduce the complexity of the space, also, the random feature selection method to minimize errors [14]. Then, we also test the logistic regression algorithm which is a method with a statistical model whose basic form uses logistic functions to represent binary dependent variables [15]. In this study, we also use the Bernoulli Naive Bayes (NB) which is a function of the Naive Bayes classifier that uses binary elements and support vector clustering (SVC) which so far has provided good results rivaling other guided learning [1].

In several types of classifiers tested, we tested the strength of each classifier method with several vectorizers. The type of vectorizer used to be tested with the four classifiers above is the count vectorizer, term frequency-inverse document frequency (TFIDF) vectorizer, and hashing vectorizer. The experimental process was carried out on data from the social media relating to community sentiment towards the service and performance of the government of a city on the Indonesian island of Sumatra.

2. EXPERIMENTAL ON SCRAPPED SOCIAL MEDIA DATA

In this study, we carried out 3 main processes that will be performed on each type of classifier with a vectorizer combination. The process includes the process of crawling data, training and testing data. The technology to perform the crawling technique will be developed using the Python programming language. This is done to provide an alternative for users to use the crawling function. This research focuses on classifier and vectorizer experiments on data crawling from Twitter social media.

For this Twitter crawling technique, Twitter has given users access to take advantage of the Twitter application programming interface (API). So, by utilizing the Twitter API, users can easily obtain data such as tweets, user data, and others. For further collected and stored in a file or database. Following is the first step flowchart from the data collection process.

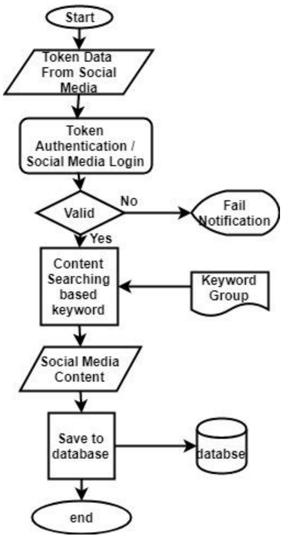

Start
Token Data From Social Media
Token Authentication / Social Media Login
Valid
No
Fail Notification
Yes
Content Searching based keyword
Keyword Group
Social Media Content
Save to database
databse
end

Figure 1. Data collection process flowchart

In Figure 1, the data retrieval is performed using the Twitter API that has been provided by Twitter to facilitate users to be able to interact with data on Twitter. These data, for example, are tweets, user IDs, location, time of tweet creation and others. To utilize Twitter API, users must use server-side scripting languages such as PHP, Python, R, and others.

By using these languages, users can make requests to the Twitter API, and the response results are forgotten in the JSON format. To secure user communication with the Twitter API, Twitter implements OAuth or Open Authorization. OAuth is an open protocol that allows users to share personal resources such as photos, videos, user data, and others stored on a website, with other sites without providing the user's name and password. OAuth allows users to provide access to third-party sites to access their information stored at other service providers without having to share access permissions or their entire data.

2.1. Social media data

Crawling is a technique used to gather information on the web. Crawling works automatically, where information is collected based on keywords provided by the user. The tool used for crawling is called a crawler. Crawlers are programs that are programmed with certain algorithms, so that they can scan web pages, according to the web address or keywords provided by the user [15]. When scanning, the crawler will read the text, hyperlinks, and various tags used on the web page. Based on this information, crawlers will index the information or store the information in a file or a database. Of these, divided into several categories of users according to the type of social media users. Social media is divided into five categories [15], namely:

a) Social networks, for example like Facebook, LinkedIn, and others
b) Microblogging, like Twitter, Tumblr, and others
c) Photo sharing, such as Instagram, Flickr, and others
d) Video sharing, such as YouTube, Vimeo, and others
e) Instant messaging, like WhatsApp, Line, and others

In this study, we use social media twitter. Because Twitter is a social media whose data is more dominant in the form of text and is the most widely used. Twitter is a social networking service, which allows users to communicate with each other. In the message sent, when mentioning the name of another user then the tweet is written with an @ followed by the user's name. Users can use the # sign (hashtag) to write messages based on topics [16], [17].

2.2. Classifier

2.2.1. Random forest

The random forest classification is a tree-based algorithm that approaches stochastic discrimination in classification. Development of trees in the random forest until it reaches the maximum size of the data tree [18]–[20]. However, the construction of random forest trees is not carried out pruning which is a method to reduce the complexity of the space. Development is carried out by applying the random feature selection method to minimize errors [21].

