The Asian Journal of Technology Management Vol. 16 No. 3 (2023): 211- 225

An Extreme Gradient Boosting Approach for Classification and Sentiment Analysis

Indah Yessi Kairupan[1], Apriandy Angdresey[*1], Hamdani Arif[2], and Kenshin Geraldy Emor[1]

[1]Department of Informatics Engineering, Universitas Katolik De La Salle Manado, Indonesia
[2]Department of Informatics Engineering, Politeknik Negeri Batam, Indonesia

Abstract. *Since 2020, when the coronavirus epidemic was at its peak, the Indonesian Ministry of Health's social media accounts have been constantly followed by a big number of individuals. The Indonesian Ministry of Health account is a fantastic resource for social media users, particularly Twitter users. The Republic of Indonesia's Ministry of Health's Twitter account publishes a wide range of content at random. As a result, it is usually difficult for Twitter users to determine the type of information provided by the Ministry of Health of the Republic of Indonesia. The positive and negative responses of Twitter users to material released by the Indonesian Ministry of Health's Twitter account are frequently noted. The decision tree algorithm is tree-based, similar to the extreme gradient boosting method (XGBoost). The extreme gradient boosting approach has been successfully implemented with high performance in the classification process. This classification is separated into two primary categories: general and essential information categorization and sentiment analysis, which is classified into three classes: positive, neutral, and negative. Both the classification work and the sentiment analysis produced outstanding accuracy levels. Based on 2243 tweets, an accuracy rate of 89.35% has been achieved for classification, supported by a precision of 88.76% and a recall value of 88.58% when using 80 data training and 20 data testing. Similarly, the maximum accuracy in sentiment analysis was achieved utilizing the same 80-20 data partitioning, with a 91.22% accuracy rate. Using 304 comments data, accuracy was calculated to be 89.17% and recall was calculated to be 89.06%. It's worth noting that an 80-20 split for training and testing consistently produced the best results for both the sentiment analysis and classification tasks.*

Keywords: *Twitter, classification, sentiment analysis, xgboost, republic of indonesia the ministry of health*

1. Introduction

The advancement of technology has enabled many tasks to be done digitally (Fatwa, 2020; Lavicza et al., 2022), leading to a surge in social media usage, particularly during the COVID-19 pandemic (Shim et al., 2021). Twitter, a prominent social network, is widely used in Indonesia, with the number of active users reaching 19.5 million, ranking sixth globally. Twitter serves as a barometer for trending topics, covering politics, religion, entertainment, scandals, and inspiring stories. The Ministry of Health of the Republic of Indonesia plays a vital role in the health sector, being accountable to the President (RI, 2021). It utilizes various platforms, including social media, to disseminate health-related information such as disease prevention, management, policies, and treatments. Given the prioritization of health issues, including communicable and non-communicable diseases, the ministry has successfully transitioned to digital platforms for information dissemination, contributing to a cultural shift towards healthier living (Tai-Seale et al., 2022). The use of social media offers advantages such as cost-effectiveness, wide reach, rapid dissemination, and opportunities for public engagement.

The official Twitter account of the Indonesian Ministry of Health features two main types of submissions, categorized as health-related information and news updates (Rathore et al.,

*Corresponding author. Email: aprhyliem@gmail.com
Received: September 12th, 2023; Revised: February 28th, 2024; Accepted: April 01st, 2024
Doi: http://dx.doi.org/10.12695/ajtm.2023.16.3.5
Print ISSN: 1978-6956; Online ISSN: 2089-791X.

2021). However, these uploads lack a discernable pattern, making it challenging for users to distinguish between the two. The content on the account covers a wide array of information presented in a haphazard manner, further complicating users' ability to differentiate between different categories. Additionally, user comments on the Ministry's tweets often reflect both supportive and opposing viewpoints, with analyzing each user's remarks being a time-consuming task due to the emotional positions expressed. Previous research (Darwis et al., 2020), on public attitudes towards the Corruption Eradication Commission (KPK) utilized Twitter data to gauge sentiments towards the KPK's performance, revealing prevalent public sentiments against the content posted by the KPK.

Effective data analysis is crucial for successful business operations (Chazal & Michel, 2021), aiding in informed decision-making and ultimately enhancing outcomes. Among various data analysis approaches, classification stands out, focusing on categorization using conventional statistical models and machine learning models. Extreme Gradient Boosting (XGBoost) is a notable technique within this domain, producing efficient solution trees for concurrent operation (Ichwanul Muslim Karo Karo, 2020). Extending this method, a study by (Cherif & Kortebi, 2019) achieves remarkable accuracy of 99.5% in traffic classifications using a supervised technique based on flow-based analysis.

