Scientific Journal of Informatics
Vol. 11, No. 4, Nov 2024

p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

Sentiment Analysis on SocialMedia Using TF-IDF Vectorization and H2O Gradient Boosting for Student Anxiety Detection

Maylinna Rahayu Ningsih[1]*, Jumanto[2]
[1, 2]Department of Computer Science, Universitas Negeri Semarang, Indonesia

Abstract.
Purpose: Mental health issues are now a concern for many people. Anxiety or often called Anxiety that is excessive and prolonged has also become the forefront of various psychological disorders that trigger impacts such as stress to suicide. People using social media platforms tend to be a medium for expressing opinions sharing information and even expressing daily emotions. Many studies have shown a correlation between expressing emotional statements on social media and mental disorders. This research aims to conduct sentiment analysis of Anxiety on social media using H2O Gradient Boosting by implementing TF-IDF Vectorization which is set to max feature.
Methods: This research utilizes 6980 post data from social media. The method applied is by conducting Exploratory Data Analysis then Data preprocessing, Tf-Idf Vectoriztion with max feature experiments 100, 250, 500, 1000 and 2000, H2O Gradient Boosting Model, Cross Validation, and Model performance evaluation.
Result: The results of this study show good model performance through max feature TF-IDF = 250 with an accuracy value of 99%, Specificity 99.57%, and Eror Rate of 0.0106.
Novelty: So that the use of the H2O Gradient Boosting model succeeded in providing good performance in classifying anxiety sentiment.

Keywords: Sentiment analysis, Anxiety, H20.ai, Gradient boosting, TF-IDF vectorization
Received January 2025 / **Revised** February 2025 / **Accepted** March 2025

INTRODUCTION

Mental health issues are now starting to become a concern for many people. In today's fast-paced and demanding world, it is common for people to be mentally disturbed, especially when they experience anxiety or worry in response to various situations. Anxiety and worry can happen to anyone, including students, because everyone has experienced unpleasant situations in various phases of life. Anxiety can be caused by various factors and certain situations that make you uncomfortable. When anxiety is at a moderate level, a person will focus on things that make them uncomfortable and begin to ignore other things. And when anxiety levels are high, a person's thinking will be disrupted and only focus on small things and ignore other things, thus not being able to think clearly [1]. Excessive anxiety is the most common mental disorder and usually occurs before or during adulthood [2].

Anxiety disorder in English is called anxiety disorder which in Latin angustus which means stiff, and anci, ango which means suffocating [3]. Anxiety disorders have become a part of everyday life that affects various aspects, whether individual or environmental. Excessive and prolonged anxiety has also been at the forefront of a variety of psychological disorders [4] which triggers impacts such as stress and suicide [5], even according to [6] around 90% of people who experience anxiety disorders are followed by stress and depression. The World Health Organization (WHO) revealed that stress and suicide are the second leading causes of death in individuals aged 15-29 years and an estimated 800 thousand people die every year [7], [8] due to suicide. According to [9] In 2020, the prevalence of anxiety disorders is around 4802.4 cases per 100,000 people, which is the highest of any mental illness.

Nowadays, social media is used to gauge public opinion with mental health being a hot-button issue these days. People using social media platforms such as Instagram, Twitter and other platforms tend to be a medium to express their opinions [10], sharing information and even expressing daily emotions. Recent

*Corresponding author.
Email addresses: maylinnarahayuningsih@students.unnes.ac.id (Ningsih)*, jumanto@mail.unnes (Unjung)
DOI: 10.15294/sji.v11i4.20582

research [11] showed a correlation between expressing emotional statements on social media and mental disorders. The use of social media as a place to express emotions has become a trend and the stigma continues to grow, thus contributing to user data about what they write. Data from social media about the expression of emotions to detect will be obtained if processed properly.

