An Investigation into Indonesian Students' Opinions on Educational Reforms through the Use of Machine Learning and Sentiment Analysis

Sarmini a, Abdullah Alhabeeb b, Majed Mohammed Abusharhah c, Taqwa Hariguna a,*, An Dhika Rafi Hananto a

a Department of Information Systems, Universitas Amikom Purwokerto, Purwokerto, 53127, Indonesia
b Department of Information Science, College of Arts, King Saud University, Riyadh, 11451, Saudi Arabia
c Ministry of Education, Saudi Arabia

Corresponding author: *taqwa@amikompurwokerto.ac.id

Abstract — An anti-Covid-19 plan with social restrictions forced all Indonesian educational institutions to implement online learning in 2020. Strategy in early 2022, a new policy brought back online learning methods. Because of the rapid change and short adaptation period, online learning, which had been accepted as a solution for approximately two years, has become controversial. There were a variety of reactions in society, particularly on social media, after the rapid shift from face-to-face learning to online learning. This study will quantify text sentiment expressed on social media through machine learning. This study used SVM, RF, DT, LR, and k-nearest neighbors to develop a sentiment analysis model for use in sentiment research (KNN). The SVM- and RF-based sentiment analysis models outperform the others in cross-validation tests using data from the same Twitter social media site. Furthermore, RF can classify public opinion into three groups: positive, negative, and neutral, with a low error rate. The f1 values of our KNN-based model were measured at 75%, 65%, and 87% for negative, neutral, and positive tweets, respectively, which are slightly more accurate than previous studies with the same method and purpose.

Keywords — Machine learning; sentiment analysis; public opinions; E-Learning.

I. INTRODUCTION

The COVID-19 virus was declared a global pandemic by the WHO on March 11, 2020 [1]. 1.2 million people died in the second week of November 2020 as a result of Covid-19, according to the World Health Organization (WHO). With the death toll reaching 15,148, Indonesia recorded 463,000 confirmed positives [2]. Since the virus spreads so quickly through physical contact, all countries are forced to implement social and physical distance measures to cut down on human contact [3]. When a vaccine is unavailable, the most widely used strategy is to implement social restrictions because the main method of transmission is through droplets released during coughing or sneezing.

Large-Scale Social Restrictions have been implemented by the Indonesian government [4]. Restricted access to public spaces like workplaces, classrooms, and campuses can affect [5]. As a result, all educational institutions are forced to cease face-to-face teaching and replace it with online education. The unpreparedness of technology, media, or students' psychological factors is all problem it causes in education. In the meantime, education will continue to be distributed unequally across Indonesia's communities and regions in the long term [6].

All educational institutions were forced to implement online learning to combat the pandemic [4], [5]. In response to the pandemic, Indonesia's Ministry of Education and Culture issued the Learning from Home policy [7], emphasizing online education's importance. This policy must use smartphones, gadgets, computers, and applications instead of face-to-face communication.

Schools that used to teach face-to-face are now forced to adapt to the online learning model. Online learning is a viable option for students as an alternative method of education that fosters self-reliance and fosters social interaction. Lecturers have more opportunities to assess and evaluate each student's learning program in an online environment [8]. In online learning, resources like documents, images, videos, and audio are combined to help students learn. Students make extensive use of the educational resources, which serve as a major
source of funding for the growth of online education. You have an obligation to make the learning experience as appealing as possible to achieve educational goals in online education courses.

An online learning strategy once considered a viable option became controversial in 2022 when the learning model began to shift back to face-to-face instruction. There has been a lot of discussion about this on Twitter. Society as a whole and students react differently to a shift from online learning to face-to-face learning after approximately two years, leading to various reactions. This study examines public opinion on online learning in Indonesia during the COVID-19 pandemic in early April 2022. There are five machine learning algorithms used in this study, including support vector machines (SVM), random forests (RF), decision trees (DT), logistic regression (LR), and K-nearest neighbors (K-NN), to classify tweets in Indonesian using the keywords' pembelajaran daring', 'kuliah', 'belajar', 'online', 'daring' and the hashtags #BelajarDariRumah, #KuliahOnline, and #BackToSchool.

Public opinion on their involvement is represented by their views on changes in learning methods on social media, which is the focus of this study. They were evaluated and classified using five machine learning algorithms: SVM, Random Forest, Decision Tree, Logistic Regression, and K-Nearest Neighbor, all of which were used in this study (KNN). In this study, we used Rumelli et al. [9] research as a benchmark for results and comparisons because it employs a similar method. Concrete concepts and guidelines in this research can help the government better understand how people feel about the value of good education in the age of social media as a communication tool.