In contrast, (Wongkar & Angdresey, 2019) utilize the Naive Bayes method to assess popular sentiment towards Indonesian presidential candidates, achieving a test result accuracy of 75.58%. However, the primary objective of this research is to utilize Extreme Gradient Boosting to identify tweet themes and analyze public responses to information from the Indonesian Ministry of Health. This approach streamlines sentiment analysis, benefiting both the Ministry by simplifying the study of public opinions and the public by facilitating easier access to classified material. However, our goal in this study is to utilize Extreme Gradient Boosting to identify tweet themes and analyze public responses to information from the Indonesian Ministry of Health. The practical application of the aforementioned method is a significant contribution of this research. The Extreme Gradient Boosting technique, when employed, allows for automatic sentiment analysis, saving time and effort. This not only benefits the Indonesian Ministry of Health by simplifying the study of public opinions but also improves public accessibility by enabling people to more readily examine classified material. Part of this research has been presented in (Angdresey et al., 2022).

The following is the structure of the article. Section II goes with scientific work relating to writing. Section III describes the approach and how it identifies tweet subjects, as well as sentiment analysis of user comments. Section IV also includes a performance review of the approach utilized. This evaluation is broken into two parts: exam preparation and the results and discussion. Section V concludes with conclusions and future work on this research.

2. Literature Review

Data mining is a scientific approach that integrates machine learning techniques, pattern recognition, statistics, and databases to overcome challenges in natural language processing. It achieves this by converting unstructured data such as text, images, sounds, and videos into a more structured format, thereby transforming it into valuable information (Musa et al., 2023). One specific type of data mining is text mining, which has been extensively studied in research contexts. Text mining involves the collection of textual data, often sourced from documents like papers. It then sifts through this data to identify terms that reflect the document's contents before analyzing the connections between various documents. This process helps extract meaningful insights from large volumes of textual information (Matrutty et al., 2023).

Sentiment analysis serves as a crucial tool for opinion mining, offering valuable insights for business intelligence. It involves evaluating digital text to discern the emotional tone of communications, whether positive, negative, or neutral. This analysis encompasses vast amounts of data sources, including emails, social media comments, and reviews, and it aims to classify the polarity of language within the content. In a study by (Li et al., 2020), the research focuses on thoroughly assessing sentiment polarity in Chinese texts. They employ a combined learning technique, integrating face-to-face and online learning with technology support. This approach aims to predict sentiment trends perceived by users and examine emotional expression in the learning process. Similarly, (Cheng & Tsai, 2019) present a methodology to assess public opinion from social media users' tweets. They utilize multiple algorithms to gauge the performance of government authorities in the health sector, particularly in addressing the Covid-19 issue. This research provides valuable insights into public sentiment regarding governmental responses to health crises.

Sentiment analysis research is frequently applied in evaluating government sector performance, as showcased by several studies. (Darwis et al., 2020) focus on the Republic of Indonesia's Corruption Eradication Commission, analyzing Twitter data using the support vector machine approach to convey public sentiment on KPK performance with an 82% accuracy rate. Similarly, (Awaludin, 2017) utilizes a multi-class support vector machine technique to assess customer happiness with airlines through Twitter data, achieving accuracy, precision, and recall rates of 80.41%, 84.33%, and 84.67% respectively. In a tourism-related context, (Wardani & Ruldeviyani, 2021) conduct research on the West Sumatera Province, utilizing sentiment analysis based on TripAdvisor hotel booking website data to assist hotel management in measuring client happiness. They employ the Naive Bayes method, achieving an average accuracy of 70.55% and a precision of 70.57% across five test examples. In another study, (Giovani et al., 2022) develop an application to gather feedback on organizational branding, incorporating the Naive Bayes technique. They attain varying results, with an accuracy of 52.66% using an 80:20 data partition and 77.78% accuracy with a 90:10 data split. Moreover, sentiment analysis is utilized to analyze student impressions of online lectures by collecting tweets from Twitter users and employing the Naive Bayes classifier (Damaratih, 2021). The authors achieve a 62% accuracy rate using different data partitions for training and testing.

XGBoost promotes efficient boosted tree generation through the use of gradient boosting decision trees that may run in parallel (Chen & Guestrin, 2016). The ensemble principle is used by XGBoost, a machine learning approach for regression and classification, to develop a strong model for robust predictions (Cherif & Kortebi, 2019). Attribute tests are represented by inner nodes in regression trees, while judgments are represented by leaf nodes with scores. This technique is widely used in a variety of research disciplines, producing cutting-edge results in machine learning tasks (Chen & Guestrin, 2016). In particular, (Luo et al., 2021) faced growing data loads where previous models fell short, recommending heightened XGBoost levels to properly anticipate short-term demands. Similarly, (Luo et al., 2021) used XGBoost to forecast probable purchasers for e-commerce apps, hence improving strategic performance and income. Machine learning techniques are extremely useful in a variety of sectors, including medicine. (Desdhanty & Rustam, 2021) used cancer data categorization to increase diagnosis accuracy by leveraging several mathematical characteristics, reaching 82% testing data accuracy.