Formation of the tree with sample data using variables that are taken at random and run classification on all trees that are formed. Random forest uses a decision tree to do the selection process. The tree that is built is recursively divided from data in the same class. Split tools are used to split data based on the type of attribute used. Making bad trees will make conflicting random predictions. Thus, some decision trees will produce good answers [22]. The advantage of using the random forest is being able to classify data that has incomplete attributes, can be used for classification and regression but not too good for regression, more suitable for classifying data and can be used to handle large sample data [23].

2.2.2. Logistic regression

The logistic regression method is a method based on predictive analysis. This logistic regression method is commonly used to describe data and explain the relationship of dependent binary variables or nominal, ordinal, interval or ratio-level independent variables [21]. The logistic regression method is a method with a statistical model whose basic form uses logistic functions to represent binary dependent variables, even though there are more complex extensions. In regression analysis, logistic regression (or logic regression) estimates the parameters of the logistic model (a form of binary regression) [19].

There are three main types of logistic regression that differ in execution and theory. Among other things, namely binary logistic regression which is suitable for classifying an object, binary logistic regression only provides two possible answers. This concept is usually represented as 0 or 1 in the encoding. For example, assessing cancer risk (high or low result) [24]. Another type is multinomial logistic regression. This model provides several classes that can be classified as items. There is a set of three or more classes that have been defined and prepared before running the model. For example, predicting whether a student will go on to college, trade school, or into the world of work. And the last type is ordinal logistic regression. This type is also a model where there are several classes that can be classified as items, but need class sorting. Classes do not need to be proportional and the distance between each class can vary. For example, rating a restaurant on a scale of 0 to 5 stars. For this study we used multinomial regression because the terms of the word feature are more suitable when classified than ordinalized.

2.2.3. Bernoulli NB

The Bernoulli NB method (Naive Bayes) is an algorithm used to group text into document classes. The results of the classification of documents by the Naive Bayes method can classify documents with a good level. This shows that the Naive Bayes method in classifying a document is not optimal. In the classification of texts using the Naive Bayes classifier, there is one model that can help us group documents, namely Bernoulli NB.

Bernoulli is a Naive Bayes classifier function uses binary elements to take the value of 1 if the word matches the document and 0 if the word does not match [25]. The Bernoulli NB method is a method for classifying a text from document categories. The results of the category document classification test using the Bernoulli-based Naive Bayes method can classify category documents with a good level of precision. Bernoulli is a Naive Bayes classifier function uses binary elements to take the value 1 if the appropriate word is found in the document and 0 if the word does not exist.

2.2.4. Support vector clustering

Large amounts of text data can potentially produce misinformation, so you need to process the text with good methods. This big data can come from various sources in the field of business processes such as finance, management, and others. Such large data can be applied to classification into sentiment analysis. One method that can be used for class classification in sentiment analysis is the data mining method of clustering data. Testing with the clustering algorithm is done to see the results of grouping using the basis of guided learning in sentiment analysis work.

Sentiment analysis work carried out using support vector clustering so far has provided good results rivaling other guided learning [21]. SVC algorithm, maps point from space to high-dimensional features using a Gaussian kernel. In the feature space, look for the constraints surrounding the data image. SVC’s limitations are compensated by SVM in a non-linear way. The difference between SVM and SVC is that the SVC hyperplane classifies datasets linearly. SVM dataset with non-linear approach. SVC returns the “best match” hyperplane then looks at what the “predicted” class is. SVC implements linear kernel functions to perform classification and works well with large sample sizes. SVC has additional features such as parameter normalization [26].

2.3. Vectorizer

Vectorization is used to represent the dimensions of words used in text processing. Many word vector formats can be formed as word features to be subjected to the analysis of sentiment classification work. The following are some of the vector formats tested in this study so that the effect of each feature format produced on the work of sentiment analysis with the methods we have determined previously [21].

For this research, we use three types of vectorizers, namely count vectorizer, TFIDF and hashing vectorizer. The basis for selecting these three vectorizers is based on the fact that the three types of vectorizers are types of text data vectorizers that have different calculation bases. In this study, all types of vectorizers that have been made will be examined for their performance in the four types of classifiers mentioned in the section 2.

2.3.1. Count vectorizer

Vectorization involves counting the number of occurrences of each word that appears in a document from various sources such as articles, books, and even paragraphs that can be used as word features. This type of vector is the simplest and most common word feature format used in various studies [21]. Count vectorizer is the simplest vectorizer, this technique is done by counting mode data on all text data tokens used.

