Sentiment Analysis for Abolition of National Exams in Indonesia using Support Vector Machine

Erlin*, Irma Suliani, Hadi Asnal, Laili Suryati and Riswan Efendi

***Abstract*—Through the Minister of Education and Culture, the Indonesian government has announced the abolition of the implementation of the National Examination for grades 6, 9, and 12 in Elementary, Junior, and Senior High Schools. The abolition of this national examination received various responses and comments from numerous groups ranging from leaders to the general public. Therefore, this study aims to apply sentiment analysis in the data mining approach to analyze the textual data of Twitter using a Support Vector Machine (SVM), explore the public's opinion toward the abolition of the national exam, and measure the policy's level of acceptance. Furthermore, this study conducted experiments using two other classifiers, namely Random Forest (RF) and Logistic Regression (LR) to deeply observe the performance of SVM. It also conducted scenarios using two different feature extraction, TF-IDF and Bag of Words, and analyzed how it impacts improved accuracy. The experimental results showed that the combination of SVM with Polynomial Kernel and TF-IDF provides performance compared to RF and LR at C=0.01 and degree=20 with accuracy, precision, recall, and F1 score values of 96.97%, 97.28%, 96.87%, and 96.90%, respectively. Furthermore, the result showed that the SVM model with a polynomial kernel provides higher algorithm performance on text classifiers. Therefore, the government can utilize this sentiment analysis as an evaluation material for the decision to abolish the national exam, with most public agreeing with this policy.**

***Index Terms*—Data mining, polynomial kernel, sentiment analysis, support vector machine, text classification**

I. Introduction

EVERY country has a unique way of determining students' passing standards at all levels of education. In Indonesia and most countries, the National Examination is one way to determine students' ability to continue to the next level of education. This standardized means of measurement has been in effect for decades.

In November 2019, the Minister of Education and Culture of Indonesia proposed abolishing the national examination to prevent detrimental consequences, such as high anxiety levels in students, parents, and teachers. In addition, the examination was unsuccessful in enhancing the standard of education, with126 reported cases of cheating across the country in 2019, approximately 59 percent from the previous year [1]-[2]. However, this idea was opposed by Yusuf Kalla, Ex. Vice President of the Republic of Indonesia argued that its removal might harm the country's education quality. Furthermore, Kalla argued that the national exam is still the essential benchmark for Indonesian students, which aligns with the opinion of numerous school principals [3].

Irrespective of these oppositions, the Ministry finally announced the abolition of the National Exams and Equality Exams in 2021 through Circular Number 1 of 2021. Instead, the examination was changed by a "minimum competency assessment and character survey," designed to measure the numeracy and literacy levels of students in grades four, eight, and eleven using Program for International Student Assessment (PISA) principles [4]. Conversely, this abolition has led to numerous pros and cons among the public and observers of education in Indonesia. Therefore, it is imperative to conduct sentiment analysis on this abolition process to determine society's sentiment tendency to provide government references and help decision-makers improve their policies.

Sentiment analysis has been one of the most active research areas in natural language processing. It is also known as opinion mining, which analyzes people's sentiments, assessments, perspectives, and emotions on objects and their attributes stated in the written text [5]. The massive availability of opinion and sentiment data on social and online media has led to its broader application.

The growing popularity of research in sentiment analysis on Twitter has prompted various surveys to determine the techniques used for supervised and unsupervised learning [6]-[7]. The expansion of the study on the subject domain, such as the level of analysis, from the document to the aspect levels, has also increased [8].

SVM is a model used for many applications and can handle linear and nonlinear problems. It provides a high degree of performance and is frequently used compared with

Manuscript received November 25, 2021; revised September 6, 2022. This work was supported by Data Science and Artificial Intelligence Team of Institut Bisnis dan Teknologi Pelita Indonesia and STMIK Amik Riau. (Corresponding Author: Erlin).

Erlin is an Associate Professor of Informatics Engineering Department, Institut Bisnis dan Teknologi Pelita Indonesia, Jln. Jend. Ahmad Yani, Pekanbaru, 28127, Indonesia; email: erlin@lecturer.pelitaindonesia.ac.id

Irma Suliani is an Alumni of Informatics Engineering Department, STMIK Amik Riau, Jln. Purwodadi Indah, Km. 10, Panam, Pekanbaru, 28294, Indonesia; email: irmasuliani1@gmail.com

Hadi Asnal is a Senior Lecturer of Informatics Engineering Department, STMIK Amik Riau, Jln. Purwodadi Indah, Km. 10, Panam, Pekanbaru, 28294, Indonesia; email: hadiasnal@sar.ac.id

Laili Suryati is a Senior Lecturer of Accounting Department, Universitas Persada Indonesia YAI, Jln. Diponegoro No. 74 Jakarta Pusat, Indonesia; email: lailisuryati@yai.ac.id

Riswan Efendi is an Associate Professor of Mathematics Department, Universiti Pendidikan Sultan Idris, Tanjong Malin, Perak, 35900, Malaysia; email: riswanefendi@fsmt.upsi.edu.my

the decision tree [9], Naïve Bayes [10]-[11], and other model classifiers [12]. Furthermore, SVMs have a high degree of generalization. It is also notable for its robustness to high-dimensional data [13]. Gosh and Sanyal [14] compared and evaluated various machine learning models, namely Naïve Bayes, Support Vector Machine, K-Nearest Neighbor, and Maximum Entropy, based on three standard datasets. The results showed that SVM performed better with a 90.24% higher level of accuracy on composite IG than NB, K-NN, and ME at 88.04%, 83.45%, and 85.62%. Moreover, Prastyo et al. [15] used SVM with a normalized poly kernel to execute sentiment analysis on Tweets in response to the Indonesian government's handling of the pandemic. The findings demonstrated that the SVM algorithm is an intelligent algorithm used to forecast sentiment on Twitter for new data in a timely and accurate manner.

This study used the SVM model to classify the opinion and sentiment of the publics in the written text into "positive," "negative," and "neutral" ternary sentiment classification. It also demonstrated the ability of the proposed method to present good performance in terms of accuracy, precision, recall, and F1 Score. Studies focusing on sentiment analysis in eliminating national exams have not been previously conducted, irrespective of the various publications dealing with multimodal data and text-based sentiment analysis. The result is significant at the C=0.01 and degree=20 with accuracy, precision, recall, and F1 score values of 96.97%, 97.28%, 96.87%, and 96.90%, respectively. This study also conducted experiments using two other classifiers, namely Random Forest and Logistic Regression to deeply observe the performance of SVM. Therefore, this study is considered the first in attempting to employ an SVM, which is the best machine learning classification, for text-based sentiment analysis in public opinion on abolishing national exams in Indonesia.