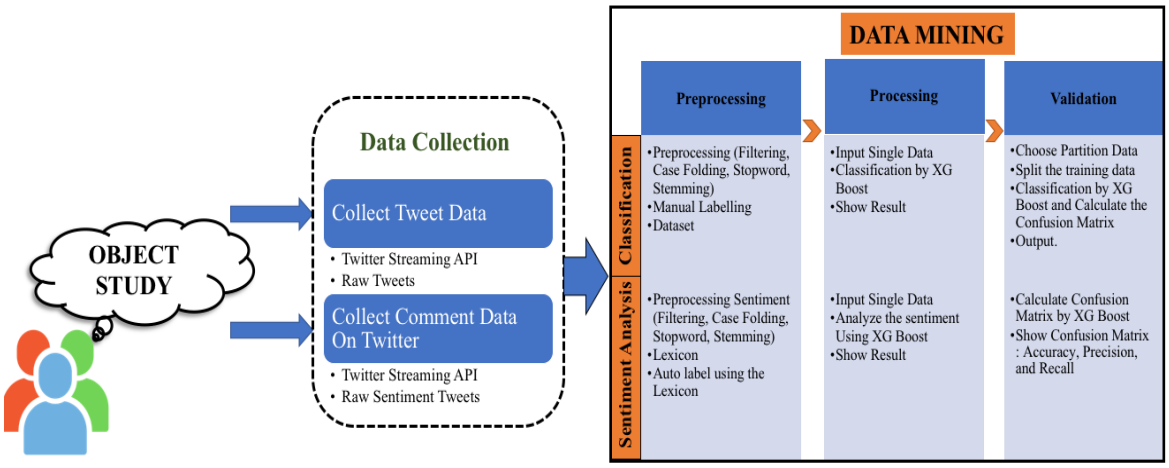


Figure 1.
System Framework

3. Methodology

As shown in Figure 1, this section highlights four critical components of the strategy used to solve current difficulties. To begin data gathering, we use a Twitter crawler to collect tweet data. Following that, we parse the obtained tweets verbatim. The preprocessing step is introduced in the second phase. This step is used to sanitize before it is processed. The initial stage of data processing is crucial as the raw data obtained at the outset is typically unstructured. This phase plays a critical role in transforming the data into a structured format, making it organized and usable for further analysis. The final component of this process involves text mining, wherein the extreme gradient boosting method is employed. Text mining enhances the extraction of meaningful insights from textual data, facilitating more informed decision-making and deeper understanding of the underlying patterns within the dataset.

A. Collected Data

To gather tweets and comments, we utilize web crawling techniques. The data collection comprises 2,243 rows of tweet data and 305 rows of comment data gathered between March and July 2022.

B. Preprocessing

After obtaining the raw data, a cleaning procedure will be carried out to eliminate any extraneous information. Any content that contains letters, symbols, punctuation marks, or numbers that are unrelated to the final conclusion will be removed. Furthermore, character consistency will be achieved in the text, since the system will convert uppercase letters to lowercase letters. The goal of this step is to make the analysis process easier while it is on the system. Following that, phrases or paragraphs will be broken down into single-word chunks. The pre-processing approach is described in depth in the following steps:

1) Cleaning; The basic cleaning concept of text can be seen in the example below: *"#RilisSehat Tetap semangat ya, agar nanti Lekas Pulih @KemenkesRI"*. After, cleaning the text will be *"Tetap semangat ya agar nanti Lekas Pulih"*.
2) Case Folding; The next stage will be a uniform character in the text. For example: *"Tetap semangat ya agar nanti Lekas Pulih"*, After the case folding is done, it becomes *"tetap semangat ya agar nanti lekas pulih"*.
3) Tokenization; In this process, sentences or paragraphs will be changed into single-word pieces. The following is an illustration of the tokenization process working: *"tetap semangat ya agar, nanti, lekas*

pulih", after tokenization becomes [*'tetap', 'semangat', 'ya', 'agar', 'nanti', 'lekas', pulih'*].

4) Stopword Removal; Removes connecting words from the text. Like this example: [*'tetap', 'semangat', 'ya', 'agar','nanti','lekas', 'pulih'*] becomes *"tetap semangat agar lekas pulih"*.
5) Stemming; The system will delete words that contain prefixes and suffixes. However, this process is not carried out because there are no words that contain the above reasons.

C. Processing

This section explains the system requirements definition as well as the method of operating the system or application that will be constructed. There are two types of system requirements specifications: functional and non-functional. We have completed two components in terms of functionality. To begin, issues inside tweets from the Indonesian Ministry of Health are classified. Second, do sentiment analysis on the same tweets.