The process of extracting hidden information or patterns from large data is called data mining. One of the implementations of data mining is sentiment analysis which is used to evaluate sentiment or understand the feelings contained in texts such as expressions of feelings on social media or app/product reviews which involves the process of data mining [12], [13]. This involves natural language processing (NLP) and machine learning algorithms such as classification techniques to classify text into certain types of labels. Classification techniques can be used as a process of understanding how a person expresses their feelings through a particular social media post or comment.

Several studies have been conducted in classifying and detecting text for anxiety, showing the potential of machine learning in mental health. Overall, research on anxiety detection makes diverse and important contributions to the understanding of early psychological and health-related anxiety. Research on anxiety detection [14] has been conducted on Twitter social media using supervised machine learning including Naïve Bayes, Random Forest and LASSO-regression algorithms. The digital footprint of self-disclosed anxiety on twitter posts shows a high frequency of negative sentiment words. The best prediction accuracy in the Naïve bayes model is 81.1% for anxiety diagnosis. Other research was also conducted [15] using Natural Language Processing (NLP) by comparing the performance of LSTM, BERT, Transformer and Ensemble models, where the best model results are Ensemble which is statistically much better than individual models with 89.3% accuracy. Other research [16] using Gradient Tree Boosting and Context features proved to provide insight into the classification of individuals diagnosed with anxiety with 97.3% accuracy.

Gradient boosting algorithm is a machine learning technique used for classification and regression problems. Gradient Boosting belongs to supervised learning which is based on decision tree [17], This algorithm optimizes the function space by selecting iteratively in the direction of the negative gradient, building an ensemble of weak decision trees through increments where the final result is obtained by adding the prediction results of all trees [18]. This approach learns a predictive model by combining M additive models (f0, f1, f2,...,fk) [19] in predicting the outcome.

From some of the previous studies mentioned, Gradient Boosting performs quite well. However, sequentially the Gradient Boosting algorithm works by adding previous predictors that do not match the prediction to the ensemble, simply correcting mistakes made previously so that it can potentially overfitting and too many outliers. Research [20] mentioned that the Gradient Boosting algorithm requires a lot of trees, high flexibility in generating many parameters which can consume time and memory space and is less interpretative. However, it is mentioned that this is easily overcome by using various tools. H20 can be an alternative tool that is open-source machine learning and uses in-memory compression [21] which is able to handle a lot of data in memory and provide fast processing time. H20 is also known as Automatic machine learning because it works by automating and optimizing the training and tuning of models, ensembles and stacking of multiple models so as to provide optimal performance for the model [22]. Research [23] conducted a review of 101 related literature and the result is that autoML is able to improve machine learning performance in a shorter time. Research [24] using autoML with gradient boosting and 8-Cross Validation and resulted in an accuracy of 90.09%. Based on the description above, research is proposed to produce a sentiment analysis model of anxiety (Anxiety) on social media using H2O Gradient Boosting by implementing TF-IDF Vectorization and Cross Validation.

METHODS

Systematically, this research takes several stages starting from the data collection stage, the data processing stage, the modeling stage and the evaluation stage. The initial stage begins with collecting data for research. then the dataset is further processed in the preprocessing stage so that the data is clean and ready for further processing. The stages used in this stage include lowercase or converting sentences into lowercase letters, then remove punctuation (removing punctuation marks) and tokenization (the process of breaking text into smaller word units) where this process helps in the assessment of each word unit so as to facilitate frequency-based analysis of word occurrence. The next stage is to explore the data using Wordcloud visualization as a stage in understanding the distribution of words in the dataset and can provide early insight into the general patterns in the data. Furthermore, the text vectorization stage uses TF-IDF as a stage in giving high weight to words that appear in certain documents. Then start the H2O environment as preprocessing by uploading and processing data through the selection of predictors and target labels, then converting them to categorical. Next, the data is divided into training data and testing data and the implementation of modeling with the gradient boosting model is carried out. Then after the model is built, the model is tested on testing data and then evaluates the performance of the model with 5-fold cross validation and displays the results of cross validation metrics such as accuracy, recall, precision, specificity, MSE. The Flowchart of the proposed model can be seen in Figure 1.