The remaining sections of this paper are organized as follows: Section II discusses related work, while section III provides explanations and illustrations of the method using SVM for sentiment analysis. Section IV discusses the experiment, while results and discussion is analyzed in section V. Finally, the conclusion is illustrated in section VI.

II. Related Works

In recent years, data mining associated with sentiment analysis has become a popular research topic. However, the type of data mined varies greatly depending on the goal and expected outcome; hence, the methods for processing data and extracting the required information differ with expanded and deepened knowledge in the domain.

Numerous literature on sentiment analysis pays significant attention to classifying the product, sales reviews, and stock market [16]-[19] with good results in this application domain [20]-[21]. Several studies focus on movie reviews using machine learning [22]-[23], including those in specific national languages [24], with robust results [25]. Research on sentiment analysis is not only concentrated on subject reviews. Still, it is also widely used to predict the battle for presidential candidates in the general election based on discussion forums on several social media [26]-[28]. Barghuthi and Said [29] stated that the proposed method has a high accuracy prediction percentage than actual election results.

Conversely, other studies focused on techniques to enhance the accuracy of the result using ensemble classifiers such as Gradient Boosted Support Vector Machine (GBSVM) for sentiment classification based on unstructured review [30] and ensemble classification with concept drifts for short text data stream [31]. The excellent result is compared to the existing model [32]. Chang et al. [33] proposed the modified cluster-based oversampling technique to overcome the imbalanced text feature, which impacts achieving high accuracy.

However, most studies have emphasized on social media content and the strategies used to obtain user views on specific topics or objects. Ruiz and Bedmar [34] compared several deep learning architectures for sentiment analysis of drug reviews using Bidirectional Encoder Representations from Transformers (BERT) and a Bi-LSTM. The experiments showed that using BERT produces the best results with a prolonged training period, while Bi-LSTM provided satisfactory results with less training time. Similarly, the performance of Bidirectional LSTM (Bi-LSTM), Long Short-Term Memory (LSTM), and Convolutional Neural Networks (CNN) in sentiment analysis of Thai children's stories was tested and compared using a combination of POS-tag, word embedding and, sentic feature [35]. The experiments showed that the CNN model that used all three features produced the best results. In addition, Umer et al. [36] suggested using a convolutional neural network combined with long short-term memory for handling sentiment analysis on Twitter datasets. The results showed that the proposed method achieved higher accuracy than other classifiers.

A more recent and well-analyzed study led to the creation of new models and approaches. For example, Zuheros et al. [37] used sentiment analysis to enable decision-making models to make expert assessments in natural language. Zuheros et al. proposed the sentiment analysis-based multi-person multi-criteria decision-making approach for a more ingenious decision aid, which constructs expert assessments from natural language reviews and numerical ratings. Similarly, Huang et al. [38] developed a novel image–text sentiment analysis via multimodal attentive fusion to exploit the discriminative features and the internal connection between visual and semantic contents with a mixed fusion framework. This approach presented the effectiveness on both poorly and manually labeled datasets. Furthermore, the study conducted by Syarif et al. [39] was used to detect the negation effect on customer reviews by proposing a modified negation approach. The results improved classification accuracy and helped in calculating negation identification.

Other studies proposed machine learning on sentiment analysis to examine the immediate impact of President Trump's tweets on two US stock markets. They found that those with strong positive or negative sentiments had positive market reactions [40]. Similarly, Sharma and Sharma [41] implemented an automated Twitter sentiment analysis framework that uses machine learning algorithms to predict public emotions. Finally, the use of machine learning for sentiment analysis was reinforced by Rustam et al. [42].

They classified tweets based on Sentiments for US Airline Companies and found that they performed better when TF-IDF was used as the feature extraction. These three works proved that machine learning is the best method for sentiment analysis.

Despite the numerous publications dealing with text-based sentiment analysis and multimodal data, no studies focused on its use to eliminate national exams by simultaneously comparing the performance of three machine learning classifiers. Therefore, this study can be viewed as the first attempt to use an SVM to conduct a text-based sentiment analysis of public opinion on abolishing national exams, especially in Indonesia.

III. Research Method

The Support Vector Machine (SVM) method was to classify a collection of opinions/sentiments from the public on abolishing the national exam in Indonesia. Classifying the polarity of a given text in the document needs a series of works, as shown in Fig. 1. The stages start from data collection using a web crawler, determining the dataset, the final stage, testing, and evaluating the SVM model.

A. Handling Skew

Of the 900 imbalanced datasets, 548 (60.89%), 203 (22.56%), and 149 (16.55%) were labeled “positive,” “negative” and "neutral." This led to many false-positive predictions for underrepresented examples. This study applied the undersampling of positive class and oversampling of the negative and neutral in enforcing comparable sizes within the training set while maintaining balanced and overcoming skewness in the original data distribution. The amount of data after resample is 1644.

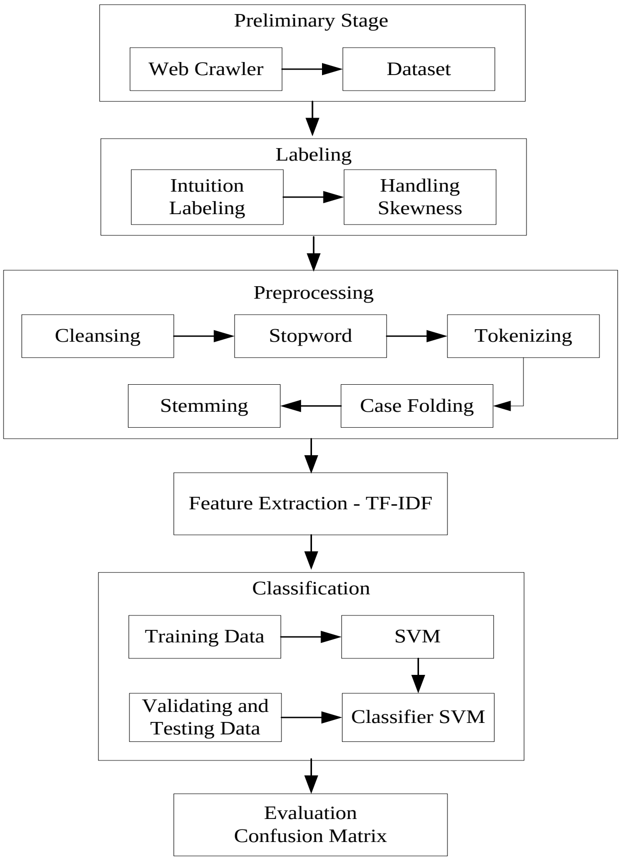


Fig. 1. A framework of sentiment analysis for text classification

B. Pre-Processing

Data pre-processing is the first and most important step in building a machine learning model for preparing raw data. Before performing any data-related procedure, it is essential to clean and format the data because it generally contains noises, missing values, and an unsuitable format that cannot be used for machine learning models through the pre-processing process. This stage is required to clean and make data suitable for a machine learning model, increasing classification accuracy and data quality. In addition, the relevant or structured data makes the process of classifying sentiment analysis easier and simpler. It involves several steps: cleansing, stopword removal, tokenizing, case folding, and stemming.