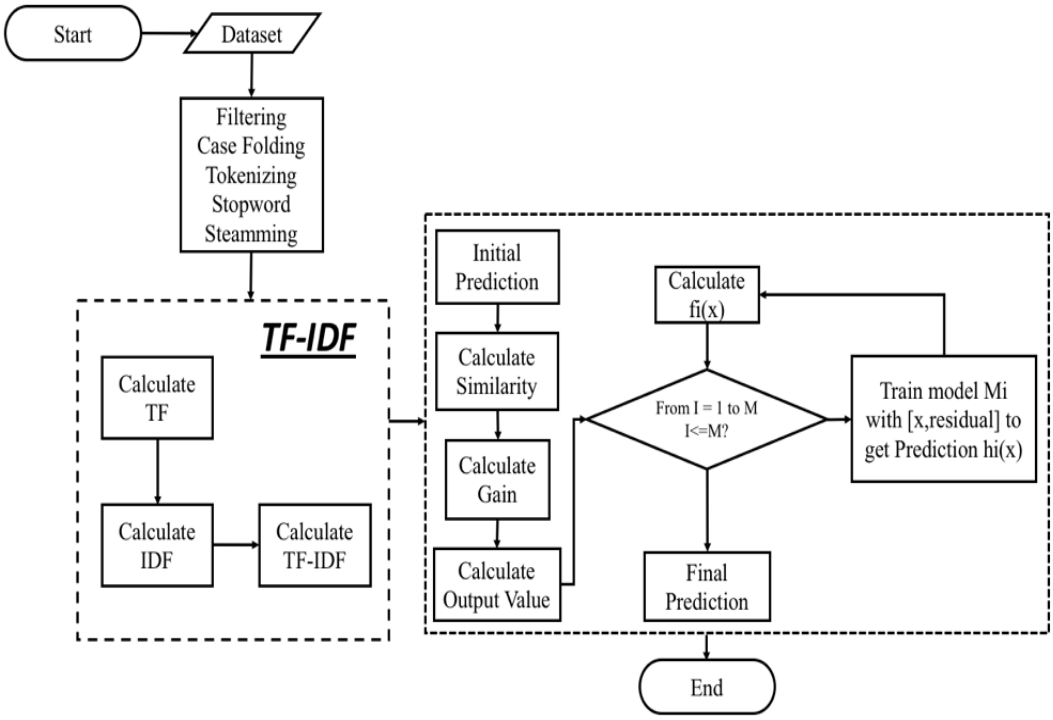


Figure 2.
Flowchart

TF-IDF

The system requirements specified result in an application with two key features: first, the categorization of tweet themes into two categories -- general and specialized information; and second, sentiment analysis of Ministry of Health tweets into positive, neutral, and negative attitudes. The Extreme Gradient Boosting approach was used in the creation of this application for sentiment analysis and categorization. An illustrated computation utilizing the extreme gradient boosting approach is offered to acquire a better grasp of how this methodology works. The first step is to compute the TF-IDF on the tweeted content.

Table 1.
An Example of Training Data

No	Comment	Label
1	*Tetap semangat agar lekas sembuh*	*Positif*
2	*Capek udah tidak harap keburu lemes minum yang ada aja*	*Negatif*
3	*Udah dapet wa tapi obat belum datang*	*Netral*
4	*Terima kasih kementerian kesehatan obat gratis pasien Isoman udah sampai*	*Positif*
5	*Wa tidak guna tidak bisa hubungi sama sekali wa respon tidak ada*	*Negatif*

Table 2 contains the training data for further processing. Table 2 shows the findings of the appearance of terms in the data. Every word in a Twitter document will be the words in a document. To find this Term Frequency (TF) value, we used the following formula:

$$\boldsymbol{tf_{ij} = \frac{f_i(i)}{\max f_d(j)}\ j \in d} \quad (1)$$

$$\boldsymbol{idf(t, D) = \log\left(\frac{N}{df(t)+1}\right)} \quad (2)$$

Table 2.
Data Occurrence

Term	D1	D2	D3	D4	D5
Tetap	1	0	0	0	0
semangat	1	0	0	0	0
Agar	1	0	0	0	0
lekas	1	0	0	0	0
pulih	1	0	0	0	0
capek	0	1	0	0	0
Udah	0	1	1	1	0
…	…	…	…	…	…
guna	0	0	0	0	1
sekali	0	0	0	0	1
respon	0	0	0	0	1

Table 3 is the *TF* value obtained based on the data in table 2. After obtaining the TF score, search for the Document Frequency (DF) value, which may then be used to determine the IDF value using Equation 2. The table below shows the IDF values.

Table 3.
Term Frequency (TF)

Term	D1	D2	D3	D4	D5
tetap	0,2	0	0	0	0
semangat	0,2	0	0	0	0
agar	0,2	0	0	0	0
lekas	0,2	0	0	0	0
pulih	0,2	0	0	0	0
capek	0	0,1	0	0	0
udah	0	0,1	0,143	0,1	0

Term	D1	D2	D3	D4	D5
tidak	0	0,1	0	0	0,25
harap	0	0,1	0	0	0
keburu	0	0,1	0	0	0
lemes	0	0,1	0	0	0
minum	0	0,1	0	0	0
yang	0	0,1	0	0	0
ada	0	0,1	0	0	0
aja	0	0,1	0	0	0
dapet	0	0	0,143	0	0
wa	0	0	0,143	0	0,167
tapi	0	0	0,143	0	0
obat	0	0	0,143	0,1	0
belum	0	0	0,143	0	0
datang	0	0	0,143	0	0
terima	0	0	0	0,1	0
kasih	0	0	0	0,1	0
Kementerian	0	0	0	0,1	0
kesehatan	0	0	0	0,1	0
gratis	0	0	0	0,1	0
pasien	0	0	0	0,1	0
isolasi	0	0	0	0,1	0
mandiri	0	0	0	0,1	0
sampai	0	0	0	0,1	0
bisa	0	0	0	0	0,083
hubungi	0	0	0	0	0,083
sama	0	0	0	0	0,083
sekali	0	0	0	0	0,083
respon	0	0	0	0	0,083
guna	0	0	0	0	0,083