A. Sentiment Analysis

It is possible to classify text polarity in a document or sentence using sentiment analysis to determine whether the sentiment is positive, negative, or neutral [9]. The use of sentiment analysis in computer science research is currently widespread. Sentiment analysis is a popular method for determining public opinion using social networks like Twitter. The focus on positive or negative opinions in sentiment analysis makes it a good fit for the term "opinion mining" [10]. An entity, such as a service or product or a person or phenomenon, can be the subject of sentiment analysis through data mining. The preprocessing of data includes tokenization, stopword, deletion, stemming, sentiment identification, and sentiment classification processes [11,12].

B. Text Mining

Text mining is the process of extracting data from a large collection of documents. Text Mining can generate useful information by processing, grouping, and analyzing large amounts of unstructured data [13]. Text mining is a technique used to extract useful information from documents with unstructured data sources in the text. Text mining can produce a feeling analysis that identifies whether a statement is positive or negative based on its retrieval process [14,15]. Unstructured or semi-structured documents can be mined as text objects [16]. As a result of text mining, relevant information can be extracted from many documents. Text mining transforms unstructured text into structured data, which is then saved in a structured database (see Fig. 1).

Fig. 1 Mining Text Model.

II. MATERIAL AND METHOD

A. Classification Method

According to the data set used to classify reviews into "positive," "negative," and "neutral," five machine learning classifiers were applied.

SVM [6] has been used to solve classification problems in a variety of ways. Classifying data by hyperplanes is the goal of this algorithm. A hyperplane is a straight line that maximizes the difference between two classes in two-dimensional space [17], [19]. SVM is based on finding line separators in the search space that can be used to identify different groups. The cost (C), epsilon (ε), and gamma (γ) parameters, as well as the type of kernel function, go into the mathematical formulation of the SVM method. In order to improve SVM performance, we use a grid search. Thus, the best values for the parameters "C" and the linear kernel function are obtained as 4, "ε" as 0.001, and "γ" as scale.

The ensemble algorithm class includes RF [17]. There are decision trees in the forest section, and bootstrap datasets are created using a random sample from the input data in the random section. Only two-thirds of the input data is included in the bootstrap dataset. In some cases, this data is repeated, and in others, it is not. Tests can be run on datasets that have been excluded from the study. In order to achieve the best model performance, the RF classifier makes use of multiple criteria.

DT [19] uses a tree-like decision model, as the name implies. It is possible to perform DT implementations without data scaling and to perform variable filtering or feature selection implicitly. In classification problems, DT can suffer from overfitting, which can lead to poor accuracy [17], [20]. The Gini index and Information Reinforcement are used to determine how the data should be divided when classifying DT data. It is part of the linear classifier group, including polynomial and linear regression. LR [13] LR is fast and simple to use, and the results are easy to understand. It is primarily a binary classification method, but it can also be used to solve multiclass problems.

Finally, the classification and regression problems are solved by KNN [17,18]. This shows that the input sample data sets have a high degree of similarity because the values are so close together. For the KNN classifier to work optimally, the number of neighbors in the prediction must be selected carefully [21]. The best KNN model can be found by experimenting with neighbor parameters ranging from 1 to 30 when using the elbow method.

Three different metrics have been used to assess the model's performance: precision, recall, and the f1-score. (1), (2), and (3) are the equations for precision, recall, and f1-score, respectively.
True positives, false positives, and false negatives are estimated in Eqs. 1 to 3. Precision is a measure of how few false positives an assay will produce, and the opposite is true for low precision. The classifier's sensitivity, or the number of positive results it returns, is calculated using recall. False negatives are reduced with a higher level of recall and predicted instances divided by accurate classifications are called recall. When precision and recall are considered, the f1-score measure is created as a harmonically weighted average of the two metrics.

C. Datasets
In this study Indonesian keywords like ‘pembelajaran daring’, ‘kuliah’, ‘belajar’, ‘online’, ‘daring’ and the hashtags #BelajarDariRumah, #KuliahOnline, and #BackToSchool collected Indonesian tweets in order to collect proposed data. The data crawling process is carried out manually using the Orange3 tools and an access token obtained from the Twitter API. Because of the limitations of manual crawling, data collection continued with Twitter API tools, resulting in 13,000 tweets relating to the chosen keywords in early April 2022.

Orange3’s data crawling process is shown in Figure 2. To connect to Twitter, enter the access token code provided by the Twitter API using some operators, such as Twitter search. While performing a data crawl, the operator for removing duplicates is used. As shown in Figure 2, the next operator, select attributes, retrieves the necessary attributes such as username and text. Last but not least, we have to write excel, which is used to save data in Excel form. Using Orange3, we collected data in the form of usernames and texts displayed in Table 1.

