Classification of Depression Expressions on Twitter Using Ensemble Learning with Word2Vec

Muhammad Reza Adi Nugraha[1*], Yuliant Sibaroni[2]

[1,2]School of Computing, Telkom University, Bandung, Indonesia
[1]muhammadrezaadi@student.telkomuniversity.ac.id (*)
[2]yuliant@telkomuniversity.ac.id

Abstract— One of the mental health disorders experienced by people is depression. Depression is a mental disorder characterized by feelings of sadness, loss of interest or pleasure in daily activities, and decreased cognitive function that can affect social life, work, and general physical health. Early detection is needed to prevent the occurrence of bad risks. One of the early detections can be done through social media. This is because social media is one of the tools used to channel expression. This research uses data taken from Twitter social media to create a machine learning model. Before model building, data pre-processing will be carried out using Word2Vec to convert text into a continuous vector representation. The algorithm used is ensemble learning by combining five algorithms: Logistic Regression, Decision Tree, K-Nearest Neighbor, Artificial Neural Network, and Support Vector Machine. The results show that different Word2Vec architectures can give the model another performance. Ensemble Learning can improve the performance of using a single model. The best results were obtained by using a data ratio of 90:10 using the Skip-gram architecture to get an accuracy value and f1-score of 94%.

Keywords—Depression Expression; Twitter Datasets; Ensemble Learning; Word2vec.

I. INTRODUCTION

Social media analysis can provide deeper insights into the behaviour's and habits of social media users and assist in understanding emerging social behaviours in society. One of the social behaviours that can be analysed is the tendency for people to express depression or other mental health problems through social media. For example, research by Jalonen found that young people tend to express themselves through social media because, compared to older people, they have more concerns about others who experience more pleasant experiences than they feel [1].

Depression is a mental disorder characterized by ongoing feelings of sadness, loss of interest or joy in daily activities, and impairments in cognitive and physical functioning that can affect social life, work, and overall physical health [2]. However, due to self-denial among some sufferers and lack of awareness of the issue in many places, depression can remain undiagnosed or untreated. Lack of diagnosis and treatment can further worsen the condition [3]. Therefore, it is necessary to detect early expressions of depression on expressions on social media to prevent worse risks.

Detecting depression through the Machine Learning-based BERT method on platforms like Twitter has been done before. This research aims to identify depressed individuals using BERT for binary classification. The optimal results were achieved with a 90:10 split between train and test data, a batch size 64, and 4 epochs, yielding a 71% accuracy. The study revealed that higher train data improves accuracy, while increased test data and batch sizes may lower accuracy. Values were obtained for accuracy (71%), precision (81%), recall (71%), and f1-Score (75%), all of which were determined through the utilization of a confusion matrix. The findings suggest more extensive datasets are needed in future BERT method research to enhance accuracy [4].

Previous research on detecting depressive expressions in social media texts was conducted using machine learning methods with supervised learning algorithms [5].

The Decision Tree algorithm has also been used to classify household appliances in the context of non-intrusive electricity load monitoring. The accuracy and f1-Score were 93.67% and 93.14%, respectively, but this algorithm has an overfitting problem caused by the high variance problem in the training process.

Unsupervised learning methods have also been carried out with the K-Nearest Neighbor (K-NN) algorithm [6]. In previous research, K-NN was used to classify the MNIST dataset. The study used error rate as an evaluation. The result was an average error rate of 23.42%. This algorithm has advantages when using non-linear data.

There is a bias and variance trade-off problem, which is when the complexity of the model increases, causing the bias to decrease and the variance to increase, which causes a U-shaped error curve. In addition, there is a problem with machine learning algorithms that cannot handle strings with plain text when performing natural language processing. Therefore, to overcome the problem of bias and variance trade-off, the author will conduct research using ensemble learning. Ensemble learning is a machine learning method that combines several algorithms intending to improve accuracy.

The previous research approach involves an ensemble classification system that evaluates the performance of eight different classifiers. Performance was assessed using various metrics, including accuracy, precision, recall, F1 score, F2 score, and F3 score. The top three classifiers were selected, and an ensemble method using a voting mechanism was introduced. The majority-based voting mechanism yielded the best F3 score,

outperforming other methods and demonstrating superiority over the state-of-the-art approaches [6]. The algorithms used in this research are Logistic Regression, Decision Tree, KNN, Artificial Neural Network, and Support Vector Machine. Then, to overcome the problem of processing strings with plain text, the author will use Word2Vec.