Table 4.
Inverse Document Frequency (IDF)

Term	Nilai DF	Nilai IDF
Tetap	1	0,699
semangat	1	0,699
Agar	1	0,699
Lekas	1	0,699
Pulih	1	0,699
Capek	1	0,699
Sudah	3	0,222
Tidak	3	0,222
harap	1	0,699
keburu	1	0,699
lemes	1	0,699
minum	1	0,699
yang	1	0,699
Ada	1	0,699
Aja	1	0,699

Term	Nilai DF	Nilai IDF
dapet	1	0,699
Wa	2	0,398
tapi	1	0,699
obat	3	0,222
belum	1	0,699
datang	1	0,699
terima	1	0,699
kasih	1	0,699
Kementerian	2	0,398
kesehatan	2	0,398
gratis	1	0,699
pasien	1	0,699
isolasi	1	0,699
mandiri	1	0,699
sampai	1	0,699
bisa	1	0,699
hubungi	1	0,699
sama	1	0,699
sekali	1	0,699
respon	1	0,699
guna	1	0,699

After obtaining the value of TF and IDF, we can determine the TF-IDF values by multiplying the TF values and IDF values. The results can be seen in Table 5. Based on the TF-IDF calculations, the first document in Table 1 has a value of 0.69 with a manual label of 1, followed by a second document with a value of 0.6 with a manual label of -1, a third document with a value of 0.51 with a manual label of 0, a fourth document with a value of 0.61 with a manual label of 1, and the last document with a value of 0.47 with a manual label of -1. The second step is to put the XGBoost approach into action. The framework of the algorithm we utilize is described in detail below.

Table 5.
TF-IDF

Term	D1	D2	D3	D4	D5
tetap	0,14	0	0	0	0
semangat	0,14	0	0	0	0
agar	0,14	0	0	0	0
lekas	0,14	0	0	0	0
pulih	0,14	0	0	0	0
capek	0	0,07	0	0	0
udah	0	0,022	0,032	0,022	0
tidak	0	0,022	0	0	0,055
harap	0	0,07	0	0	0
keburu	0	0,07	0	0	0
lemes	0	0,07	0	0	0
minum	0	0,07	0	0	0
yang	0	0,07	0	0	0
ada	0	0,07	0	0	0

Term	D1	D2	D3	D4	D5
aja	0	0,07	0	0	0
dapet	0	0	0,099	0	0
wa	0	0	0,057	0	0,066
tapi	0	0	0,099	0	0
obat	0	0	0,032	0,022	0
belum	0	0	0,099	0	0
datang	0	0	0,099	0	0
terima	0	0	0	0,07	0
kasih	0	0	0	0,07	0
Kementerian	0	0	0	0,039	0
kesehatan	0	0	0	0,039	0
gratis	0	0	0	0,07	0
pasien	0	0	0	0,07	0
isolasi	0	0	0	0,07	0
mandiri	0	0	0	0,07	0
sampai	0	0	0	0,07	0
bisa	0	0	0	0	0,058
hubungi	0	0	0	0	0,058
sama	0	0	0	0	0,058
sekali	0	0	0	0	0,058
respon	0	0	0	0	0,058
guna	0	0	0	0	0,058
dari	0	0	0	0	0
sesuai	0	0	0	0	0
kebutuhan	0	0	0	0	0
Total	***0,69***	***0,60***	***0,52***	***0,61***	***0,47***

Extreme Gradient Boosting (XGBoost)

The Extreme gradient boosting approach is used by taking the value received from the TF-IDF and utilizing it as the x value. XGBoost, like the Decision Tree method, is a tree-based method that employs the ensemble principle, which means that it combines numerous inadequate learning sets to construct a brand-new design that is both powerful and create powerful forecasts (Husada & Paramita, 2021). The first step after entering the dataset is to generate an initial model using the original data. The initial predicted value and residual score will then be calculated from the initial model using Eq. 3.

$$\widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{h_0(X)} \quad (3)$$

The second model will be constructed using the first residual error and will generate second model predictions. The third model will be carried out in the same manner until the iteration process matches the *n_estimator* numbers provided. The XGBoost tree is built differently from a traditional decision tree. Some of the formulas utilized before building the XGBoost tree illustrated in Eq. 4 are shown below.

$$Gain = Left_{Similarity} + Right_{Similarity} - Root_{Similarity} \quad (4)$$

$$\boldsymbol{SC} = \frac{(\sum \boldsymbol{Residual_i})^2}{\sum [\boldsymbol{Previous\ Probability_i\ x\ (1-Previous\ Probability_i)+\lambda}} \quad (5)$$