<table>
<thead>
<tr>
<th>Username</th>
<th>Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>@schfess</td>
<td>Sch! Jujur gaes, kalian lebih pilih sekolah offline apa online? Temen2 ku ko pda pengen online trs ya pdaal q pusing klu daring tuh</td>
</tr>
<tr>
<td>@AmbarwatiRexy</td>
<td>Eh meeting online itu ternyata gak se-rek meeting beneran yak? Jadi sekarang bisa ngebayangin gimana rasanya belajar online</td>
</tr>
<tr>
<td>@Bai_haqiii</td>
<td>Berkah kuliah online : gak perlu repot&quot; buat nyatet, jadi bisa fokus dengerin materi</td>
</tr>
<tr>
<td>@ladaa</td>
<td>Kendala kuliah online gue tiap bimbang gue paling soebel kalau pas lagi assistensi terus tiba2 emak gue buka pintu triak nyuruh cuci piring pdhl lagi video call</td>
</tr>
</tbody>
</table>
When data is being preprocessed, processes like tokenization and data conversion are used to reduce clutter and make important features easier to see. After removing unnecessary attributes like URLs and other types of punctuation (such as the @ symbol), this stage is about determining the data class. When the main process is completed, this process will turn a text into data that the system can easily accept [18]. Before implementing the algorithm, the preprocessing stage is required. As shown in table 3, lowercase and removed URL links are folded to make the letters used to lowercase homogeneous.

When compared to the precision, recall, and f1 scores for SVM, RF, DT, LR, and KNN-based models, it was found that SVM-based models consistently outperformed the RF-based models in terms of their f1 scores. In terms of precision and recall value, there is no clear hierarchy between SVM-based and RF-based prediction models. A comparison of the precision, recall, and f1-scores for the results of preprocessing data yields a list of words that appear the most frequently. Using the word clouds shown in Figure 3, we can see how frequently these words are used in our dataset.

### TABLE III

<table>
<thead>
<tr>
<th><strong>Before</strong></th>
<th><strong>After</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Kendala kuliah online gue tiap bimbingan gue paling sebel kalau pas lagi asistensi terus tiba2 emak gue buka pintu triak nyuruh cuci piring pdhl lagi vidio call sama dospem.</td>
<td>Kendala kuliah online gue tiap bimbingan gue paling sebel kalau pas lagi asistensi terus tiba2 emak gue buka pintu triak nyuruh cuci piring pdhl lagi vidio call sama dospem.</td>
</tr>
<tr>
<td>&quot;kuliah&quot;, &quot;online&quot;, &quot;bimbingan&quot;, &quot;asistensi&quot;, &quot;emak&quot;, &quot;pintu&quot;, &quot;cuci&quot;, &quot;vidio&quot;, &quot;call&quot;</td>
<td>&quot;lecture&quot;, &quot;online&quot;, &quot;guidance&quot;, &quot;assistance&quot;, &quot;mother&quot;, &quot;door&quot;, &quot;washing&quot;, &quot;video&quot;, &quot;call&quot;</td>
</tr>
</tbody>
</table>

Grouping the results of preprocessing data yields a list of words that appear the most frequently. Using the word clouds shown in Figure 3, we can see how frequently these words are used in our dataset.

### FIG. 3
Positives and negatives of word cloud.
obtained using the DT, LR, and LR-based models, SVM- and RF-based models have the highest or in the worst case, are similar. Based on the KNN, in the case of neutral text, the SVM-based model's precision score is lower than the LR-based model's precision score, which is the only exception. The same holds true for RF-based models' precision scores and recall. In most cases, these models had higher precision and recall values than the DT-based, LR-based, and KNN-based models. There are, however, some exceptions to this rule. Precision values for positive text are lower with the RF-based model compared to the LR-based model's precision score, which is the only exception. The same holds true for RF-based models' precision scores and recall. In most cases, these models had higher precision and recall values than the DT-based, LR-based, and KNN-based models is lower than that obtained from the RF-based model. For both neutral and negative texts, the RF-based model's recall scores were lower than the corresponding recall scores from the DT-based model.

When compared to the scores obtained using the DT-based, LR-based, and KNN-based models, the SVM- and RF-based models typically have the highest or, in the worst case, similar results. In the case of neutral text, the SVM-based model's precision score is lower than the LR-based model's precision score, which is the only exception. The same holds true for RF-based models' precision scores and recall. In most cases, these models had higher precision and recall values than the DT-based, LR-based, and KNN-based models. There are, however, some exceptions to this rule. Precision values for positive text are lower with the RF-based model compared to DT and KNN models, respectively. For negative texts, the recall score obtained using the DT-based, LR-based, and KNN-based models is lower than that obtained from the RF-based model.

The hyper-parameters were set by cross-validation with the SVM method, which took about 2 hours and 36 minutes. The RF-based model required 2 hours 46 minutes of training time. To train these models, the training time ranged from 1 to 9 seconds for each of the DT, KNN, and LR models, respectively.