Although this calculation is very simple, this method is still included as a vectorizer candidate which produces good scores in document classification. Count vectorizer, is a way to vectorize sentences. The goal is to take the words from each sentence and create a vocabulary of all the unique words in the sentence [27]. This vocabulary can then be used to create a feature vector of the number of words. Count vectorizer, uses the scikit-learn library to vectorize sentences. This library will take the words from each sentence and create a vocabulary of all the unique words in the sentence.

2.3.2. Term frequency-inverse document frequency vectorizer

The TFIDF method is a way of weighting the relationship of a word to each document [9], [28], [29]. TFIDF is a statistical method used to evaluate a word is in a document. In large documents, the most successful and widely used scheme for assigning term weights is the TFIDF term weighting scheme [21]. The thing to note in finding information from a heterogeneous collection of documents is the weighting of terms.

The term can be in the form of words, phrases or other. TFIDF vectorizer has units of indexing results in a document that can be used to determine the context of the document, then for each word given an indicator, namely the term weight. Basically, this TFIDF works to find the relative frequency of a word so that it can be compared with the proportion of said word in the entire document file.

2.3.3. Hashing vectorizer

This text vectorizer implementation uses hashing tricks to find the string token. The hash function used is a signed 32-bit version of Murmurhash3. One method of matching words using the hashing method is the Rabin-Karp algorithm. Word matching methods are still quite often used today by using word weights based on hash values, one of which is the Rabin-Karp method [30]. The hashing vectorizer method is still used today in experiments on the application of document classification, sentiment analysis and other text data processing.

Like the two methods previously mentioned, the count vectorizer and the TFIDF vectorizer, the hashing vectorizer is still a method capable of providing high relevance to data class labels with good accuracy. However, semantically speaking, TFIDF as a word that occurs in many documents is not a good differentiator, and should be given less weight than it appears in many documents. The hashing vectorizer works by applying a hash function to the features and then serving the hash value as an index directly, rather than looking up the index in an associative array. Hashing is the process of converting a certain key or character set into another value. This is usually represented by a shorter, fixed-length value or key that represents and makes it easier to find or use the original string. The most popular use of hashing is the implementation of hash tables. One of the main uses of hashing is to compare two files for similarities. Without opening two document files to compare them verbatim, the hash values calculated from these files will allow the owner to know immediately if they are different.

2.4. Experimental of vectorizer and classifier

This section describes the experiments carried out in this study by testing several vectorizer and classifier methods. The types of methods we use as vectorizers are count vectorizer, TFIDF vectorizer and hashing vectorizer. Each type of vectorizer method is then tested on each of the following classifier methods, namely random forest, logistic regression, Bernoulli NB, and SVC. In general, the classifier method that we do is a supervised learning technique, divided into training and testing stages. The following image is an overview for the training section.

Figure 2 illustrates the training flow of Twitter social media data used for the training process. The data obtained in the form of text data that has been structured into a dataset to be manually labeled analysis of sentiment analysis. The labels used in the classification in this study are positive, negative sentiment labels and neutral labels. The experimental process was carried out on data from social media relating to community sentiment towards the service and performance of the government of a city on the Indonesian island of Sumatra. Text data from Twitter tweets related to this is then carried out manual labeling to be chosen so that it becomes a training model used for the next process.

This general training flow is carried out on all classifiers with each vectorizer that has been determined. The selection of random forest is because this algorithm carries out bagging of classes produced by leaves from a set of decision trees with the most optimal class results. In this research, the decision tree that we use is C4.5. Logistic regression selection for a method based predictive analysis. NB Bernoulli for comparison with probability-based methods and SVC, as a clustering-based SVM method where each vector is labeled randomly until it reaches convergence. Furthermore, the testing phase can be seen in the next section.

In Figure 3 the flow of the data testing process from Twitter social media is illustrated using the model generated by the training process. The data obtained in the form of text data that has been structured to be used as test data does not yet have a sentiment class label. This testing process is carried out to determine the label of the data class-tested to find out whether the label has a positive, negative or neutral sentiment class on the services and performance of the tested city government. As with the testing process in general, the data tested will be adapted to each type of test for all classifier variants used. And of course, all the datasets tested have been changed in the form of each vectorizer that has been initialized.