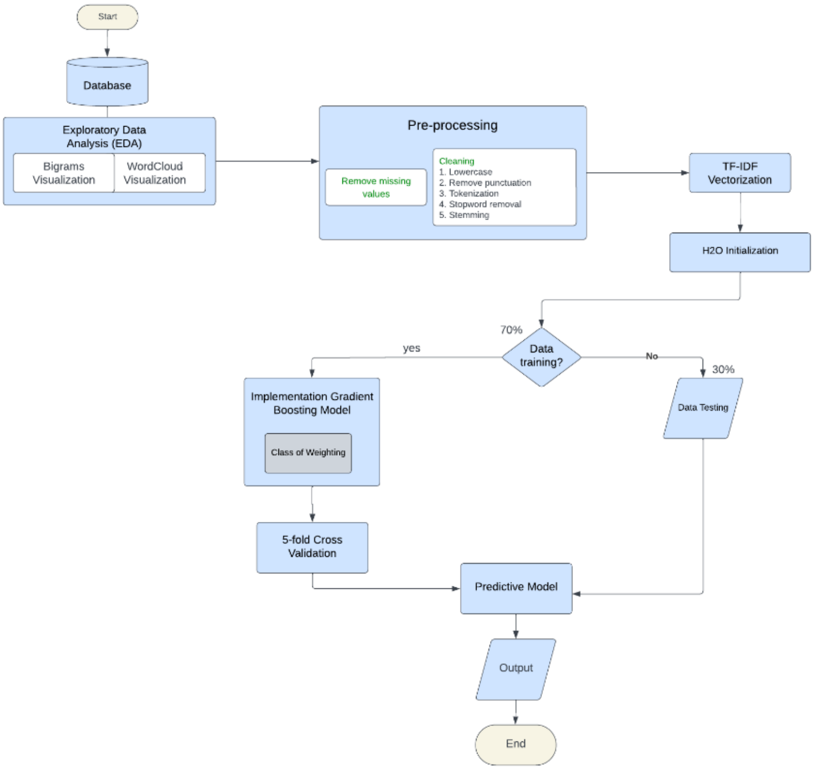

Figure 1. Research methodology

Data

In this study, the dataset was taken from the Kaggle Repository, namely the Students anxiety and depression dataset. The dataset can be accessed at the link https://www.kaggle.com/datasets/sahasourav17/students-anxiety-and-depression-dataset Students anxiety and depression dataset consisting of 6982 data with 2 columns, namely sentiment and text. Where the sentiment label consists of a label that shows anxiety with label 1 and normal with label 0. The dataset is taken from social media Facebook, Instagram, and other social media uploads regarding undergraduate student uploads.

Exploratory data analysis (EDA) & preprocessing data

At the Exploratory Data Analysis (EDA) stage, an exploration of the data is carried out to see and understand the data and patterns that exist in the data. This EDA approach also helps to see patterns that need to be resolved before further analysis. In the research conducted, preprocessing begins with the lowercase process, which changes the arrangement of words into lowercase letters. Then the remove punctuation process is a process to remove punctuation from the text and the process of removing unwanted characters so as not to affect the analysis. Then the process continues tokenization which is the process of separating sentences in the text into pieces of words or tokens that stand alone, where word splitting is based on spaces in the sentence that separate words. Next, apply stopword removal to eliminate irrelevant words by comparing them with existing stopwords, the stopwords used in the study are 'english' stopwords. Stopwords in short are a list of words that are not used in natural language processing due to lack of useful information. And finally, stemming performs the process of converting words into their basic form.