II. RESEARCH METHODOLOGY

In this study, the system built can classify expressions of depression on Twitter. Based on Figure 1, we developed a research flowchart that had been carried out by Ade by adjusting it to the methods used in this study [7].
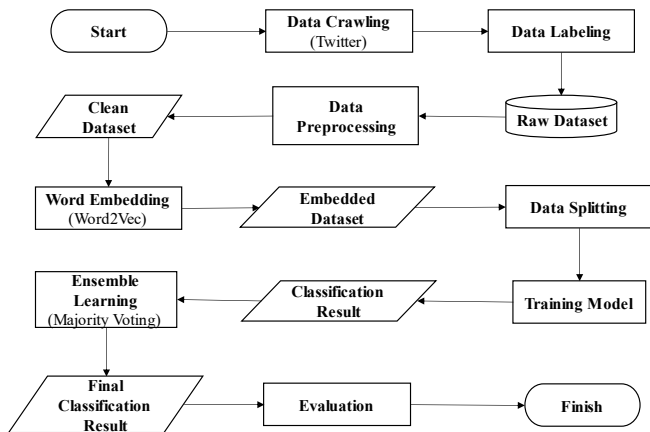

Figure 1. System Architecture

A. Data Crawling

To obtain tweet data, we distributed the Depression Anxiety Stress Scale (DASS) questionnaire to respondents. DASS is a scale for the measurement of three states: anxiety, depression, and stress [8]. We only used 14 questions to measure depression. We successfully crawled 14853 raw tweets from 48 Twitter accounts. Table I shows the questions we used to measure depression. Respondents will answer these questions on a scale of 0-3. Respondents will get a score that calculates their level of depression.

TABLE I
QUESTIONARE QUESTIONS

No	Question
1	Unable to see the positive in events.
2	Feeling like you can't do something anymore.
3	Pessimistic.
4	Feeling sad and depressed.
5	Loss of interest in many things (e.g., eating, socializing)
6	Feeling unworthy.
7	Feeling that life is worthless.
8	Not being able to enjoy the things I do.
9	Feeling lost and hopeless.
10	It's hard to be enthusiastic about things.
11	Feeling worthless.
12	There is no hope for the future.
13	Feeling like life is meaningless.
14	It is difficult to increase initiative in doing things.

After getting the level of depression from the respondent, crawling tweet data on the respondent's account is carried out. Table II shows examples of tweet data that was successfully crawled.

TABLE II
EXAMPLE OF CRAWLING RESULTS

User	Tweet
User 1	*Persetan kenyamanan ini, aku harus bisa melakukan semua tanpa kenyamanan. jangan pernah lari hanya karena keringat bercucuran, internet selambat siput, dan telinga dipekakan. itu ketidaknyamanan yang sepele! hei bajingan, teruslah coba kerjakan dan lanjutkan!* Never mind this comfort. I should be able to do everything without comfort. Never run away just because the sweat is pouring, the internet is slow as a snail, and the ears are cracking. It's a trivial inconvenience! Hey, asshole, keep trying to do it and keep going!
User 2	*Itu udah diluar kendali kita, kita cuma bisa ngelakuin apa yang kita bisa, selebihnya biarkan orang lain menilai.* It's already out of our control; we can only do what we can, and the rest let others judge.

B. Data Pre-processing

Data We perform the following steps in the data pre-processing process: case folding, stopword removal, typo word correction, elongation replacement, emoji change, cleaning, lemmatizing, and tokenization. Case folding is converting text to be all capitalized or non-capitalized. Stopword removal is the process of removing words in the text that have no significant meaning. We use the Indonesian stopwords dictionary from the Natural Language Toolkit (NLTK) library. Table III stop word examples based on the NLTK library. Table III shows stop word examples based on the NLTK library. The word in the tweet that contains the stopword will be removed.

TABLE III
STOP WORDS EXAMPLE

Process	Words
Stopwords Removal	*daripada, ujarnya, beginian, suatu, sebesar*

Typo word correction corrects typographical errors in a text. We used the corpus of Indonesian Wikipedia articles as a reference. Elongation replacements are replacing repeated characters (e.g., sending to sensing). Emoji change is replacing emojis with their corresponding textual representations. Cleaning is removing symbols, numbers, punctuation, and white space. Lemmatizing is text conversion by removing the affixes on a word into its base word only. We use the Indonesian lemmatize provided by the NLTK library. Tokenization is converting each base word into smaller units called tokens.

Table IV shows examples of typo words, elongations, and emojis. Each word in the tweet that contains a word in the example in the table will be corrected or changed into a more meaningful form.