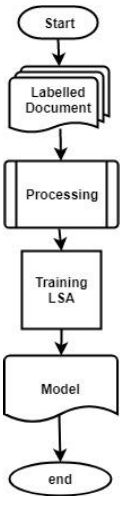


Figure 2. Flowchart for training sentiment analysis

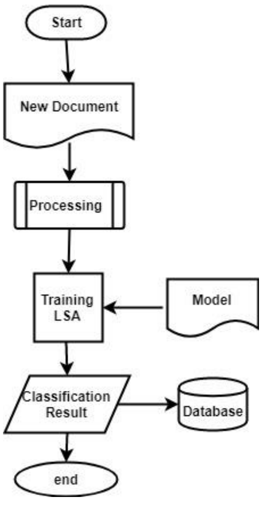


Figure 3. Flowchart for testing sentiment analysis

3. RESULTS AND DISCUSSION

This section describes the results and discussion of the experiments carried out based on the flow of research patterns that have been designed in the previous section on testing social media text data using several classifier methods and vectorizer methods. The tests conducted in this study include testing classifiers with variations of vectorizers. Classifier which is a method of classifying sentiments classifies sentiments for the text data used in this study. Following are the types of classifier methods used, among others:

- Random forest
- Logistic regression
- Bernoulli NB
- Support vector clustering

For the vectorizer method used, among others:

- Count vectorizer
- TFIDF vectorizer
- Hashing vectorizer

The following is a table of the results of the accuracy and confusion matrix produced on the random forest classifier algorithm for each variation of the vectorizer tested. Table 1 is the result of testing text data from Twitter related to sentiment analysis of Jambi city government with the random forest algorithm experiment. All types of process feature with 3 vectorizer variants were tested. Test results are stored in the confusion matrix with each evaluation result being true positive (TP), false positive (FP), true negative (TN) and false negative (FN).

Table 1 shows the results of the sentiment process resulting from an analysis of the government which has not yielded a maximum value. The evaluation carried out in the confusion matrix contains a type 1 or FN evaluation error which is still involved in calculating accuracy. Therefore, the results of sentiment analysis accuracy are still in the range of values that are not optimal. The Table 1 shows the experimental results of twitter data that have been scrapped for sentiment towards the Jambi city government as many as 500 tweets using a classifier algorithm with three types of vectorizers. In the random forest classifier algorithm, the result of using the highest vectorizer is generated by TFIDF. The following is Table 2, this table contains the results of the accuracy and confusion matrix produced on the logistic regression algorithm for each variation of the vectorizer tested.

Table 1. Random forest classifier accuracy in vectorizer variations

No	Random forest classifier		
	Vectorizer	Confusion matrix	Accuracy
1, A	Count vectorizer	[[2451 724] [120 599]]	0.783256291730868
1, B	TFIDF vectorizer	[[2449 713] [122 610]]	0.785567539804828
1, C	Hashing vectorizer	[[2492 809] [79 514]]	0.7719568567026194

The Table 2 the results of the twitter data experiment that has been scrapped for sentiment towards the Jambi city government as many as 500 tweets using a classifier algorithm with three types of vectorizers. In the logistic regression algorithm, the result of using the highest vectorizer is generated by TFIDF. The classifier and vectorizer test results have better results than the random forest vectorizer test. All vectorizer results produce accuracy values above 86% in all. Although there has not been an increase in the value of accuracy. Next, the results of testing text data for sentiment analysis of Jambi city government using the Bernoulli NB classifier are shown in Table 3. The Table 3 contains the results of the accuracy and confusion matrix produced on the Bernoulli NB classifier algorithm on each variation of the vectorizer tested.

Table 2. Accuracy of logistic regression in vectorizer variations

No	Logistic regression		
	Vectorizer	Confusion matrix	Accuracy
2, A	Count vectorizer	[[2372 346] [199 977]]	0.8600410888546481
2, B	TFIDF vectorizer	[[2418 325] [153 998]]	0.8772470467385721
2, C	Hashing vectorizer	[[2412 321] [159 1002]]	0.8767334360554699

The Table 3 shows the experimental results of twitter data that have been scrapped for sentiment towards the Jambi city government as many as 500 tweets using a classifier algorithm with three types of vectorizers. In the Bernoulli NB algorithm, the highest vectorizer results are produced by TFIDF. The value of the evaluation results with the classifier and vectorizer has unstable results. Count and TFIDF vectorizers have values above 80%, while Hashing vectorizers reach 66%. The following is a table of the results of the accuracy and confusion matrix produced on the SVC algorithm for each variation of the vectorizer tested.