TF-IDF vectorization

The TF-IDF Vectorization stage is performed in the research in representing the importance of each word in the document. The TF-IDF process calculates and determines the frequency of words in a document by dividing the number of words by the total number of terms in the document. TF-IDF calculation can be seen as follows:

$$TF - IDF_{score\ for\ term\ i\ in\ document\ j} = TF(i,j) \times IDF(j) \quad (1)$$

$$TF(i,j) = \frac{Frequency\ of\ term\ i\ in\ document\ j}{total\ words\ in\ document\ j} \quad (2)$$

$$IDF(i) = log(\frac{Total\ documents}{number\ of\ documents\ with\ term\ i}) \quad (3)$$

From equations (1), (2) and (3) $TF(i,j)$ represents the frequency of occurrence of a word i in a particular document j. A high $TF(i,j)$indicates a greater significance of the word in the document. Then for $IDF(i)$ represents the number of documents where word i appears. A high IDF(i) frequency indicates that the word is found more often in several documents.

Class weighting

In overcoming the problem of class imbalance, suitable techniques are expressed [25] is using class weighting. In this method, the weight applied to each class is adjusted to the number of samples. Where classes that have small samples are given a greater weight, and classes with larger samples will be given a small weight. This is done in order to reduce classification errors and prevent the model from being biased towards the majority class [26], [27]. The imbalance of classes in the dataset certainly needs to be addressed with techniques used in balancing the data. In this case, class weighting is used to overcome the problem by weighting during the training process [28]. The weight applied to each class is adjusted to the number of samples. Where classes that have small samples are given greater weight, and classes with larger samples will be given a small weight. The calculation of the ratio comparing samples in classes can be seen in the following equation.

$$Rasio = \frac{number\ of\ minority\ classes\ (class\ 1)}{number\ of\ majority\ classes\ (class\ 0)} \quad (4)$$

5-fold cross validation

Cross Validation as a method is used to evaluate the predictive performance of the model. This evaluation will be used on the training data where the data will be divided into two parts, the training data while the predictive performance of the model is tested on the testing data. In the research, this method randomly divides the dataset with k-folds worth 5-fold separations of approximately the same size, and each fold in turn is used to test the model indicated from the other k-folds. At each iteration [29], This process is repeated for a set number of 5-fold values.

Modelling

In this research, we will use H2O AutoML's Gradient Booting model, but the process is not fully automated, because in this case H2O AutoML is used to call H2O's Gradient boosting model. Gradient boosting is part of the ensemble method. In machine learning ensemble classifiers improve results by combining multiple models (Dutta et al., 2020). Boosting means the process of improving predictive accuracy by applying the function continuously in a series and then combining the output of each function (Nassif, 2017). Data that has gone through a series of processes such as Exploratory Data Analysis, Preprocessing and text vectorization using TF-IDF Vectorization will be prepared to be loaded on the H2O library before starting the modeling analysis process. This process starts by initializing the H2O environment by activating it using the h2o.init() function. After the H2O environment is successfully initialized, feature and predictor selection is performed. The TF-IDF matrix results generated in the previous process will be converted into H2Oframe format using the h2o. H2Oframe() function so that it can be processed in the H2O ecosystem or environment for further processing.

Model evaluation

The testing and evaluation stage of the model is carried out with Confusion matrix and cross validation metrics which are table metrics to measure the performance and performance of the classification model. Confusion matrix represents the predicted and actual results of the applied machine learning model. From the Confusion Matrix method, it will be able to calculate classification reports such as accuracy. Precision, F1-Score and recall which can be calculated by Equation. Model evaluation is also done by looking at error rate, specificity, MSE and RMSE.

RESULTS AND DISCUSSIONS

Result exploratory data anlysis dan preprocessing

After exploration, the data contains missing values that must be checked again. After the missing values are removed, the number of data becomes 6972 data and there is no duplicate data. Then the labels in the data were analyzed for proportion as shown in Figure 2.