TABLE IV
PRE-PROCESSING EXAMPLE

Process	Example
Typo Word	*snang, sdih, mals, perig*

Process	Example
Elongation	*senangg, sedihh malass, peergii,*
Emoji	*:) , :(, :O, \(>_<\)*

Table V shows some examples of the lemmatizing process. Every word in the tweet that is still not in its basic form will be converted into its basic form by lemmatizing.

TABLE V
LEMMATIZING EXAMPLE

Before	After
panasnya	*panas*
berjalan	*jalan*
bermain	*main*

C. Data Libelling

Each DASS question was answered using a 0-3 scale, with 0 equal to not applicable at all and three equals to very applicable most of the time. The range of possible scores for each scale is 0-42. We categorized the depression category into two groups: scores in the range of 0-13 as non-depressed and 21 and above as depressed [8]. Based on the 48 people we used in this study, the number of people categorized as depressed was 22 people, and non-depressed was 26 people. We use the transformers API of the hugging face to perform positive and negative sentiment labelling on the crawled dataset.

Once the dataset has a label, we take the negative sentiment tweets of Twitter accounts that belong to the depressed category as the depression class, and we take the positive sentiment tweets as the non-depression class. After the process, the number of tweets that can be used is 10876, with the same amount of data for depression and non-depression classes.

D. Data Splitting

The data that has been word-embedded will then be divided into two parts, namely training data and testing data. Data division will use the sickit-learn library. In this study, a data ratio of 90:10 is used with the same amount of data for depression and non-depression classes.

E. Word Embedding

Representing words as numerical vectors based on the context in which they appear has become a necessary method for analyzing text with machine learning [9]. This process is called word embedding. One of the words embedding models is Word2Vec. Word2Vec provides semantics for each word in a vector valuable representation in machine-learning text classification [10].

Based on Figure 2, Word2Vec can find analogies, syntax, and semantic analysis of words. There are two architectures in Word2Vec, namely Continuos Bag-of-Word (CBOW) and Skip-gram. The main difference between the two is that CBOW predicts the target word based on the context of the surrounding words, while Skip-gram predicts the context word based on the target word. Figure 2 shows the different architectural forms of CBOW and Skip-gram.
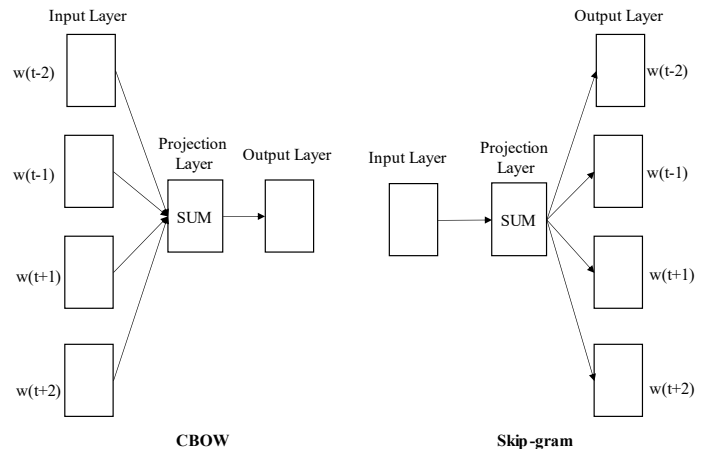

Figure 2. CBOW and Skip-gram Architecture

F. Training Model

Decision Tree: A decision Tree is a machine learning algorithm that is supervised learning. Supervised learning means that the data used already has a label in building a model. The Decision Tree can be used to perform classification by creating a flowchart that will determine the class that matches the given input value [11]. The decision tree model is constructed with a Gini impurity criterion. Gini impurity is one such measure and is commonly used. It quantifies the likelihood of misclassifying a randomly chosen element in the dataset. A maximum depth of 50 means that the tree will continue splitting nodes until it reaches a depth of 50 or until further splitting does not improve the impurity. Allowing leaf nodes with a minimum of 1 sample specifies the minimum number of samples required for a node to be a leaf. Requiring a minimum of 2 samples for node splitting helps prevent the tree from making splits on very small subsets of the data, which can lead to overfitting.
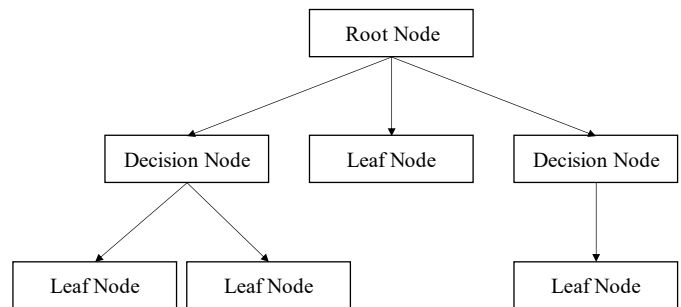

Figure 3. Decision Tree Architecture

Based on Figure 3, the classification process starts from the root node of the Decision Tree and then recursively continues until it reaches a leaf node with a certain class label. A split condition is applied at each node to decide whether the input value should continue to the left or right until it reaches the leaf node. The final decision will be in the leaf node [12].