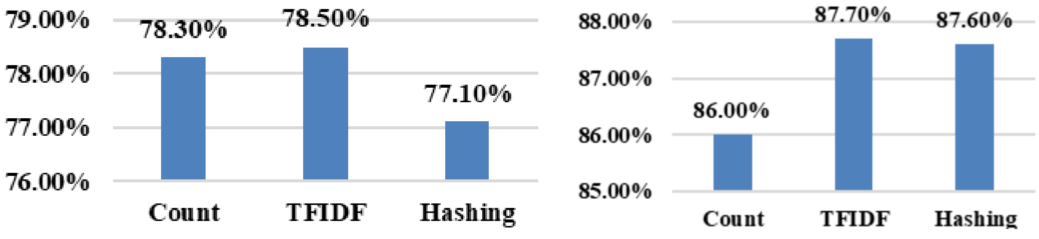
The Table 4 the experimental results of twitter data that has been scrapped for sentiment towards the Jambi city government as many as 500 tweets using a classifier algorithm with three types of vectorizers. In the SVC algorithm, the result of using the highest vectorizer is generated by the count vectorizer. The evaluation results with the SVC classifier have unfavorable values, even though they use the same vectorizer as the other tests in Table 1 to Table 3. From all the classifier tests conducted, you can see the results of the comparison of the performance of each type of vector classifier used. The following Figure 4 and Figure 5 shows the results of the performance of the random forest classifier, logistic regression, Bernoulli NB and SVC on the vectorizer method used, namely the count vectorizer, TFIDF vectorizer, and hashing vectorizer.

Table 3. Bernoulli NB accuracy in vectorizer variations

No	Bernoulli NB		
	Vectorizer	Confusion matrix	Accuracy
3, A	Count vectorizer	[[2323 491] [248 832]]	0.810220852593734
3, B	TFIDF vectorizer	[[2323 491] [248 832]]	0.810220852593734
3, C	Hashing vectorizer	[[2567 1311] [4 12]]	0.6623009758602979

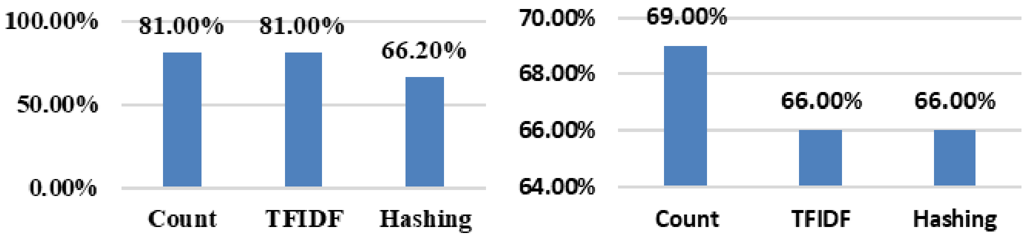
Table 4. SVC accuracy in vector variation

No	SVC		
	Vectorizer	Confusion matrix	Accuracy
4, A	Count vectorizer	[[2548 1182] [23 141]]	0.6905495634309193
4, B	TFIDF vectorizer	[[2571 1323] [0 0]]	0.6602465331278891
4, C	Hashing vectorizer	[[2571 1323] [0 0]]	0.6602465331278891



(a) (b)

Figure 4. Comparison graph of vectorizer on (a) random forest classifier and (b) logistic regression

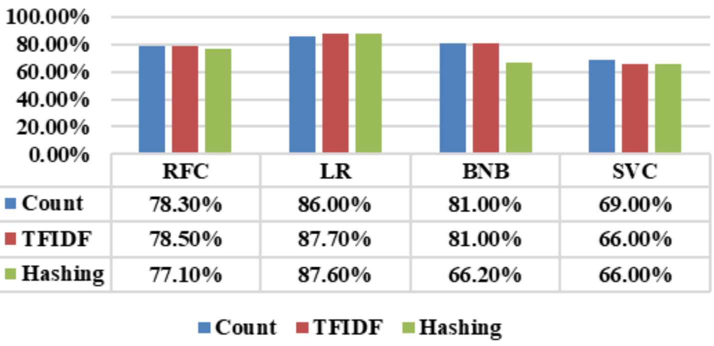


(a) (b)