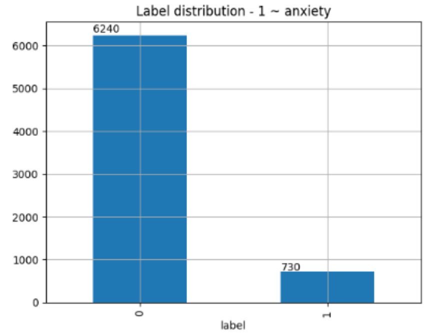

Figure 2. Label distribution

It can be seen from Figure 2. graph the number for each class 0 or normal is 6240 data and for class 1 or anxiety is 730 data. In this case it is said that the data has a class imbalance and needs further handling which will be handled in the modeling process by applying more weight to the minority class or called class of weighting. After that the data is also explored by visualizing the wordcloud or words that often appear in the data shown in Figure 3.

really
want
will
people
still
make
time
feel
come
know
one
today
morning
day
love

(a) (b)

Figure 3. (a) Wordcloud for class 0, (b) Wordcloud for class 1 (Anxiety)Wordcloud for class 1 (Anxiety) (normal)

Differences in the word representation of each class indicate explicit differences in language characteristics that are specific to the emotional context. The preprocessing result is the result of Students anxiety and depression dataset which has been processed through several stages. These stages consist of data cleaning by checking for missing values, then lower case, remove punctuation, tokenization, stopword removal, and stemming. Sample results of this process can be seen in Table 1.

Table 1. Number of attributes in the Dataset

Original Text	Results lower case *and* remove punctuation	Result *tokenization*, *stopword removal* and *stemming*
All wrong, back off dear, forward doubt. Stay in a restless and restless place	all wrong back off dear forward doubt stay in a restless and restless place	['wrong', 'back', 'dear', 'forward', 'doubt', 'stay', 'restless', 'restless', 'place']
No need to prove anything, thats enough for mee	no need to prove anything thats enough for mee	['need', 'prove', 'anyth', 'that', 'enough', 'mee']
In the morning the smell of praying	in the morning the smell of praying	['morn', 'smell', 'pray']

Thus, the overall text preprocessing process on the data becomes more focused on words that have important meanings for further analysis in extracting information from the data. This preprocessing process not only simplifies the data but also improves the quality of the data itself in order to produce analysis that is centered on the important elements of the data.

Result modelling

The application of TF-IDF Vectorization uses the TfidfVectorizer(max_features=maxfeat) function. This parameter allows researchers to limit the number of words to be used based on the highest TF-IDF value, thus helping to reduce the risk of overfitting and control the complexity of the model. In this case, the max_features=maxfeat parameter represents the determination of the maximum number of features or words considered. In this research, TF-IDF will be tested with max_features scenarios of 100, 250, 500, 1000 and 2000 to find out how the max_features parameter affects the H2O Gradient Boosting model. The test results can be seen in Table 2. And Figure 4.

Table 2. Number of attributes in the Dataset

Number of Features	Accuracy (%)
100	98,41
250	98,94
500	98,94
1000	98,94
2000	98,94

This indicates that the optimal number of features is around 250, so using a max_features value of more than that will not provide significant information but instead requires more computation so the modeling will continue with a TF-IDF max_features value of 250. The Cross Validation results in this experiment are shown in Table 3.

Table 3. Training results from cross validation metrics if max feature = 250

	H2O Gradient Boosting model performance when max feature TF-IDF =100	H2O Gradient Boosting model performance when max feature TF-IDF =250
accuracy	98,31%	98,94%
f1-score	92,0%	95,13%
mse	99,57%	99,57%
precision	0,0169	0,0106
recall	0,0168	0,0106
specificity	98,31%	98,94%

At this stage, the calculation of the model is carried out by displaying the confusion matrix and then calculating the value of other metrics needed to assess the performance of the model. Confusion Matrix results can be seen in Figure 4.

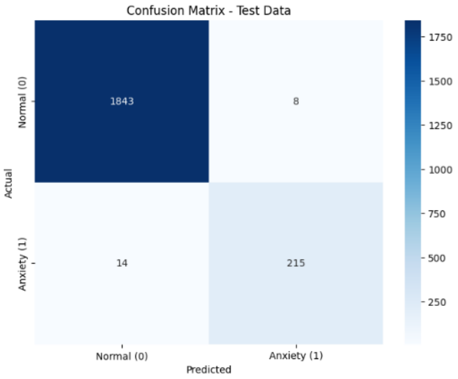

Figure 4. Result Confusion Matrix model with max feature TF-IDF=250

Then other metrics such as accuracy, precision, recall, F1-score, specificity, error rate, MSE, and RMSE are displayed to see the performance of the model when using max feature TF-IDF 100 and 250 which can be seen in Table 4 below.