Logistic Regression: Logistic Regression is one of the statistical methods used in the application of machine learning. This method can perform analysis for binary outputs that have

the same level. Logistic Regression works by calculating the probability of an event that can occur divided by the probability of an event that cannot occur [13]. The following is the Equation (1) used in Linear Regression. The logistic regression model is configured with a regularization strength (C) of 0.1 c. A smaller value of C corresponds to stronger regularization. In this case, a value of 0.1 suggests relatively strong regularization, and the L1 penalty encourages sparsity in the model by adding the absolute values of the coefficients to the cost function. This can lead to some coefficients being exactly zero, effectively performing feature selection, and the 'lib-linear' solver is the algorithm used for optimization during the training of the logistic regression model. It enables a warm start to reuse the existing solution for increased efficiency in fitting the model to subsequent data. The *P* variable is the probability of an event occurring, *e* is the base of the natural logarithm (approximately 2.71828), *a* and *b* are constants, and *X* is the independent variable. The function ensures that the predicted probabilities lie between 0 and 1, making it suitable for modelling probabilities in a logistic regression context.

$$P = \frac{e^{a+bX}}{1+e^{a+bX}} \qquad (1)$$

K-Nearest Neighbor (K-NN): K-NN is one of the machine learning algorithms in modelling using data with no previous label or unsupervised learning. K-NN is used for classification, but it is also chosen when the parameters of the probability density are unknown or difficult to determine. The KNN model is specified with the 'auto' algorithm for neighbor search, which automatically selects the most appropriate method for finding the nearest neighbors based on the input data and other parameters. The Euclidean metric is a measure of distance between points in Euclidean space. Setting the number of neighbors to 23 means the model will consider the labels of the 23 nearest neighbors to the query point when making a prediction. Employing distance-based weights for neighbors points means closer neighbors influence the prediction more than those farther away.
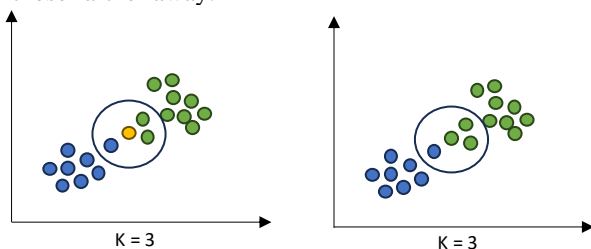

Figure 4. K-NN Architecture

Based on Figure 4, the K-NN algorithm performs classification works by determining the distance between the latest input data point and all existing data points in the K-NN model. It then selects the *K* closest data points. Finally, determine the class corresponding to the highest number of K nearest neighbors [14].

Artificial Neural Network (ANN): A computational model that draws inspiration from the architecture and operations of the human brain is known as an artificial neural network (ANN). The ANN model is configured with a rectified linear unit (ReLU) activation function, and it introduces non-linearity by outputting the input for all positive values and zero for all negative values. A regularization parameter (alpha) of 0.0001, a smaller alpha value, such as 0.0001 in this case, indicates relatively light regularization: two hidden layers, each containing 50 neurons. The number of neurons in a layer determines the capacity of the network to learn complex patterns from the data. Having two hidden layers allows the model to capture hierarchical representations of the input features.
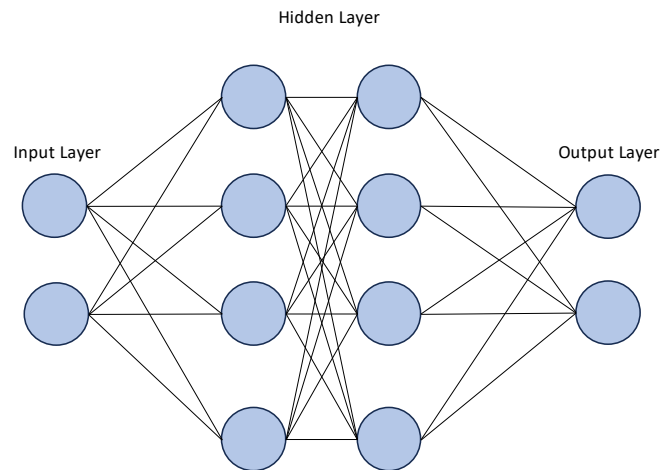

Figure 5. ANN Architecture

Based on Figure 5, the ANN comprises networked nodes that process and send information, referred to as neurons. Digital neurons are joined into a complex network that mimics the brain's structure [15].