While the Eq. 5 below is the formula used to calculate the *similarity score* (SC), the Eq. 6 is the formula used to calculate the *output value* (OV). In addition, the calculation *i* is calculated till it equals *n_estimator*. If *i* is less than *n_estimator*, model *i* will be turned off again until it is equal to *n_estimator*. The Sigmoid equation used to train the model is shown in Eq. 7.

$$OV = \frac{(\sum Residual_i)}{\sum [Previous\ Probability_i\ x\ (1-Previous\ Probability_i]+\lambda} \quad (6)$$

$$Sigmoid = \frac{1}{(1 + Exp\left((-h_0(x)) + (-h_i(x))\right)} \quad (7)$$

Table 6.
Initial Prediction of Data Training

X	Y	$F_0(x)$	$Y = Y - F_0(x)$
0.69	1	0.5	0.5
0.60	-1	0.5	-1.5
0.51	0	0.5	-0.5
0.61	1	0.5	0.5
0.47	-1	0.5	-1.5

The initial prediction or *base_score* is set to *0.5* for all data points, namely $F_0(x)=0.5$. The first training model, which corresponds to the construction tree, is then generated. The value of the *n_estimator* in the development model that follows the *n_estimator* obtained in this calculation example is 2. The trees are constructed by partitioning the data into two halves *0.69*.

Table 7.
Calculation of n_estimator =1

X	Y	$F_0(x)$	$Y = Y - F_0(x)$
0,60	-1	0,5	-1,5
0,51	0	0,5	-0,5
0,61	1	0,5	0,5
0,47	-1	0,5	-1,5

Then, the Eq. 4 is applied to determine the similarities and benefits of the trees that have been constructed. The formula's output is shown on the following page.

Table 8.
Value Similarity and Gain Part 1

X	L,R	Similarity	Gain
0,645	[-1,5.-0,5.0,5.-1,5]	9	10
	[0,5]	1	
0,555	[-0,5.-1,5]	8	8,3333333
	[0,5.-1,5.0,5]	0,333333	
0,56	[-0,5.-1,5]	8	8,3333333
	[0,5.-1,5.0,5]	0,333333	
0,54	[-0,5.-1,5]	8	8,3333333

After determining the gain of each tree, the tree with the highest gain will be chosen, $X < 64$ being the tree with the highest gain of 10. The trees $X < 64$ will then be separated further until they match the *max_depth* value, completing the tree construction. We use Eq. 5 to obtain the output value.

Table 9.
Calculation of n_estimator = 2

X	Y	$F_0(x)$	$Y = Y - F_0(x)$
0.60	-1	0.5	-1.5
0.51	0	0.5	-0.5
0.61	1	0.5	0.5
0.47	-1	0.5	-1.5

Table 10.
Value Similarity and Gain Part 2

X	L,R	Similarity	Gain
0,555	[-0,5.-1,5] [-1,5.0,5]	8 2	10
0,56	[-0,5.-1,5] [-1,5.0,5]	8 2	10
0,54	[-0,5.-1,5] [-1,5.0,5]	8 2	10

Following the determination of the *n_estimator*, Eq. 4 and Eq. 5 are used to compute similarity and gain. Table 10 illustrates the results of the preceding computations. Because some leaves contain more than one residue, $X<56$ with a gain of 10 will calculate the output values for all leaves to obtain the final tree in model -1. Eq. 6 shows the computation of the output value. Now, run all of the data points through the Final Tree (Model-1) to obtain $h_1(x)$, and then calculate $F_1(x)$ and the residual. Table 11 displays the findings of $F_1(x)$ and residues.

Table 11.
Prediction Results $F_1(x)$

X	Y	$h_1(x)$	$F_1(x)$	$Y = Y - F_1(x)$
0,69	1	2	0,88	0,12
0,60	-1	-2	0,12	-1,12
0,51	0	-4	0,02	-0,02
0,61	1	-2	0,12	0,88
0,47	-1	-4	0,02	-1,02

The following is the outcome of model-2's output value. To obtain $h_2(x)$, run all of the data points through the Final Tree (Model-2) and then calculate the value of $F_2(x)$ and the residual. Table 12 displays the findings of $F_2(x)$ and residues.

Table 12.
Prediction Results $F_2(x)$

X	Y	$H_2(x)$	$F_2(x)$	$Y = Y - F_2(x)$
0,69	1	-0,37	0,84	0,16
0,60	-1	-0,37	0,09	-1,09
0,51	0	-1,02	0,01	-0,01
0,61	1	-0,37	0,09	0,91
0,47	-1	-52,04	0	-1