Table 4. Model performance when using max feature TF-IDF 100 and 250

	mean	**sd**	**cv-1**	**cv-2**	**cv-3**	**cv-4**	**cv-5**
accuracy	0.989195	0.003948	0.988142	0.983968	0.988764	0.994918	0.990185
f1-score	0.947280	0.019117	0.939393	0.922330	0.945812	0.974093	0.954773
precision	0.950492	0.028872	0.93	0.931372	0.950495	1.0	0.940594
recall	0.944500	0.020241	0.948979	0.913461	0.941176	0.949495	0.969387
specificity	0.9942969	0.0032980	0.9923413	0.9921700	0.9942988	1.0	0.992674

In Table 4, it can be seen that the model performance tends to be better when using max feature TF-IDF =250. The accuracy increase is about 0.63 with a performance accuracy of 98.94%. Furthermore, a comparison with previous research is made to find out that this research is better than previous research which can be seen in Table 5.

Table 5. Comparison Result

Author	Method	Result
Zarate et al., (2023) [14]	*Naïve Bayes, Random Forests, LASSO-Regression*	Accuracy best model *Naïve Bayes* 81,1%
Dixit & College, (2023) [15]	*LSTM, BERT, Transformer, Ensemble*	Accuracy best model *Ensemble 89,3%*
A. Jain & Kumar, (2024) [16]	*Gradient Tree Boosting*	Accuracy *97,3%*
Varmann et al., (2024) [30]	*H2O Gradient Boosting*	MAE 0,38 MSE 0,32 RMSE 0,57
Y. V. Modha, (2021) [31]	*H20 Deep Learning Model + Word2vec-500 vector*	Accuracy 86.46%
Proposed Method	***H2O Gradient Boosting + TF-IDF (max feature=250)***	Accuracy **98,94%**

From Table 5. It can be seen that the proposed model has better performance than previous studies. So that the proposed model succeeds in providing good performance.

CONCLUSION

Based on the discussion and results of the research conducted previously, the initial stage begins with data collection for research. then the dataset is further processed by removing missing values, in the preprocessing stage (lowercase, remove punctuation, tokenixation, stopword removal and stemming) so that the data is clean and ready for further processing. The next stage is to explore the data using Wordcloud visualization as a stage in understanding the distribution of words in the dataset and can provide early insight into the general patterns in the data. Furthermore, the text vectorization stage uses TF-IDF as a stage in giving high weight to words that appear in certain documents. At the TF-IDF stage, experiments are carried out by comparing the weighting of max feature100, 250, 500, 1000, 2000. Then modeling is done using H2O Gradient Boosting by applying 5-fold cross validation on the training model. Then after the model is built, the model is tested on testing data and then evaluated the performance of the model where the results of applying TF-IDF by determining the max feature considered aim to focus on the data with the highest value. The best model performance is obtained when using max feature =250. The accuracy result in the experiment of max feature value TF-IDF =100 is 98.31% and the accuracy result with max feature value TF-IDF =250 is 98.94%. So from these results, the model is stable with the use of max feature TF-IDF around the number 250 to produce the best model performance with an accuracy of 98.94%, precision 96.41%, recall 93.88%, F1-score 95.13%, specificity 99.57%, and an error rate value of 0.0106.

REFERENCES

[1] B. Storer *et al.*, “The prevalence of anxiety in respiratory and sleep diseases : A systematic review and meta-analysis,” *Respir. Med.*, vol. 230, no. February, p. 107677, 2024, doi: 10.1016/j.rmed.2024.107677.