Figure 5. Vectorizer comparison chart on the (a) Bernoulli NB and (b) SVC

From Figure 4(a), Figure 4(b), Figure 5(a), and Figure 5(b) the highest accuracy result is obtained by a linear regression algorithm on the TFIDF vectorizer with a value of 87.70% (Figure 4(b)) and the lowest accuracy by SVC on hashing vectorizer type with a value of 66.20% (Figure 5(b)). The best classifier performance on the count type vectorizer is occupied by a linear regression classifier with a value of 86.00% (Figure 4(b)). For the best classifier performance on the TFIDF vectorizer type occupied by a linear regression classifier with a value of 87.70% (Figure 4(b)). In hashing type vectorizer, classifier performance is occupied by linear regression with a value of 87.60% (Figure 4(b)). It can be said that the best classifier is logistic regression with stable accuracy results in all types of vectorizers. All tests carried out still tolerate type 1 prediction errors (false positive) and type 2 prediction errors (false negative). Each classifier has also not done hyperparameter tuning. All algorithms used are still the original parameters. This could be a gap for further research such as reducing the prediction results of type 1 errors or suppressing FP values. Gaps for increasing accuracy can also be achieved by setting or tuning all hyperparameters. One thing that needs to be explored is why the SVC results have a very low value. Indeed, the SVC type classifier is an attempt to unsupervised a labeled text data that removes all of its classes. However, the sentiment analysis work on the Jambi city government that was tested was not suitable for use with SVC unsupervised learning.

The comparative accuracy in the Figure 6, on average, the best vectorizer produced in the experiment is TFIDF, the second position is occupied by the count vectorizer and the last position is the hashing vectorizer. This, very in line with the way each method works, TFIDF can produce the best accuracy value because the basis for calculating TFIDF is to calculate every word frequency in each document, and this is certainly mathematically more relevant in finding the connection between words in class labels. Compared to just counting the frequency of words as the count vectorizer does. For hashing vectorizer, the calculation has a different way, namely by adding a hash value to each word feature. The result of the lowest hashing calculation in the experiment could be because of this, in addition to word processing, the use of hashing vectorizer is very rarely applied as a vectorizer.



	RFC	LR	BNB	SVC
🟦 Count	78.30%	86.00%	81.00%	69.00%
🟥 TFIDF	78.50%	87.70%	81.00%	66.00%
🟩 Hashing	77.10%	87.60%	66.20%	66.00%

Figure 6. Comparative accuracy of classifier results against vectorizer

4. CONCLUSION

Based on testing that has been done on the variation of classifiers and vectorizers variation, it is obtained that the best classifier is a linear regression algorithm based on predictive adaptive compared to random forest method based on the decision tree, probability-based Bernoulli NB, and SVC that works by doing clustering. Whereas from the test results on the vectorizer count, TFIDF, and hashing, the best accuracy is achieved on the TFIDF vectorizer. In this case, it means that the TFIDF vectorizer has a better value in presenting the word feature dimensions. The determination of the training data can affect the test results, because the pattern of the training data is used as a rule to determine the class in the testing data. So that the percentage of precision, recall, and accuracy is also influenced by the determination of the training data. Suggestions for future research can be done to reduce the prediction results of type 1 errors or suppress FP values. It can also be traced to a gap in which improvement in accuracy can also be achieved by setting or tuning all of the classifier hyperparameters.

REFERENCES

[1] S. Saad and B. Saberi, “Sentiment Analysis or Opinion Mining : A Review,” *International Journal on Advanced Science, Engineering and Information Technology*, vol. 7, no. 5, pp. 1660-1666, 2017, doi: 10.18517/ijaseit.7.4.2137.

[2] K. Park, J. S. Hong, and W. Kim, “A Methodology Combining Cosine Similarity with Classifier for Text Classification,” *Applied Artificial Intelligence*, vol. 34, no. 5, pp. 396–411, 2020, doi: 10.1080/08839514.2020.1723868.

[3] H. T. Sueno, B. D. Gerardo, and R. P. Medina, “Dimensionality Reduction for Classification of Filipino Text Documents based on Improved Bayesian Vectorization Technique,” *International Journal of Advanced Trends in Computer Science and*

Engineering, vol. 9, no. 5, pp. 7526–7531, 2020, doi: 10.30534/ijatcse/2020/87952020.

[4] X. Yang, K. Yang, T. Cui, M. Chen, and L. He, “A Study of Text Vectorization Method Combining Topic Model and Transfer Learning,” *Processes*, vol. 10, no. 2, 2022, doi: 10.3390/pr10020350.

[5] M. Lupei, A. Mitsa, V. Repariuk, and V. Sharkan, “Identification of authorship of ukrainian-language texts of journalistic style using neural networks,” *Eastern-European Journal of Enterprise Technologies*, vol. 1, no. 2, pp. 30–36, 2020, doi: 10.15587/1729-4061.2020.195041.