The cleansing task is associated with removing unnecessary tweets to reduce noise. This method removes punctuation and special characters on Twitter, such as retweet/RT, username (@username), hashtag (#), URLs, and Unicode emojis. Additionally, this study transformed Indonesian slang terms and informal abbreviations into their formal synonyms.

Stopword removal eliminates words that do not add much meaning to a sentence or convey any meaningful message. A stoplist was created to keep track of the stopwords in the dataset and eliminate similar words or tokens. Approximately 357 words are classified as stopwords, a vocabulary comprising conjunction and adverb in Indonesian languages, including about, this, and, after, etc.

Tokenizing is associated with splitting a sentence into several words or breaking the sequence of strings into pieces called tokens. In the tokenizing process, spaces function as delimiters, while the word tokenize is the NLTK package for Python used to complete this operation.

Case folding is converting all the characters in a document into upper or lower cases to speed comparisons during the indexing process. The characters processed are only letters 'a' to 'z' with the exclusion of punctuations and unnecessary symbols.

Stemming is the last step of pre-processing, which removes prefixes, suffixes, and inserts and combines both prefixes and suffixes from a word and reduces it to its root word. The stemming used in this study is in accordance with the Sastrawi Library provided by Python.

C. Feature Extraction

Machine learning algorithms cannot directly work on the raw text; therefore, this study employed the term frequency-inverse document frequency (TF-IDF) method to convert text into a feature matrix (or vector). It highlights a specific issue, which is not frequent in a corpus, irrespective of its importance. The TF–IFD value increases proportionally to the number of times a word appears in the document and decreases with the number of documents in the corpus. TF-IDF is the product of TF and IDF formulated as in Eq. (1), tf (t, d) in Eq. (2), and idf (t, D) in Eq. (3).

$$tf - idf(t,d,D) = tf(t,d).idf(t,D) \qquad (1)$$

where:

$$tf(t,d) = \log(1 + freq(t,d)) \qquad (2)$$

$$idf\,(t,D) = \log\left(\frac{N}{count(d \in D) : t \in d}\right) \tag{3}$$

tf (t, d) is the term frequency that specifies how frequently a term (t) appears in the entire document (d) and idf (t, D) is the inverse document frequency from the term (t). N is the total number of documents in the corpus, and df is the number of documents containing the term t.

The scikit-learn library of Python in-build function TF-IDF vectorizer was used in this study to calculate the TF-IDF score of any corpus. It was also used to obtain a high and low TF-IDF score in the document and corpus. For a word that appears in almost all documents, the idf value approaches 0; hence the TF-IDF is closer to 0. On the other hand, the TF-IDF value is high when both idf and tf values are high. This high score means that the word is rare in the whole document and frequent in some parts.

D. Support Vector Machine

The next step after the pre-processing stage is modeling the text classifier using SVM, which must be trained before the categorization process [43][44]. The phase determines the kernel parameter value using a constant variable that allows to trade off the influence of the higher and lower order terms, which is a factor to consider when changing C values between 0.01 and 1. The degree of the polynomial is indicated by the variable d and values used in this study between 1 and 20.

The values of d impact performance accuracy, while C is chosen as a constraint based on the C function. Therefore, a higher value of C indicates a more significant penalty for classification errors. The Polynomial kernel's C and d had the highest accuracy of parameter pairings during the training period. After conducting a series of experiments using pairs of C and d values, the best pair was C = 0.01, with a value of d = 20. A polynomial is a kernel used for three or more dimensions. The data used in this study has more than three dimensions or features; therefore, the kernel used is polynomial.

Polynomial Kernel Function

The kernel is a way of computing the dot product of two vectors x_i and $x_{j,}$ in some (very high dimensional) feature space. The kernel function is sometimes called "generalized dot product" and comes to the major part of the SVM for which the kernel trick is most famous. The kernel function calculation is performed in Eq. (4).

$$k(x_i, x_j) = \Phi(x_i).\Phi(x_j) \tag{4}$$

The kernel function k is substituted into the dual of the lagrangian, allowing the determination of a maximum margin hyperplane in the (implicity) transformed space Φ(x). The standard kernel functions for SVM are Linear function, Polynomial kernel, Gaussian or Radian Basis Function (RBF), and Sigmoid function. This study uses the polynomial kernel as a popular kernel function for SVM [45][46]. The polynomial kernel is defined in Eq. (5).

$$k(x_i, x_j) = (x_i.x_j + c)^d \tag{5}$$

where c is the constant term, and d is the polynomial degree.

In the polynomial kernel, the process calculates the dot product by increasing the kernel's power. For example, originally 2-dimensional vectors $x = (x_i, x_j)$; and $k(x_i, x_j) = (x_i.x_j + 1)^2$. Need to show that

$$k(x_i, x_j) = \Phi(x_i).\Phi(x_j).k(x_i, x_j) = (x_i.x_j + 1)^2 = (x_i, x_j + 1).(x_i, x_j + 1) =$$
$$2.x_i x_j + x_i{}^2 x_j{}^2 + 1 = (\sqrt{2x_i}, x_i{}^2, 1).\sqrt{2x_j}, x_j{}^2, 1)$$

Thus, a kernel function implicitly maps data to a high dimensional space without explicitly needing to compute each Φ(x). This kernel is very useful for rewriting training data using more complex features, and datasets not linearly separable in original space may be linearly separable in higher dimensional space.

E. Evaluation

The model evaluation metrics used in this research are accuracy, precision, recall, and F1 Score, consistent with those used in other preliminary studies. The accuracy of the SVM modeling is calculated using the confusion matrix. In contrast, the work process of this confusion matrix is obtained by comparing the label from the predetermined data test with the label's prediction from the grouped test data. The result of applying this evaluation technique is determining errors in values during the classification process.

IV. Experiment

A total of 1000 tweets were split into two parts. The first part, which serves as the "training data" contain 900 tweets manually classified into positive, negative, and neutral classes. The second part contains 100 tweets manually checked and classified into three classes. This part is a test set called the "testing data."

Fig. 2 displays the proportion of classification on abolishing the national exam from 900 data tweets containing positive, negative, and neutral is 60.89%, 22.56%, and 16.55%, respectively, whereas, Fig. 3 shows a proposed framework for classifying text document data on Twitter for sentiment analysis using SVM. It contains several process sequences carried out at the initial stage after the pre-processing and manual labeling.

a. Weight and normalize data for each term by calculating the frequency of occurrence;
b. Split the training data and testing data;
c. Determine the input (x), the weighted term, and the target output (y) for the positive, neutral, and negative classes;
d. Map the nonlinearly separable data to higher features using kernel functions;
e. Determine the best SVM parameter values used in the classification process. From a series of tests for the value of C = (0.01 - 1) with degree = (1 - 20) the best value of C = 0.01 with degree = 20;
f. Create a classifier model in the training process;
g. Classification of testing and validating data based on training data modeling.