Figure 4 shows the SVM-based model's confusion matrix results. Classifier correctly predicted 6142 out of 6437 positive tweets, but incorrectly predicted neutral and negative tweets in the same number of cases. The classifier correctly predicted 1142 of the 1576 neutral tweets, while incorrectly classifying 682 and 144 as positive and neutral, respectively. The classifier correctly predicted 2136 of the 2467 tweets as negative, while 205 and 126 of the classifier incorrectly classified as positive and neutral.

Similar results are shown in Fig. 5, which shows the confusion matrix for the RF-based model. The classifier correctly predicted 6387 out of 6437 positive tweets, while only 50 of the negative tweets were incorrectly classified. The classifier correctly predicted 750 of the 1576 neutral tweets while incorrectly classifying 682 and 144 as positive and negative, respectively. Negative tweets accounted for 1792 of the 2467 tweets classified by the classifier as such, while positive tweets accounted for 673 and neutral tweets accounted for 2.

Each positive and negative tweet yielded 13,000 labeled samples. After 100-fold cross-validation, the best KNN-based model yields a f1 score of approximately 0.73. As part of our research, we used KNN to separate the tweets into three

### TABLE V
<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>f1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.81</td>
<td>0.72</td>
<td>0.77</td>
</tr>
<tr>
<td>Positive</td>
<td>0.93</td>
<td>0.95</td>
<td>0.94</td>
</tr>
</tbody>
</table>

### TABLE VI
<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>f1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>0.90</td>
<td>0.73</td>
<td>0.80</td>
</tr>
<tr>
<td>Neutral</td>
<td>1.00</td>
<td>0.48</td>
<td>0.64</td>
</tr>
<tr>
<td>Positive</td>
<td>0.82</td>
<td>0.99</td>
<td>0.90</td>
</tr>
</tbody>
</table>

### TABLE VII
<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>f1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>0.81</td>
<td>0.79</td>
<td>0.80</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.77</td>
<td>0.73</td>
<td>0.75</td>
</tr>
<tr>
<td>Positive</td>
<td>0.91</td>
<td>0.93</td>
<td>0.92</td>
</tr>
</tbody>
</table>

### TABLE VIII
<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>f1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>0.84</td>
<td>0.79</td>
<td>0.82</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.87</td>
<td>0.38</td>
<td>0.53</td>
</tr>
<tr>
<td>Positive</td>
<td>0.84</td>
<td>0.97</td>
<td>0.90</td>
</tr>
</tbody>
</table>

### TABLE IX
<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>f1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>0.64</td>
<td>0.83</td>
<td>0.75</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.65</td>
<td>0.65</td>
<td>0.65</td>
</tr>
<tr>
<td>Positive</td>
<td>0.91</td>
<td>0.83</td>
<td>0.87</td>
</tr>
</tbody>
</table>
categories: negative, positive, and neutral. Our KNN-based model’s f1 scores were 0.75, 0.65, and 0.87 for negative, neutral, and positive tweets, respectively, which are marginally better than Rumelli et al. [9]’s average. Using SVM and RBF to build our model further enhances classification performance. To represent these results, we calculated f1 scores of 0.87 for negative tweets, a neutral 0.77, and a positive 0.94 for our SVM-based model. In addition, our RF-based model’s f1 scores were 0.80, 0.64, and 0.90 for negative and neutral tweets.

IV. CONCLUSION
Numerous educational institutions use sentiment analysis to process large amounts of data more efficiently and cost-effectively. An educational institution’s ability to quickly gauge the general opinion of their students’ wants and needs is made possible by using sentiment analysis. People’s arguments, social media chatter, and more can be sorted out automatically so that one can make faster and more accurate decisions. Data and competitive analysis are powered by sentiment analysis. To discover new knowledge types or anticipate future trends, sentiment analysis can be extremely useful. There are numerous benefits to employing a sentiment analysis tool, including a reduction in both time and money. A supervised machine learning method was used in this study to develop a model for predicting the sentiment expressed in the text on social media. A number of different classifiers have been used to sort tweets into three categories: positively received, negatively received, and neutral. The f1 values of our KNN-based model were measured at 75%, 65%, and 87% for negative, neutral, and positive tweets, respectively, which are slightly more accurate than previous studies with the same method and purpose. People’s tweets on Twitter data are used to predict the sentiment of their tweets using sentiment analysis. Prediction models based on sentiment analysis SVM and RF were found to outperform other sentiment analysis models built based on KNN. SVM and RF can be used to classify students’ needs and opinions into positive, negative, and neutral groups within an acceptable error rate.

As a follow-up project, we intend to divide our training dataset into a k-fold number to enhance the sentiment analysis model’s efficiency. It is possible to improve the accuracy of the training data by incorporating more words with polarity labels into it. There will be a search for new words to be added to the training data, and data from all Indonesian abbreviated words will be used to improve accuracy in future research.

REFERENCES