[6] J. Chen, H. Chen, Z. Wu, D. Hu, and J. Z. Pan, “Forecasting smog-related health hazard based on social media and physical sensor,” *Information Systems*, vol. 64, pp. 281-291, 2017, doi: 10.1016/j.is.2016.03.011.

[7] P. Chen, Z. Sun, L. Bing, and W. Yang, “Recurrent attention network on memory for aspect sentiment analysis,” in *Proceedings ofthe 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 452–461, doi: 10.18653/v1/d17-1047.

[8] H. Yousuf and S. Salloum, “Survey analysis: Enhancing the security of vectorization by using word2vec and CryptDB,” *Advances in Science, Technology and Engineering Systems Journal*, vol. 5, no. 4, pp. 374–380, 2020. [Online]. Available: https://www.researchgate.net/publication/343301593_Survey_Analysis_Enhancing_the_Security_of_Vectorization_by_Using_word2vec_and_CryptDB

[9] C. Baziotis, N. Pelekis, and C. Doulkeridis, “DataStories at SemEval-2017 Task 4: deep LSTM with attention for message-level and topic-based sentiment analysis,” in *Proceedings ofthe 11th International Workshop on Semantic Evaluations (SemEval-2017)*, 2017, pp. 747–754. [Online]. Available: https://aclanthology.org/S17-2126.pdf

[10] Z. E. Khatab, A. Hajihoseini, and S. A. Ghorashi, “A fingerprint method for indoor localization using autoencoder based deep extreme learning machine,” *IEEE Sensors Letters*, vol. 2, no. 1, pp. 1-4, 2018, doi: 10.1109/lsens.2017.2787651.

[11] A. B. Adege, H. P. Lin, G. B. Tarekegn, Y. Y. Munaye, and L. Yen, “An indoor and outdoor positioning using a hybrid of support vector machine and deep neural network algorithms,” *Journal of Sensors*, 2018, doi: 10.1155/2018/1253752.

[12] H. R. Moon and M. Weidner, “Dynamic linear panel regression models with interactive fixed effects,” *Econometric Theory*, vol. 33, no. 1, pp. 158–195, 2017, doi: 10.1017/S0266466615000328.

[13] M. A. Nishi and K. Damevski, “Scalable code clone detection and search based on adaptive prefix filtering,” *Journal of Systems and Software*, vol. 137, pp. 130–142, 2018, doi: 10.1016/j.jss.2017.11.039.

[14] Fachruddin, Y. Pratama, E. Rasywir, D. Kisbianty, Hendrawan, and M. R. Borroek, “Real time detection on face side image with ear biometric imaging using integral image and Haar-like feature,” in *2018 International Conference on Electrical Engineering and Computer Science (ICECOS)*, 2018, pp. 165-170, doi: 10.1109/ICECOS.2018.8605218.

[15] M. Saeidi, G. Bouchard, M. Liakata, and S. Riedel, “SentiHood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods,” in *COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016*, 2016, doi: 10.48550/arXiv.1610.03771.

[16] D. Tang, B. Qin, X. Feng, and T. Liu, “Effective LSTMs for target-dependent sentiment classification,” in *COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers*, 2016, doi: 10.48550/arXiv.1512.01100.

[17] M. Cliche, “BB_twtr at SemEval-2017 Task 4: Twitter sentiment analysis with CNNs and LSTMs,” in *Proc. of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017, pp. 573–580, doi: 10.18653/v1/s17-2094.

[18] A. E. Maxwell, T. A. Warner, and F. Fang, “Implementation of machine-learning classification in remote sensing: An applied review,” *International Journal of Remote Sensing*, vol. 39, no. 9, pp. 2784–2817, 2018, doi: 10.1080/01431161.2018.1433343.

[19] Y. Zeng, Y. Lan, Y. Hao, C. Li, and Q. Zheng, “Leveraging multi-grained sentiment lexicon information for neural sequence models,” *arXiv Computation and Language*, 2018, doi: 10.48550/arXiv.1812.01527.

[20] A. H. Salamah, M. Tamazin, M. A. Sharkas, and M. Khedr, “An enhanced WiFi indoor localization System based on machine learning,” in *2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2016, pp. 1-8, doi: 10.1109/IPIN.2016.7743586.