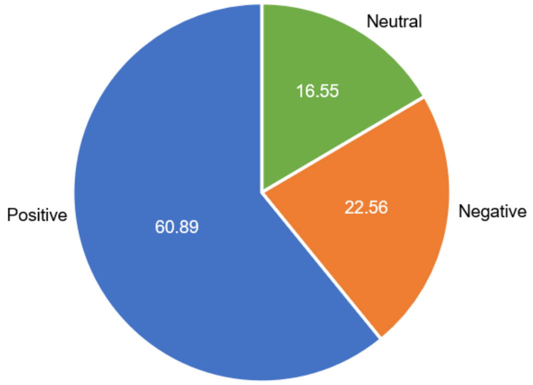


Fig. 2. Distribution of positive, negative, and neutral tweets

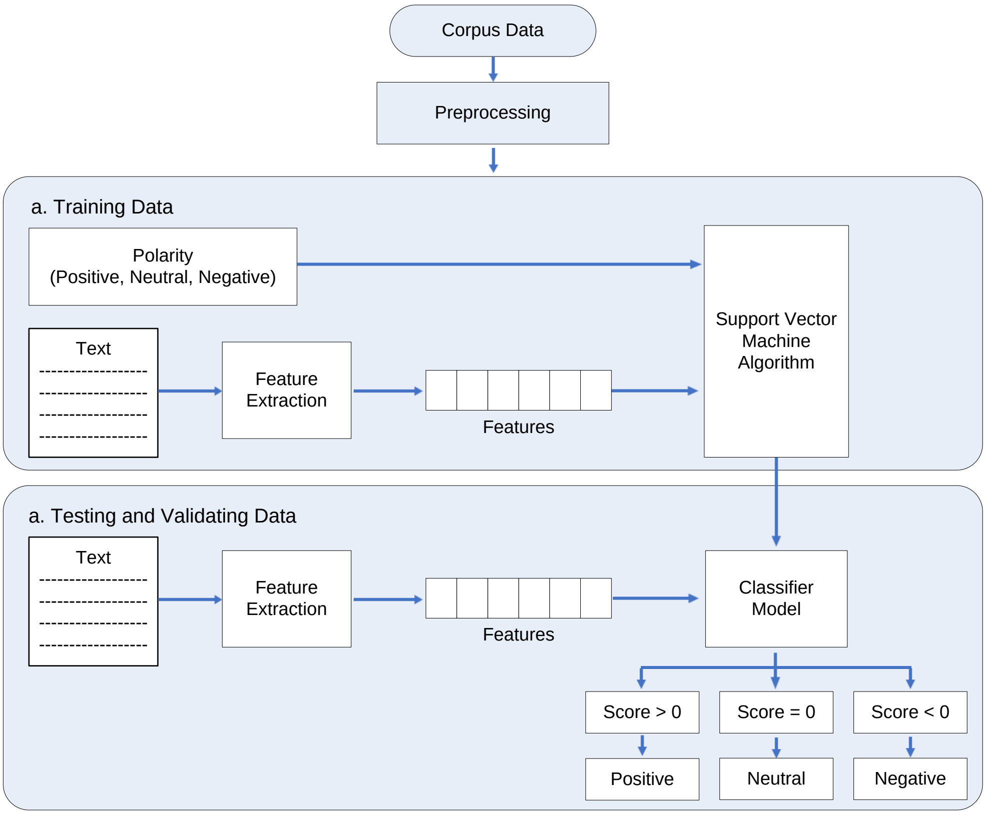


Fig. 3. SVM classification on sentiment analysis

Table I shows the excerpts of original data obtained through crawling technology on Twitter. Unfortunately, many tweets are still in the form of unstructured text and informal language containing particular words and symbols, including emoticons and slang. In addition, the acquired document is a labeled text that has not been cleaned, while the training data to be processed is still mixed with other characters attached to the data.

Data is further processed in the pre-processing phase, with several steps used to remove noise, clarify features, convert original data to fit needs, and enlarge and reduce relevant data. The acquired documents cleansed special characters such as html, hashtag, site address (http: www/site.com), username (@username), punctuation marks (.,?! [] /%:; < > () *), numbers (0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 0) and other characters. This phase is important to obtain valid data to be processed in the next steps.

The stopword phase removes unnecessary words, such as conjunctions e.g, then, which, with, or, will, to, as, by, and words that have no meaning. This phase is tricky because public comments on social media in Indonesia are free and unstructured. For example, the deleted word in document 1, which is taken as a sample is "saya (I) " "ketika (went)” and "saja (still)" according to the Indonesian stopword dictionary.

The tokenizing or parsing phase breaks down a document into words, which are analyzed by separating the words and determining their syntactic structure. This phase is followed by the case folding phase. Then, in the uniformity of letterforms in documents, capital letters are changed to lowercase, and uniform letters from A to Z, while others are removed because they are considered a delimiter. The last phase, known as stemming, is the pre-processing phase, which converts words into basic structures by removing affixes, i.e., prefixes, infixes, suffixes, and combinations of prefixes and suffixes. All pre-processing steps are finished using Python's scikit-learn library, making the processing more efficient. Table II shows a sample of training data before and after pre-processing.

TABLE I
THE EXCERPTS OF ORIGINAL DATA ON TWITTER

No	*Tweets*
1	Saya ketika sekolah ikut UN, tetap aja lembek. #ujiannasional #hapusUN (*When I was in school, I took the national exam, and my score was low. #nationalexam #deletenationalexam*)
2	Tul. Buktinya skripsi juga minta dihapus. Lemah su! bendinone ambyar. Nom noman gatel #ujiannasional #ujianhapus #Unhapus (*That's right. The final project also needs to be deleted. So weak #nationalexams #deleteexams #deletenationalexams*)
3	UN ganti dengan Ujian Internasional saja #ujiannasional #hapusUN (*Just replaced national exams with international exams #deleteexams #deletenationalexams*)
4	UN bukan TOLOK UKUR. Kalo pak JK jd orang tua yang ikut stres pasti nggak bicara seperti ini. Pihak sekolah, siswa serta orang tua mencari cara bagaimana supaya lulus UN. Dr sini sudah jelas bahwa UN tidak lagi bisa dijadikan tolok ukur. Dulu jaman saya NEM jada standart masuk .. sekarang? #ujiannasional #hapusUN (*The national exam is not a benchmark. Assuming Mr. Jusuf Kalla was a stressful parent, he would not have talked like this. The school principals, students, and parents are looking for ways to make their children pass the national exams; therefore, it is clear that it no longer acts as a benchmark. Back in my day, the pure evaluation score became the entry standard … Now? #nationalexams #deletenationalexams*)
…	…
1000	Bukan menolak tetapi mengingatkan bedalah maksudnya #ujiannasional (*Not refusing but reminding, it is the meaning of #nationalexam*)