Support Vector Machine (SVM): A machine learning algorithm for regression analysis and classification is called Support Vector Machine (SVM). A supervised learning technique that works with both linear and non-linear data. The SVM model was configured with a regulation value of 0.1, which controls the trade-off between achieving a low training error and a low testing error. A smaller value of C (such as 0.1 in this case) indicates a higher regularization, meaning the model is penalized more for misclassifying training examples. This can help prevent overfitting. A 'poly' kernel maps the input data into a higher-dimensional space, making it capable of capturing non-linear relationships between features., and a gamma parameter of 0.1. A small gamma value (0.1) makes the decision boundary more flexible, allowing it to capture intricate patterns in the training data. It essentially defines the influence of a single training example's features on the decision boundary.

Based on Figure 6 SVM, the optimal hyperplane to divide the data into distinct classes must first be identified. The margin between the two classes is maximized by the selection of the hyperplane [16].
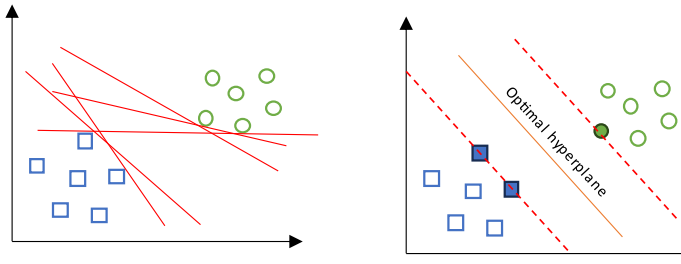
Figure 6. SVM Architecture

The model will provide output in binary classification results whether the tweet is an expression of depression or non-depression. The models' classification results developed in the past will be merged with ensemble learning and majority voting criteria based on the classes that occur the most frequently. Ensemble Learning is one machine learning method that combines several machine learning methods to improve accuracy.

There are three ways to do ensemble learning: bagging, Boosting, and Stacking [17]. But there are also ensemble learning methods that use majority voting rules. Bagging involves training the same algorithm multiple times using different subsets sampled from the training data. Boosting multiple models is trained sequentially, and each model learns from the errors of its predecessor. Stacking combines heterogeneous machine learning models (base learners) using other data mining techniques [18]. Meanwhile, majority voting rules take advantage of the shortcomings of one algorithm, which can be an advantage for other classifiers.
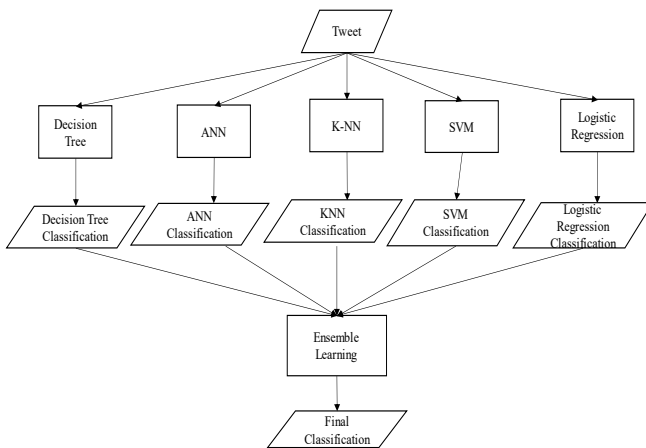

Figure 7. Ensemble Learning Flow Chart

Based on Figure 7, Majority voting rules combine classifier predictions into a final prediction output based on the majority of the output [19]. Majority Voting Rules provide a simple and interpretable method for ensemble learning. In contrast to Bagging and Boosting, which often entail repeated training of the same base algorithm or sequential adjustments based on errors, Majority Voting Rules enable the fusion of diverse base classifiers. This approach promotes a robust ensemble that is less susceptible to overfitting. Figure 3 shows the flow chart of our Ensemble Learning.