[21] L. Zheng, H. Wang, and S. Gao, “Sentimental feature selection for sentiment analysis of Chinese online reviews,” *International Journal of Machine Learning and Cybernetics*, vol. 9, pp. 75–84, 2018, doi: 10.1007/s13042-015-0347-4.

[22] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, “A survey of deep neural network architectures and their applications,” *Neurocomputing*, vol. 234, pp. 11–26, 2017, doi: 10.1016/j.neucom.2016.12.038.

[23] R. Socher *et al*., “Recursive deep models for semantic compositionality over a sentiment treebank,” in *Proc. of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1631–1642. [Online]. Available: https://aclanthology.org/D13-1170.pdf

[24] A. Y. A. Amer and T. Siddiqui, “Detection of Covid-19 fake news text data using random forest and decision tree classifiers,” *International Journal of Computer Science and Information Security (IJCSIS)*, vol. 18, no. 12, pp. 88–100, 2020, doi: 10.5281/zenodo.4427204.

[25] Z. Wu, Q. Xu, J. Li, C. Fu, Q. Xuan, and Y. Xiang, “Passive indoor localization based on CSI and Naive Bayes classification,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 9, pp. 1566-1577, 2018, doi: 10.1109/TSMC.2017.2679725.

[26] S. Sohangir, D. Wang, A. Pomeranets, and T. M. Khoshgoftaar, “Big Data: Deep Learning for financial sentiment analysis,” *Journal of Big Data*, vol. 5, no. 3, 2018, doi: 10.1186/s40537-017-0111-6.

[27] Y. Wang, A. Sun, J. Han, Y. Liu, and X. Zhu, “Sentiment Analysis by Capsules,” in *Proc. of the 2018 World Wide Web Conference*, 2018, pp. 1165–1174, doi: 10.1145/3178876.3186015.

[28] R. E. Putri, A. Putera, and U. Siahaan, “Examination of Document Similarity Using Rabin-Karp Algorithm,” *International Journal of Recent Trends in Engineering and Research*, vol. 3, no. 8, pp. 196–201, 2017. [Online]. Available: https://www.researchgate.net/publication/319272358_Examination_of_Document_Similarity_Using_Rabin-Karp_Algorithm

[29] M. Azimpourkivi, U. Topkara, and B. Carbunar, “Camera Based Two Factor Authentication Through Mobile and Wearable Devices,” in *Proc. of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2017, vol. 1, no. 3, pp. 1–37, doi: 10.1145/3131904.

[30] E. Rasywir, Y. Pratama, Hendrawan, and M. Istoningtyas, “Removal of Modulo as Hashing Modification Process in Essay Scoring System Using Rabin-Karp,” in *2018 International Conference on Electrical Engineering and Computer Science (ICECOS)*, 2018, pp. 159-164, doi: 10.1109/ICECOS.2018.8605211.

BIOGRAPHIES OF AUTHORS

Setiawan Assegaff received the Doctor of Philosophy (Ph.D.), Information System Doctor of Philosophy (Ph.D.) from Information System Department at Universiti Teknologi Malaysia Universiti Teknologi Malaysia in 2010–2014. In addition, He is serving as Member of ISRG (Information System Research Group), Information System Department Coordinator, Information System Postgraduate Program Coordinator, Rector of Universitas Dinamika Bangsa Jambi. He research interests are in Information System, IT Government, COBID Framework, and Information Technology. He can be contacted at email: setiawanassegaff@unama.co.id.

Errissya Rasywir received the Bachelor degree (S.Kom) in computer science from the Sriwijaya University. She received the Master degree (M.T) in Informatics Master STEI from the Institut Teknologi Bandung (ITB). She is a Lecture of Computer Science in the Informatics Engineering, Dinamika Bangsa University (UNAMA). In addition, she is serving as Head of the research group (LPPM) on UNAMA. Her research interests are in data mining, artificial intelligent (AI), natural languange proccessing (NLP), machine learning, and deep learning. She can be contacted at email: errissya.rasywir@gmail.com.

Yovi Pratama received the Bachelor degree (S.Kom) in computer science from the Sriwijaya University. He received the Master degree (M.T) in Informatics Master STEI from the Institut Teknologi Bandung (ITB). He is a Lecture of Computer Science in the Informatics Engineering, Dinamika Bangsa University (UNAMA). In addition, he is serving as Information Technology Division (IT Division) on UNAMA. His research interests are in data mining, artificial intelligent (AI), natural languange proccessing (NLP), machine learning, and deep learning. He can be contacted at email: yovi.pratama@gmail.com.