TABLE II
SAMPLE OF TRAINING DATA

Label	Tweet (Before pre-processing)	After pre-processing
Positive	Senang mendengar informasi ini, setujuh saia dengan keputusan mas menteri. Gebrakan baru, maju terusss. #HapusUN #UjianNasional	senang dengar informasi setuju putusan menteri
Neutral	Dihapus atau tidak ujian nasional bagi saya sama saja. Tetep ga ngaruh tuh ke masa depan saya, tul gak, hehe	hapus tidak ujian nasional sama saja tetap pengaruh masa depan
Negative	Menghapus ujian nasional berarti akan menurunkan kualitas pendidikan di Indonesia. Ini berbahaya lho, gimana dunk SDM +62 nantinya? wong selama ini udah memble tambah memble lagi. #ujiannasional #hapusUN	hapus ujian nasional arti turun kualitas didik indonesia bahaya sumberdaya manusia selama sudah memble tambah

V. RESULT AND DISCUSSION

The processing stage is the next phase after the pre-processing, resample, and word weighting (TF-IDF) phases. There are two essential sub-phases: classification with the support vector machine (SVM) method and data accuracy.

The number of trained and tested data after the pre-processing phase is 1644, divided into training and testing data in a ratio of 90%: 10%, as shown in Table III. The split feature vector is classified using SVM with a polynomial kernel function that maps nonlinear data, thereby providing a new learning model dataset for each experiment.

The most crucial task in experimenting is selecting the parameters of the SVM learning machine and the kernel polynomial function, namely parameters C and d (degree). The value C> 0 is the independent parameter in the polynomial. Therefore, the kernel is homogeneous when C = 0.

The classification learning model is summarised in Table IV, with a 3x3 matrix used to represent the actual and predicted classes. The products of the learning model were tested using new data which had not been previously trained.

TABLE III
SPLITTING TRAINING DATA AND TESTING DATA

Ratio		A total of data	
Training data	Testing data	Training data	Testing data
90%	10%	1479	165

TABLE IV
TERNARY CLASSIFICATION CONFUSION MATRIX

Actual class	Predicted class		
	Category positive	Category negative	Category neutral
Category positive	tp	pn	pnu
Category negative	np	tn	nnu
Category neutral	nup	nun	tnu

Several measures in information retrieval and machine learning have been defined based on this classification confusion matrix. For instance, the evaluation measures of precision, recall, F1 Score, and accuracy to determine text classifiers are shown in Eqs. 6a-6c, 7a-7c, 8, and 9.

$$ClassPositive\,\mathrm{Pr}\,ecision = \frac{tp}{tp+np+nup} \tag{6a}$$

$$ClassNegative\,\mathrm{Pr}\,ecision = \frac{tn}{pn+tn+nun} \tag{6b}$$

$$ClassNeutral\,\mathrm{Pr}\,ecision = \frac{tnu}{pnu+nnu+tnu} \tag{6c}$$

$$ClassPositive\,\mathrm{Re}\,call = \frac{tp}{tp+pn+pnu} \tag{7a}$$

$$ClassNegative\,\mathrm{Re}\,call = \frac{tn}{np+tn+nnu} \tag{7b}$$

$$ClassNeutral\,\mathrm{Re}\,call = \frac{tnu}{nup+nun+tnu} \tag{7c}$$

$$F1Score = \frac{2(recall.precision)}{(recall+precision)} \tag{8}$$

$$Accuracy = \frac{tp+tn+tnu}{tp+tn+tnu+fp+fn+fnu} \tag{9}$$

TABLE V
CONFUSION MATRIX OF CLASSIFICATION RESULTS

Actual class	Predicted class		
	positive	negative	neutral
positive	58	0	1
negative	3	48	0
neutral	1	0	54

TABLE VI
PERFORMANCE MEASURE

Precision (%)			Recall (%)			F1 Score (%)		
Positive	Negative	Neutral	Positive	Negative	Neutral	Positive	Negative	Neutral
93.7	100	98.2	98.3	94.1	98.2	95.8	96.9	98.0
Average: 97.28			Average: 96.87			Average: 96.90		

$$Accuracy = \frac{58+48+54}{58+48+54+3+1+1} = \frac{160}{165} = 0,9697 = 96,97\%$$

Table V shows that 160 data (96.97%) are in the precise classification, while 5 data (3.03%) are on the contrary. Three negative classes were predicted as a positive class, 1 of neutral class was predicted as a positive class, and 1 of a positive class was predicted as a neutral class.

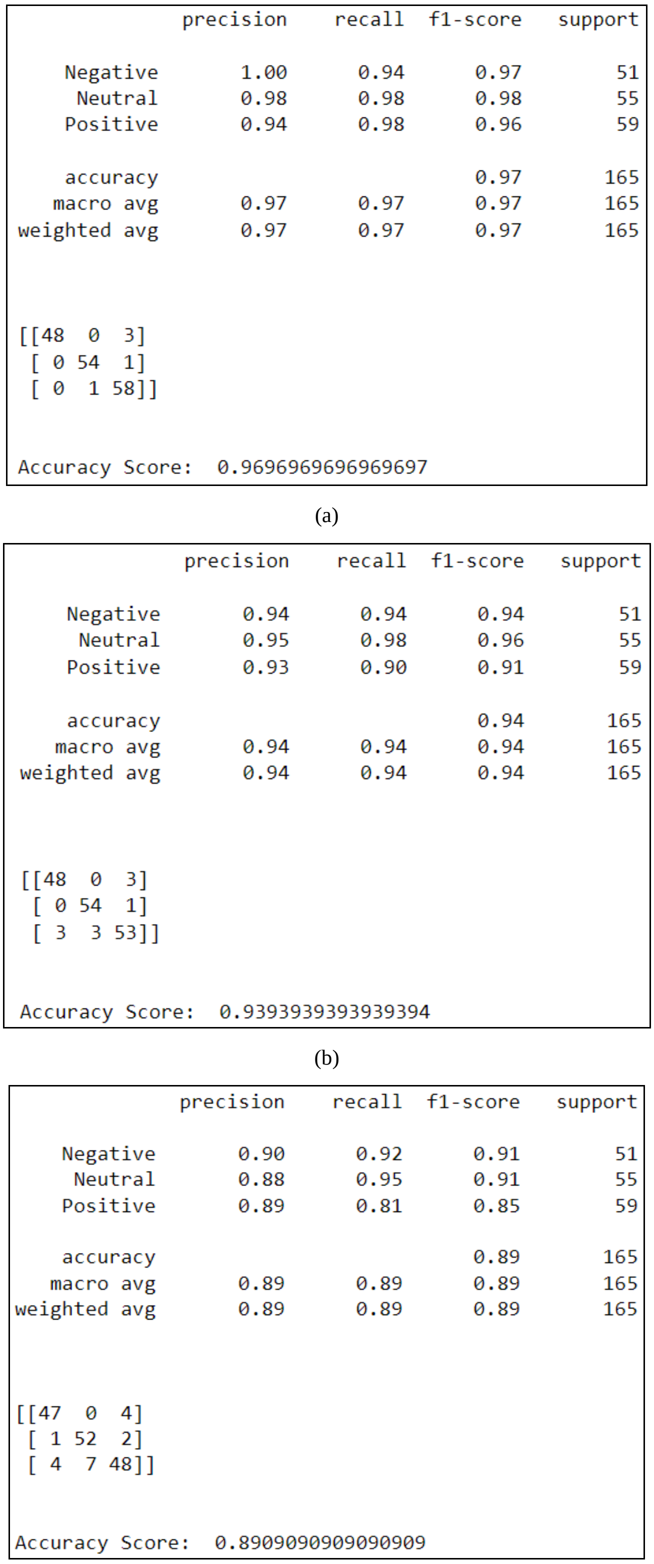
Table VI highly performs SVM as a text classifier on sentiment analysis. In precision term, class positive=93.66%, class negative=100%, and class neutral=98.18%, with average=97.28%. In recall term, class positive=98.31%, class negative=94.12%, and class neutral=98.18%, with average=96.87%. Performance measure for F1 Score also have a great result with class positive=95.83%, class negative=96.91%, and class neutral=97.96%, with average=96.90%. As for accuracy, the value is 96.97%, which is an excellent value for accuracy in text classification.