G. Evaluation

In machine learning and specifically for classification, the Confusion Matrix is a special table layout that allows visualizing the performance of an algorithm in making predictions. Each row of the matrix represents an instance in the actual class, while each column represents an instance in the prediction class [20]. There are four elements in the Confusion Matrix, namely True Positive (TP), a positive prediction class that corresponds to a positive actual class; True Negative (TN), a negative prediction class that corresponds to a negative actual class; False Negative (FN) a negative prediction class that does not correspond to a positive actual class, and False Positive (FP) a positive prediction class that does not correspond to a negative actual class.

Evaluation is carried out after the model is successfully created. The aim is to determine the performance of the classification model that has been made. The evaluation results will determine whether the model still needs improvement or not. This process is measured using a confusion matrix with four performance measures: accuracy, recall, precision, and f1-score. In evaluating the performance of a model, four critical measures provide different insights. Accuracy, which measures the ratio of correct and incorrect predictions, provides a general idea of the model's accuracy level.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \qquad (2)$$

Precision becomes relevant when False Positives impact predictions, focusing on how precise the model is in identifying positives.

$$Precision = \frac{TP}{TP+FP} \qquad (3)$$

Conversely, recall becomes significant when False Negatives have a crucial impact, indicating the model's ability to detect all positive cases.

$$Recall = \frac{TP}{TP+FN} \qquad (4)$$

F1-Score, as a harmonious blend of recall and precision, seeks the optimal balance between the two. By understanding the role of each measure, performance evaluation can be done holistically, providing an in-depth understanding of how effective a model or system is in handling a particular task or problem [21].

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision+ Recall} \qquad (5)$$

III. RESULT AND DISCUSSION

We perform the word embedding process with Word2Vec using the Gensim library by pre-training using a corpus of the

latest Indonesian Wikipedia articles taken on October 1, 2023. Our Word2Vec model uses a vector size of 300, the maximum number of words around the target word is 5, and the minimum number of occurrences of the considered word is 1.

In this study, we used CBOW and Skip-gram architectures for comparison. Using the Word2Vec model, we converted each token in the dataset into a vector form, and then we calculated the average value of all tokens in the tweet. It is planned to use the vector's average value as a vector representing the entire tweet. This vector will be utilized in the modelling process.

The first experiment was to perform parameter tuning to determine the best parameters for each machine learning model we used. This study used Decision Tree, Logistic Regression, K-NN, ANN, and SVM models before combining them with Ensemble Learning.

We conducted further experiments by performing classification using the previous model. We used a 90:10 data ratio and performed 10-fold cross-validation. 10-fold cross-validation is a technique used to assess the performance of a predictive model. The dataset is divided into ten equal sections, nine of which are used to train the model, and the remaining is used for testing. Every component is used as the test set precisely once during this operation's ten repetitions to ensure the model is not overfitting the data. [22]
We then took the average of the 10-cross-validation as the accuracy metric. Table VI shows the experimental results using CBOW architecture, and Table VII shows the results using Skip-gram architecture. Table VI shows that machine learning algorithms perform differently classifying depressive expressions using the CBOW architecture. The ANN model obtains the best performance.

TABLE VI
CBOW RESULT

Model	Accuracy	F1	Recall	Precision
Decision Tree	0.90	0.91	0.91	0.91
Logistic Regression	0.69	0.70	0.70	0.70
K-NN	0.92	0.92	0.92	0.92
ANN	*0.92*	*0.93*	*0.93*	*0.93*
SVM	0.91	0.91	0.91	0.91

Table VII shows that machine learning algorithms perform differently in classifying depressive expressions using the Skip-gram architecture. The ANN model also obtains the best performance.

TABLE VII
SKIP-GRAM RESULT

Model	Accuracy	F1	Recall	Precision
Decision Tree	0.91	0.91	0.91	0.91
Logistic Regression	0.70	0.70	0.70	0.70
K-NN	0.92	0.92	0.92	0.92
ANN	*0.93*	*0.93*	*0.93*	*0.93*
SVM	0.92	0.93	0.93	0.93

Each classification resulted from the five models before it was saved to determine which class occurred the most

frequently. The ensemble learning model's primary classification result is the class that appears the most often with the highest frequency. The Ensemble Learning classification results get higher accuracy when compared to the use of only a single model. Figure 4 shows the accuracy comparison of Ensemble Learning with the five models using CBOW, and Figure 5 shows the accuracy comparison of Ensemble Learning with the five models using Skip-gram. Based on Figure 8, Ensemble Learning gets the highest accuracy value compared to a single machine learning method with a value of 0.93.