4. Findings and Discussion

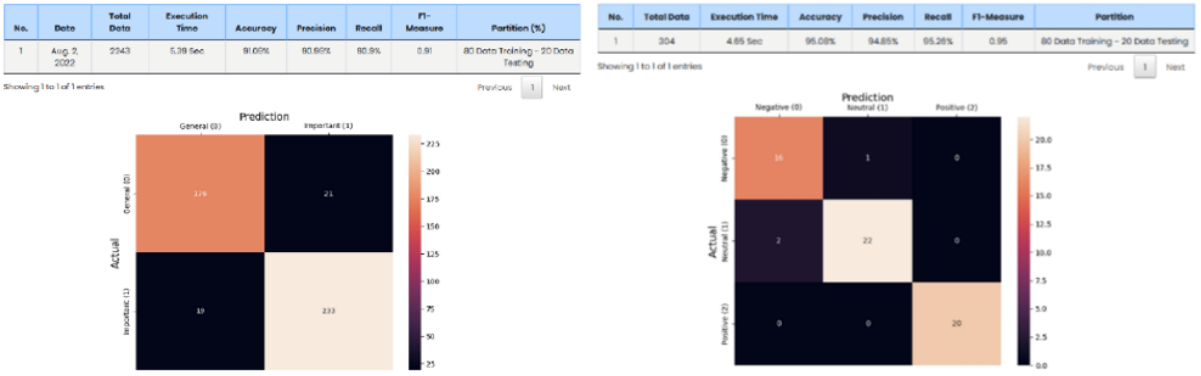
Performance Evaluation
This section demonstrates how the extreme gradient boosting method was used to emulate the classification and sentiment analysis applications created for this research. The application simulation is given below: The Twitter API is first utilized to obtain post data, which is then saved in the database. Second, after initially storing the posting data in the database, it is manually retrieved, cleaned, and tagged for each post before being saved again in the database. Third, do a single data test in which you use the extreme gradient boosting approach to predict the types of single-category data based on the data currently in the database. Fourth, choose which division will be tested twice. Posting data into the database is partitioned based on the incoming partition. After executing the computation and showing the results, the extreme gradient boosting process is used for testing. Sixth, the saved comment data is retrieved, cleaned, and automatically tagged using the vocabulary given in the database. New copies of labeled data have been added to the database.

Additionally, employ the extreme gradient boosting approach to test only the data and predict the sentiment of each tweet by studying the database's comment data. As a result, do a second test using XGBoosting algorithm. Following the execution, a graph of the confusion matrix and the confusion matrix findings for each division are displayed. The sentiment analysis validation page is seen in Figure 3. When the user clicks the process button on this page, the system processes three data divisions and displays their accuracy together with the associated pictures. In addition, if the user clicks the show button, the system will display a popup confusion matrix for each data point, as seen in Figure 4.

Figure 3.
Validation of Sentiment Analysis

Testing

Testing is done with the goal of verifying that the functionality and all elements of the program work properly, just like in the predefined design phases. The test criteria used in testing the application serve as a standard for determining the overall success of the program's functioning.



(a) Classification

(b) Sentiment Analysis

Figure 4.
Confusion Matrix

In Figure 5 demonstrates a multiple-data test in which the data was separated based on clean and labeled Twitter data and then processed using XGBoosting approach. The outcomes of the double data test with a total of 2243 Twitter data using extreme gradient boosting can run well and produce high accuracy values, namely an accuracy value of 91.09% with an execution time of 5.88 seconds for 80 training data and 20 test data; an accuracy value of 89.9% with an execution time of 5.0 seconds for testing 70 training data and 30 test data; and an accuracy value of 89.9% with an execution time of 5.0 seconds for testing 60 training data and 40 test data.

6. Conclusions

The extraction of data from Twitter can take a long time up until the stage of cleaning the raw Twitter data into clean Twitter data that is ready to be processed. Automatic sentiment analysis labeling works effectively by labeling comment data with a database dictionary. The extreme gradient boosting approach has also been successfully implemented and tested with single and multiple data sets. The prediction of single data in the classification process and sentiment analysis can be well forecasted using XGBoost. Get accurate and good findings via double data testing. In the classification process, with a total of 2243 posting data, the partition of 80 training data and 20 test data obtains an accuracy value of 91.09% with an execution time of 5.88 seconds; after that, in the test, 70 training data and 30 test data obtain an accuracy value of 89.9% with an execution time of 5.0 seconds; and in the test, 60 training data and 4 test data obtain an accuracy value of 89.9% with an execution time of 5.0 seconds.

Meanwhile, using three categories of sentiment, namely negative, neutral, and positive sentiment, and a total of 304 commentary data, the sentiment analysis process yielded the following results: in testing 80 training data and 20 test data, an accuracy value of 95.08% was obtained; The 70:30 partition data yields an accuracy result of 82.61% in the test case. Meanwhile, an accuracy rating of 77.05% was attained in the 60:40 test. The difficulty in testing multiple data points on sentiment analysis is that the process of testing multiple data points takes 88.90 seconds to execute due to the testing process based on partitions is executed one by one, and the more data that is processed, the longer the execution time is required.

For future works, classify photographs and videos uploaded by the Ministry of Health, Republic of Indonesia, and classify content from additional social media platforms, such as Facebook, Instagram, TikTok, and others.

References

Angdresey, A., Kairupan, I. Y., & Emor, K. G. (2022). Classification and Sentiment Analysis on Tweets of the Ministry of Health Republic of Indonesia. *2022 Seventh International Conference on Informatics and Computing (ICIC)*, 1–6.

Awaludin, A. A. R. (2017). Akreditasi Sekolah sebagai Suatu Upaya Penjaminan Mutu Pendidikan di Indonesia. *SAP (Susunan Artikel Pendidikan)*, *2*(1).