Measurement of accuracy, precision, recall, and F1 Score of SVM using the Python programming language is illustrated in Fig. 4(a). It shows that the precision, recall, and F1 Score for the negative and neutral classes are 100%, 94%, and 97%, respectively. Meanwhile, the positive class values are 94%, 98%, and 96%, respectively. The figure also shows that the SVM based on the polynomial kernel is an excellent method for classifying texts in sentiment analysis, with an average yield of 97% for four performance measures. This study used the scikit-learn version 1.0 that supports Python 3.7+. This tool is a robust library for machine learning, including classification.

Additionally, the performance measure of SVM is compared with Random Forest (RF) and Logistic Regression (LR) to evaluate the effectiveness of SVM significantly. Fig. 4(b) and Fig. 4(c) show that the RF and LR classifier performance still falls behind SVM. For example, the performance measure of Random Forest for Precision, Recall, F1 Score, and Accuracy is 93.95%, 94.04%, 93.99%, and 93.94%, respectively. In comparison, the performance measure of the Linear Regression classifier for the four main metrics are 89.14%, 89.35%, 89.24%, and 89.09% successively.

Table VII summarizes the performance comparison of the three classifiers, namely SVM, RF, and LR, in terms of precision, recall, F1 Score, and accuracy. The table shows that SVM outperforms RF and LR with percentage values of 96.97%, 93.94%, and 89.09%, respectively.

Furthermore, this study also experimented using two feature extraction methods: TF-IDF and Bag of Words (BoW). Table VIII shows a performance comparison based on four main metrics between SVM using TF-IDF and SVM using BoW. The experimental result shows that SVM with TF-IDF outperforms BoW in terms of precision, recall, F1 Score, and Accuracy. Therefore, SVM with Polynomial Kernel and TF-IDF became the chosen model because they perform better than other classifiers and feature extraction.



```
              precision    recall  f1-score   support

    Negative       1.00      0.94      0.97        51
     Neutral       0.98      0.98      0.98        55
    Positive       0.94      0.98      0.96        59

    accuracy                           0.97       165
   macro avg       0.97      0.97      0.97       165
weighted avg       0.97      0.97      0.97       165



[[48  0  3]
 [ 0 54  1]
 [ 0  1 58]]


Accuracy Score:  0.9696969696969697
```

(a)

```
              precision    recall  f1-score   support

    Negative       0.94      0.94      0.94        51
     Neutral       0.95      0.98      0.96        55
    Positive       0.93      0.90      0.91        59

    accuracy                           0.94       165
   macro avg       0.94      0.94      0.94       165
weighted avg       0.94      0.94      0.94       165



[[48  0  3]
 [ 0 54  1]
 [ 3  3 53]]


Accuracy Score:  0.9393939393939394
```

(b)

```
              precision    recall  f1-score   support

    Negative       0.90      0.92      0.91        51
     Neutral       0.88      0.95      0.91        55
    Positive       0.89      0.81      0.85        59

    accuracy                           0.89       165
   macro avg       0.89      0.89      0.89       165
weighted avg       0.89      0.89      0.89       165



[[47  0  4]
 [ 1 52  2]
 [ 4  7 48]]


Accuracy Score:  0.8909090909090909
```

(c)

Fig. 4. Performance measure, (a) Support Vector Machine (SVM), (b) Random Forest (RF), (c) Logistic Regression (LR)

TABLE VII
PERFORMANCE COMPARISON OF THREE MACHINE LEARNING CLASSIFIERS

Classifier	Sentiment Label	Precision (%)	Recall (%)	F1Score (%)	Accuracy (%)
Support Vector Machine	Positive	94	98	96	97
	Negative	100	94	97	
	Neutral	98	98	98	
Random Forest	Positive	93	90	91	94
	Negative	94	94	94	
	Neutral	95	98	96	
Logistic Regression	Positive	89	81	85	89
	Negative	90	92	91	
	Neutral	88	95	91	

TABLE VIII
PERFORMANCE COMPARISON OF SVM WITH TF-IDF AND SVM WITH BOW

Four main metrics	SVM with TF-IDF (%)	SVM with BoW (%)
Precision	97.28	86.69
Recall	96.87	86.22
F1 Score	96.90	85.37
Accuracy	96.97	85.45

Based on the created modeling, new data are classified as positive, negative, and neutral in agreeing, not agreeing, and not engaging in abolishing, with percentage values of 71%, 23%, and 6%, respectively.

Sentiment Result Analysis

The sentiment results on the abolition of the national exam showed that more people left positive sentiment tweets than negative and neutral sentiment, with 60,89%, 22,56%, and 16,55%, respectively. In other words, more people agree about the abolition of the national exam than those who disagree and are neutral. Further positive, negative, and neutral sentiment analysis was conducted using a word cloud. The presentation of the word cloud can be used as an illustration of understanding the intent of the reviews and comments written by the general public.