[2] B. W. Penninx, D. S. Pine, E. A. Holmes, and A. Reif, “Anxiety disorders,” *Lancet*, vol. 397, no. 10277, pp. 914–927, 2021, doi: https://doi.org/10.1016/S0140-6736(21)00359-7.

[3] Q. Jumrotul, Aqobah, and D. Rhamadian, “The Impact Of Anxiety In Sports On Athletes,” *J. Sport Sci. Tour. Act.*, vol. 1, no. 1, pp. 33–39, 2022, doi: http://dx.doi.org/10.52742/josita.v1i1.

[4] J. Wang *et al.*, “The moderating role of psychological resilience in the relationship between falls, anxiety and depressive symptoms,” 2023, doi: 10.1016/j.jad.2023.08.060.

[5] L. M. Prichett, R. H. Yolken, and E. G. Severance, “COVID-19 and Youth Mental Health Disparities : Intersectional Trends in Depression , Anxiety and Suicide Risk-Related Diagnoses,” *Acad. Pediatr.*, vol. 24, no. 5, pp. 837–847, 2024, doi: 10.1016/j.acap.2024.01.021.

[6] J. W. G. T. MD, FRACP, and FRANZCP, “Depression and anxiety,” *Med. J. Aust.*, no. October, pp. 1–4, 2013, doi: 10.5694/mja12.10628.

[7] H. Sueki, “The association of suicide-related Twitter use with suicidal behaviour : A cross-sectional study of young internet users in Japan,” *J. Affect. Disord.*, vol. 170, pp. 155–160, 2015, doi: 10.1016/j.jad.2014.08.047.

[8] H. Won *et al.*, “Predicting National Suicide Numbers with Social Media Data,” vol. 8, no. 4, pp. 1–6, 2013, doi: 10.1371/journal.pone.0061809.

[9] A. Maria, M. Herrera, J. Shadid, P. Zheng, and D. M. Pigott, “Articles Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic,” *Lancet*, vol. 398, no. 10312, 2021, doi: 10.1016/S0140-6736(21)02143-7.

[10] M. Deshpande, “Depression Detection using Emotion Artificial Intelligence,” *2017 Int. Conf. Intell. Sustain. Syst.*, no. Iciss, pp. 858–862, 2017.

[11] J. Zhu, Z. Li, X. Zhang, Z. Zhang, and B. Hu, “Public attitudes towards anxiety disorder on Sina

Weibo : content analysis Table of Contents," *J. Med. Internet Res.*, vol. 25, 2023, doi: 10.2196/45777.

[12] M. R. Ningsih, "Classification Email Spam using Naive Bayes Algorithm and Chi-Squared Feature Selection," vol. 9, no. 1, pp. 74–87, 2024.

[13] M. R. Ningsih, J. Unjung, D. Ananda, A. Pertiwi, B. Prasetiyo, and M. Aziz, "Optimized support vector machine with particle swarm optimization to improve the accuracy amazon sentiment analysis classification," *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control*, vol. 4, no. 1, 2024.

[14] D. Zarate, M. Ball, M. Prokofieva, V. Kostakos, and V. Stavropoulos, "Identifying self-disclosed anxiety on Twitter : A natural language processing approach," *Psychiatry Res.*, vol. 330, no. September, p. 115579, 2023, doi: 10.1016/j.psychres.2023.115579.

[15] K. K. Dixit and L. L. College, "Analyzing Textual Data for Mental Health Assessment : Natural Language Processing for Depression and Anxiety," *2023 10th IEEE Uttar Pradesh Sect. Int. Conf. Electr. Electron. Comput. Eng.*, vol. 10, pp. 1796–1802, 2023, doi: 10.1109/UPCON59197.2023.10434291.

[16] A. Jain and R. Kumar, "Machine Learning based Anxiety Detection using Physiological Signals and Context Features," *Int. Conf. Adv. Comput. Comput. Technol.*, 2024.

[17] S. E. Suryana, B. Warsito, and S. Suparti, "Penerapan Gradient Boosting Dengan Hyperopt Untuk Memprediksi Keberhasilan Telemarketing Bank," *J. Gaussian*, vol. 10, no. 4, pp. 617–623, 2021, doi: 10.14710/j.gauss.v10i4.31335.