Based on Figure 9, Ensemble Learning gets the highest accuracy value compared to a single machine learning method with a value of 0.94.
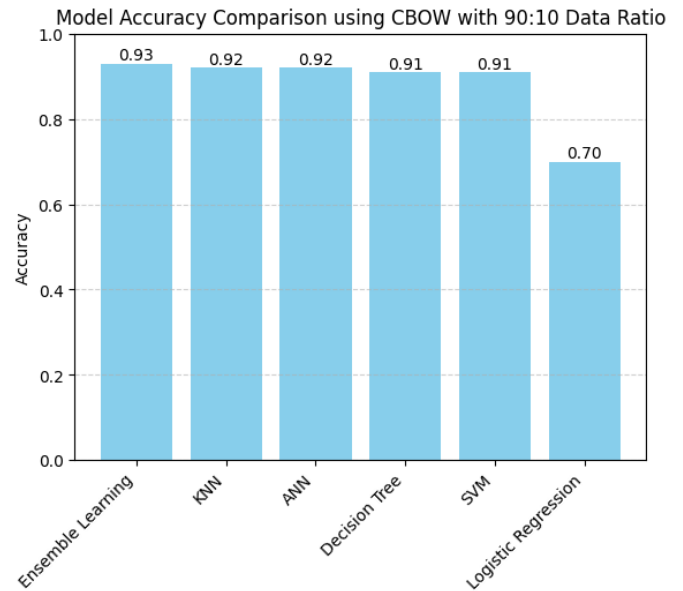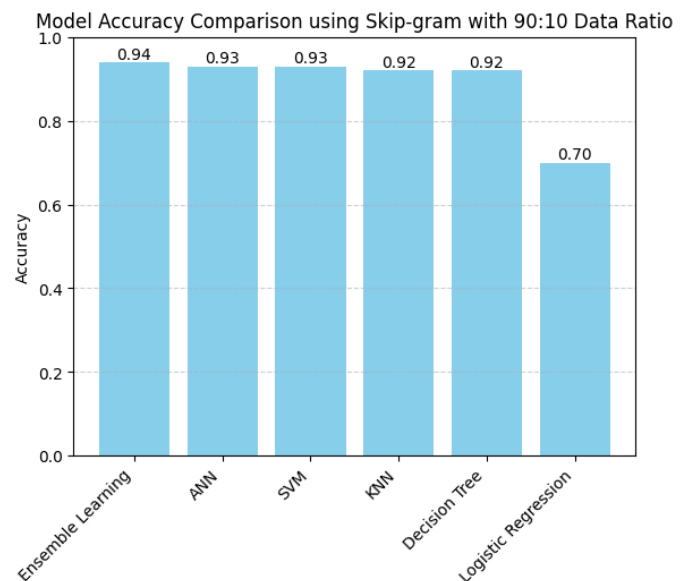

Figure 8. Ensemble Learning Result Using CBOW.

Figure 9. Ensemble Learning Result Using Skip-Gram

IV. CONCLUSION

In conclusion, our study successfully classifies depressive expressions on Twitter by word embedding using Word2Vec with CBOW and Skip-gram architectures on the Indonesian Wikipedia corpus. Afterward, ensemble learning was used to test the models. Decision Tree, Logistic Regression, K-NN, ANN, and SVM were the methods that were utilized. For each model, the parameters were tuned to their optimal values, and compared to the ANN model, the CBOW and Skipgram architectures displayed more excellent performance. Compared to single models, ensemble learning performed better. Using ensemble learning with the skip-grain architecture, which achieved accuracy and an f1-score of 0.94, was the method that yielded the most outstanding results. Ensemble learning methods such as bagging, boosting, and stacking are some approaches that could be utilized for future research. Word2Vec can also be compared to a variety of other Word embedding methods to determine which way is the most effective.

REFERENCES

[1] H. Jalonen, "Negative Emotions in Social Media as a Managerial Challenge," *Proc. 10Th Eur. Conf. Manag. Leadersh. Gov. (Ecmlg 2014)*, pp. 128–135, 2014.

[2] P. Tong *et al.*, "Clinical traits of patients with major depressive disorder with comorbid borderline personality disorder based on propensity score matching," *Depress. Anxiety*, vol. 38, no. 1, pp. 100–106, 2021, doi: 10.1002/da.23122.

[3] Y. M. Alsagri HS, "Machine Learning-Based Approach for Depression Detection in Twitter Using Content and Activity Features," *IEICE Trans. Inf. Syst.*, vol. E103-D, 2020, doi: 10.1587/transinf.2020EDP7023.

[4] F. Ilham and W. Maharani, "Analyze Detection Depression In Social Media Twitter Using Bidirectional Encoder Representations from Transformers," *J. Inf. Syst. Res.*, vol. 3, no. 4, pp. 476–482, 2022, doi: 10.47065/josh.v3i4.1885.