The word cloud and the word frequency of the positive sentiment is shown in Fig. 5 and Fig. 6. Numerous words convey positive sentiment, such as "ujian (exam)," "nasional (national)," "setuju (agree)," "adil (fair)," "dukung (support)," "bagus (good)" etc. The top five words that appear most frequently in tweets with positive sentiment are "ujian (exam)," "nasional (national)," "beban (burden)," "kasus (case)," and "guru (teacher)" are the five nouns that most often appear in positive sentiment tweets. These nouns are an object that gets a positive sentiment from the general public. In addition, several reviews contain words with positive sentiments, such as:

- Agree. Abolish the national exam, and focus on the interests and talents of students;
- Abolish the national exam to reduce the psychological burden on students, parents, and teachers;
- Teachers are the person who knows the competence of their students, not the government;
- Great minister, improving the quality of education does not have to go through a national exam, and
- Every year, the number of cases of foul and cheating on the national exam continues to increase.

All tweets conveyed their agreement and support for Indonesia's abolition of the national exam.

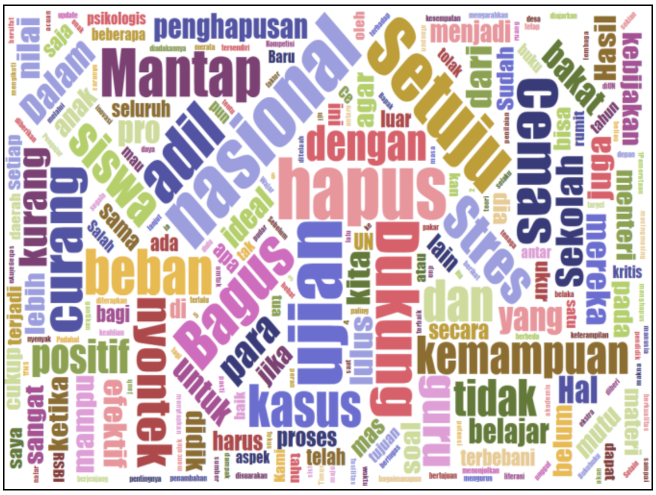


Fig. 5. Word cloud sentiment positive

On the other hand, Fig. 7 and Fig. 8 show the word cloud and the word frequency of the negative sentiment. The words that often appear include "penting (important)", "kualitas (quality)", "perlu (need)", "kompetensi (competence)", "pendidikan (education)" and "evaluasi (evaluation)". Some examples of tweets related to this word such as:

- National exams are important to measure the ability of students and teachers;
- Maintaining the quality of education requires a standard, namely national exams;
- Evaluating student achievement through national exams;
- National exams are essential for measuring student competence.

All these comments indicate disapproval of the abolition of the national exam.

Fig. 9 and Fig. 10 demonstrate the word cloud and the word frequency of the neutral sentiment. The words "ujian (exam)", "nasional (national)", "ada (there is)", "sama (same)", "solusi (solution)" and "pemerintah (government)" are words that often appear in neutral sentiment tweets. This neutral sentiment word is followed by other words such as:

- yes or no about national exams, the learning is still based on the curriculum;
- eliminated or not, it's the same case for me;
- eliminate or not, the most important thing is the government more focuses on the quality of human resources;
- The national exam does not guarantee a person's competence, but eliminating the national exam also risks the standards that must be achieved by Indonesian education.

All of these tweets show impartiality to either the pro or con options.

However, from the tweets, it can also be seen that apart from giving a neutral opinion, the public wants the government to focus on improving the quality of education by finding alternative solutions to the abolition of the national exam.