Chazal, F., & Michel, B. (2021). An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. In *Frontiers in Artificial Intelligence* (Vol. 4).

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-August-2016*.

Cheng, L. C., & Tsai, S. L. (2019). Deep learning for automated sentiment analysis of social media. *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019*.

Cherif, I. L., & Kortebi, A. (2019). On using eXtreme Gradient Boosting (XGBoost) Machine Learning algorithm for Home Network Traffic Classification. *IFIP Wireless Days*, *2019-April*.

Damaratih, D. A. (2021). Sentiment Analysis of Online Lecture Opinions on Twitter Social Media Using Naive Bayes Classifier. *2021 International Conference on Computer Science, Information Technology, and Electrical Engineering, ICOMITEE 2021*.

Darwis, D., Pratiwi, E. S., & Pasaribu, A. F. O. (2020). Penerapan Algoritma SVM Untuk Analisis Sentimen Pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia. *Edutic - Scientific Journal of Informatics Education*, *7*(1).

Desdhanty, V. S., & Rustam, Z. (2021). Liver Cancer Classification Using Random Forest and Extreme Gradient Boosting (XGBoost) with Genetic Algorithm as Feature Selection. *2021 International Conference on Decision Aid Sciences and Application, DASA 2021*.

Fatwa, A. (2020). Pemanfaatan Teknologi Pendidikan di Era New Normal. *Indonesian Journal of Instructional Technology*, *1*(2), 20–31.

Giovani, A. P., Ardiansyah, A., Haryanti, T., Kurniawati, L., & Gata, W. (2022). Implementasi Metode Multinomial Naïve Bayes Untuk Sentiment Analysis Terhadap Data Ulasan Produk Colearn Pada Google Play Store. *Prosiding Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)*, *1*(1).

Husada, H. C., & Paramita, A. S. (2021). Analisis Sentimen Pada Maskapai Penerbangan di Platform Twitter Menggunakan Algoritma Support Vector Machine (SVM). *Teknika*, *10*(1).

Ichwanul Muslim Karo Karo. (2020). Implementasi Metode XGBoost dan Feature Importance untuk Klasifikasi pada Kebakaran Hutan dan Lahan. *Journal of Software Engineering, Information and Communication Technology*, *1*(1).

Lavicza, Z., Weinhandl, R., Prodromou, T., Anđić, B., Lieban, D., Hohenwarter, M., Fenyvesi, K., Brownell, C., & Diego-Mantecón, J. M. (2022). Developing and Evaluating Educational Innovations for STEAM Education in Rapidly Changing Digital Technology Environments. *Sustainability (Switzerland)*, *14*(12).

Li, G., Zheng, Q. S., Zhang, L., Guo, S. Z., & Niu, L. Y. (2020). Sentiment Infomation based Model for Chinese text Sentiment Analysis. *2020 IEEE 3rd International Conference on Automation, Electronics and Electrical Engineering, AUTEEE 2020*.

Luo, S., Zhang, S., & Cong, H. (2021). Research on Consumer Purchasing Prediction Based on XGBoost Algorithm. *2021 IEEE International Conference on Artificial Intelligence and*

Computer Applications, ICAICA 2021.

Matrutty, J. P., Adrian, A. M., & Angdresey, A. (2023). Sentiment Analysis of Visitor Reviews on Star Hotels in Manado City. *Journal of Information Technology and Computer Science*, *8*(1).

Musa, U., Adebiyi, M. O., Adebiyi, A. A., & Adebiyi, A. A. (2023). Development of a Machine Learning Model For Big Data Analytics. *2023 International Conference on Science, Engineering and Business for Sustainable Development Goals (SEB-SDG)*, *1*, 1–6.

Rathore, A. K., Maurya, D., & Srivastava, A. K. (2021). Do policymakers use social media for policy design? A Twitter Analytics Approach. *Australasian Journal of Information Systems*, *25*.

RI, K. (2021). PMK No 10 Tahun 2021 Tentang Pelaksanaan Vaksinasi dalam Rangka Penanggulangan Pandemi Corona Virus Disease 2019 (COVID-19). *Permenkes RI*, *2019*.

Shim, J. G., Ryu, K. H., Lee, S. H., Cho, E. A., Lee, Y. J., & Ahn, J. H. (2021). Text mining approaches to analyze public sentiment changes regarding covid-19 vaccines on social media in korea. *International Journal of Environmental Research and Public Health*, *18*(12).

Tai-Seale, M., May, N., Sitapati, A., & Longhurst, C. A. (2022). A learning health system approach to COVID-19 exposure notification system rollout. *Learning Health Systems*, *6*(2).

Wardani, S. K., & Ruldeviyani, Y. (2021). Sentiment Analysis of Visitor Reviews on Hotel in West Sumatera. *Proceedings - IWBIS 2021: 6th International Workshop on Big Data and Information Security*.

Wongkar, M., & Angdresey, A. (2019). Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter. *Proceedings of 2019 4th International Conference on Informatics and Computing, ICIC 2019*.