[5] T. T. H. Le, H. Kang, and H. Kim, "Household appliance classification using lower odd-numbered harmonics and the bagging decision tree," *IEEE Access*, vol. 8, pp. 55937–55952, 2020, doi: 10.1109/ACCESS.2020.2981969.

[6] A. S. Assiri, S. Nazir, and S. A. Velastin, "Breast Tumor Classification Using an Ensemble Machine Learning Method," *J. Imaging*, vol. 6, no. 6, 2020, doi: 10.3390/JIMAGING6060039.

[7] A. Bastian, Ardi Mardiana, and Dinda Sri Wulansari, "Sentiment Analysis Against Political Figure's Billboard During Pandemic Using Naïve Bayes Algorithm," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 7, no. 1, pp. 138–145, 2023, doi: 10.29207/resti.v7i1.4643.

[8] N. Bilgel and N. Bayram, "Depresyon anksiyete stres ölçeğinin (DASS-42) Türkçeye uyarlanmiş şeklinin psikometrik özellikleri," *Noropsikiyatri Ars.*, vol. 47, no.

2, pp. 118–126, 2010, doi: 10.4274/npa.5344.

[9] F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, and F. Rudzicz, "A survey of word embeddings for clinical text," *J. Biomed. Informatics X*, vol. 4, no. October, p. 100057, 2019, doi: 10.1016/j.yjbinx.2019.100057.

[10] S. Sivakumar, L. S. Videla, T. Rajesh Kumar, J. Nagaraj, S. Itnal, and D. Haritha, "Review on Word2Vec Word Embedding Neural Net," *Proc. - Int. Conf. Smart Electron. Commun. ICOSEC 2020*, no. Icosec, pp. 282–290, 2020, doi: 10.1109/ICOSEC49089.2020.9215319.

[11] S. Tangirala, "Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 2, pp. 612–619, 2020, doi: 10.14569/ijacsa.2020.0110277.

[12] Y.-S. . Loh, W.-Y., and Shih, "Split selection methods for classification trees," *Stat. Sin.*, pp. 815-840., 1997.

[13] M. P. LaValley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008, doi: 10.1161/CIRCULATIONAHA.106.682658.

[14] K. Taunk, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," *2019 Int. Conf. Intell. Comput. Control Syst. ICCS 2019*, no. Iciccs, pp. 1255–1260, 2019.

[15] K. Kar, S. Kornblith, and E. Fedorenko, "Interpretability of artificial neural network models in artificial intelligence versus neuroscience," *Nat. Mach. Intell.*, vol. 4, no. 12, pp. 1065–1067, 2022, doi: 10.1038/s42256-022-00592-3.

[16] Muhammad Afif Raihan and Erwin Budi Setiawan, "Aspect Based Sentiment Analysis with FastText Feature Expansion and Support Vector Machine Method on Twitter," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 4, pp. 591–598, 2022, doi: 10.29207/resti.v6i4.4187.

[17] A. Chauhan, A. Agarwal, and R. Sulthana, "Genetic Algorithm and Ensemble Learning Aided Text Classification using Support Vector Machines," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 8, pp. 260–267, 2021, doi: 10.14569/IJACSA.2021.0120830.

[18] L. Wen and M. Hughes, "Coastal wetland mapping using ensemble learning algorithms: A comparative study of bagging, boosting and stacking techniques," *Remote Sens.*, vol. 12, no. 10, 2020, doi: 10.3390/rs12101683.

[19] R. S. Khairy, A. S. Hussein, and H. T. H. S. ALRikabi, "The Detection of Counterfeit Banknotes Using Ensemble Learning Techniques of AdaBoost and Voting," *Int. J. Intell. Eng. Syst.*, vol. 14, no. 1, pp. 326–339, 2021, doi: 10.22266/IJIES2021.0228.31.

[20] S. Haghighi, M. Jasemi, S. Hessabi, and A. Zolanvari, "PyCM: Multiclass confusion matrix library in Python," *J.*

Open Source Softw., vol. 3, no. 25, p. 729, 2018, doi: 10.21105/joss.00729.

[21] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit.*, vol. 91, pp. 216–231, 2019, doi: 10.1016/j.patcog.2019.02.023.

[22] R. M. and S. Meena, "Empirical Validation of cross-version and 10-fold cross-validation for Defect Prediction," *2021 Second Int. Conf. Electron. Sustain. Commun. Syst.*, 2021, doi: 10.1109/ICESC51422.2021.9533030.