[18] W. Li, W. Wang, and W. Huo, "RegBoost : a gradient boosted multivariate regression algorithm," *Int. J. Crowd Sci.*, vol. 4, no. 1, pp. 60–72, 2020, doi: 10.1108/IJCS-10-2019-0029.

[19] R. Fadhilah, S. D. Budiwati, D. R. Wijaya, P. R. Oktranida, Z. Q. Hijriana, and A. Firmansyah, "Comparison of Bandung Social Media-based Sentiment Classifier using Multinomial Logistic Regression and Gradient Boosting Models," *2023 Int. Conf. Data Sci. Its Appl. ICoDSA 2023*, pp. 83–87, 2023, doi: 10.1109/ICoDSA58501.2023.10276762.

[20] T. Z. Jasman, M. A. Fadhlullah, A. L. Pratama, and R. Rismayani, "Analisis Algoritma Gradient Boosting, Adaboost dan Catboost dalam Klasifikasi Kualitas Air," *J. Tek. Inform. dan Sist. Inf.*, vol. 8, no. 2, pp. 392–402, 2022, doi: 10.28932/jutisi.v8i2.4906.

[21] H2O.ai, "Starting H2O," 2024.

[22] A. Garg and A. Chaudhary, "Analysis of IPL Auction Dataset Using Explainable Machine Learning with Lime and H2O AutoML," *4th Int. Conf. Intell. Eng. Manag. ICIEM 2023*, no. Iciem, pp. 1–4, 2023, doi: 10.1109/ICIEM59379.2023.10167124.

[23] J. Waring, C. Lindvall, and R. Umeton, "Automated machine learning: Review of the state-of-the-art and opportunities for healthcare," *Artif. Intell. Med.*, vol. 104, no. October 2019, p. 101822, 2020, doi: 10.1016/j.artmed.2020.101822.

[24] V. Joshi, A. Ayushi, and N. Agarwal, "Forest Cover Type Prediction using Automatic Machine Learning," *2023 14th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2023*, pp. 1–5, 2023, doi: 10.1109/ICCCNT56998.2023.10307426.

[25] B. Bakirar and A. H. Elhan, "Class weighting technique to deal with imbalanced class problem in machine learning: methodological research," *Turkiye Klin. J. Biostat.*, vol. 15, no. 1, pp. 19–29, 2023, doi: 10.5336/biostatic.2022-93961.

[26] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, p. 27, Dec. 2019, doi: 10.1186/s40537-019-0192-5.

[27] J. He and M. X. Cheng, "Weighting methods for rare event identification from imbalanced datasets," *Front. Big Data*, vol. 4, no. December, pp. 1–11, 2021, doi: 10.3389/fdata.2021.715320.

[28] H. Akbar and W. K. Sanjaya, "Kajian performa metode class weight random forest pada klasifikasi imbalance data kelas curah hujan," *J. Sains, Nalar, dan Apl. Teknol. Inf.*, vol. 3, no. 1, 2023, doi: 10.20885/snati.v3i1.30.

[29] S. Yadav and S. Shukla, "Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification," *Proc. - 6th Int. Adv. Comput. Conf. IACC 2016*, no. Cv, pp. 78–83, 2016, doi: 10.1109/IACC.2016.25.

[30] S. S. M. Varmann, G. Hariprasath, and I. Kadirova, "Optimizing Educational Outcomes : H2O Gradient Boosting Algorithm in FMDB Transactions on Sustainable Techno Learning Optimizing Educational Outcomes : H2O Gradient Boosting Algorithm in Student Performance Prediction," *FMDB Trans. Sustain. Techno Learn.*, vol. 1, no. March, pp. 165 – 178, 2024.

[31] Y. V. Modha, "Machine Learning to aid mental health among youth during COVID-19 Yash Vijay Modha MSc Data Analytics," 2021.