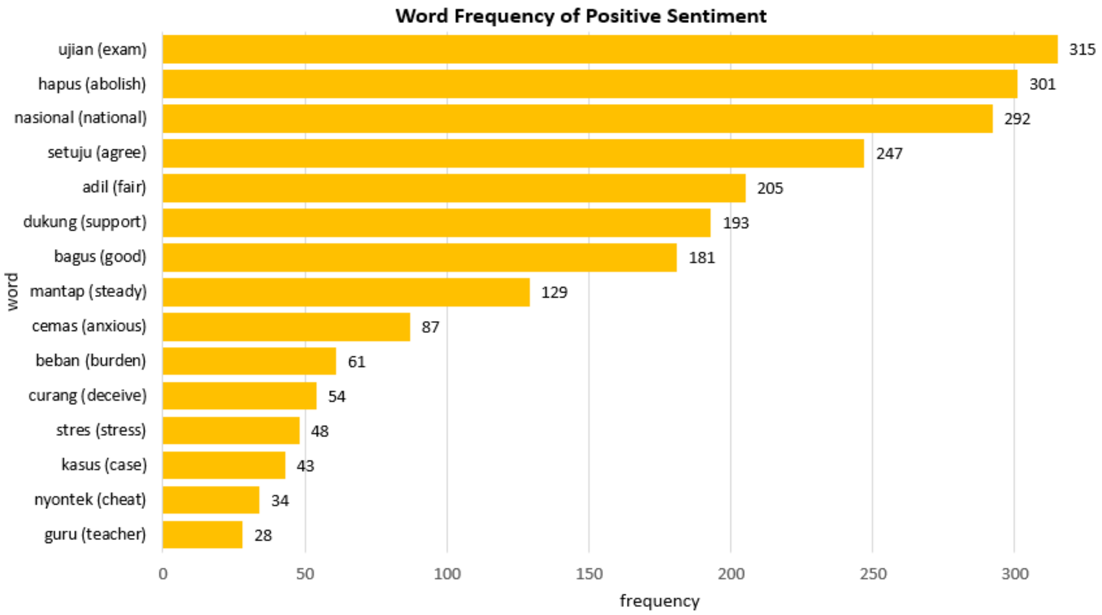


Fig. 6. Frequency of positive sentiment words

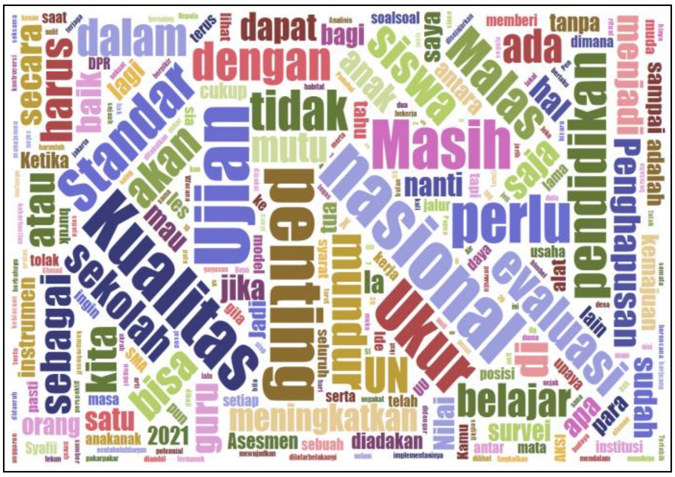


Fig. 7. Word cloud sentiment negative

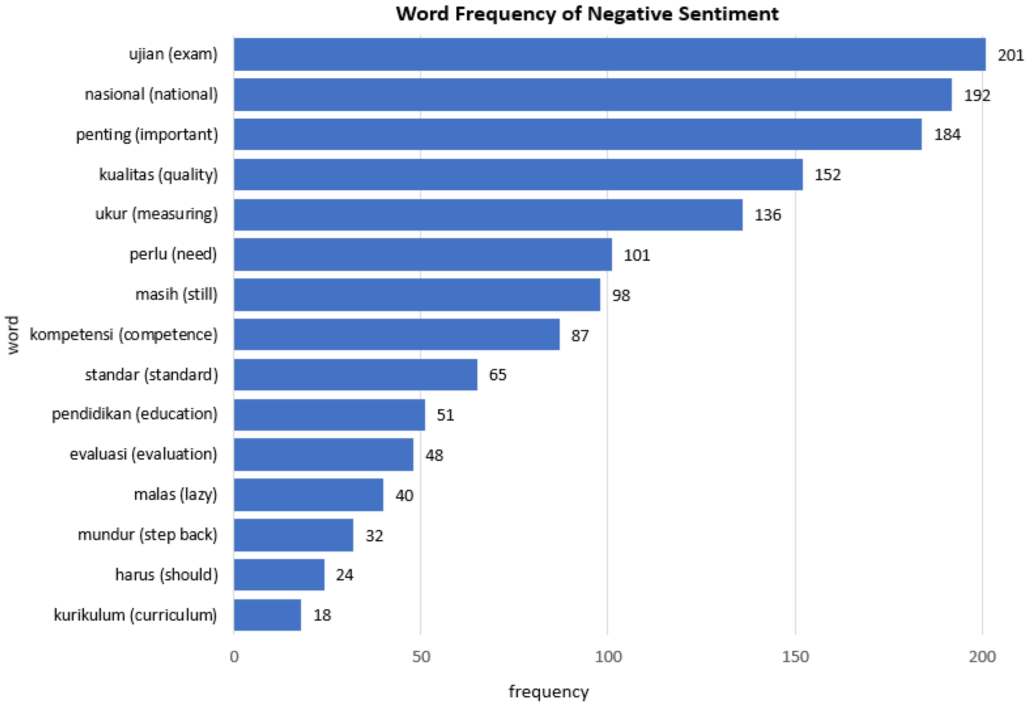


Fig. 8. Frequency of negative sentiment words

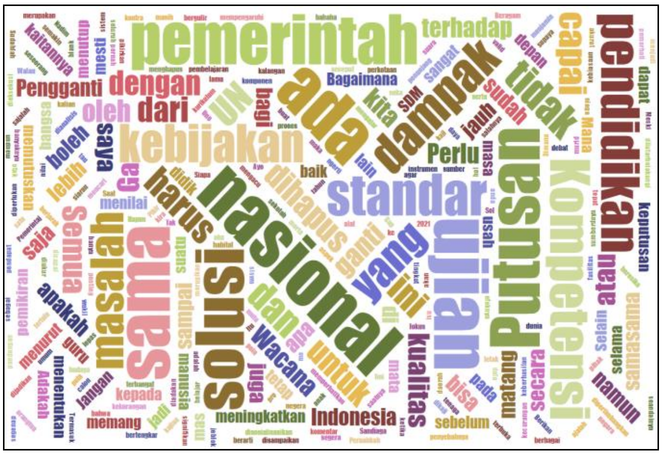


Fig. 9. Word cloud sentiment neutral

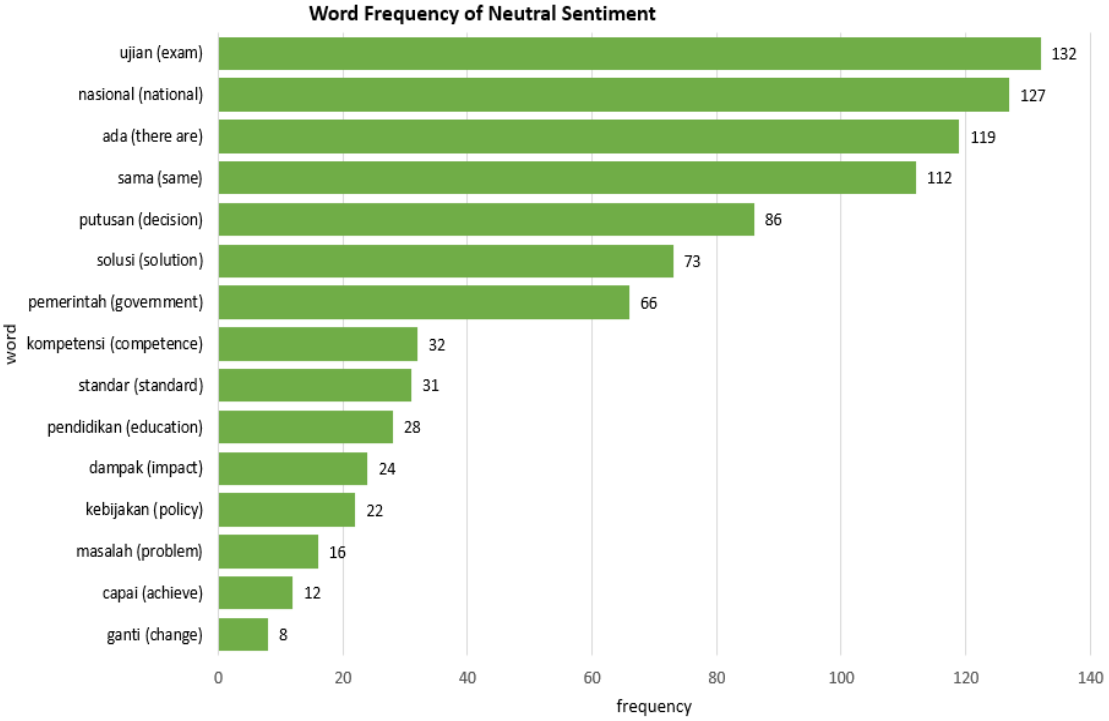


Fig. 10. Frequency of neutral sentiment words

VI. Conclusion

Indonesia has used national exams to determine student graduation in education units for decades. However, the Minister of Education and Culture currently issued a circular to abolish the national exam and replace it with a minimum competency assessment and character survey. This policy has received serious attention from various social circles, with some agreeing to its abolition. The increase in the number of Twitter opinions discussing eliminating the national exam indicates the importance of this policy. Sentiment analysis using a Support Vector Machine was conducted to determine public perceptions of this policy. The comparison analysis of the Support Vector Machine with Random Forest and Logistic Regression proves that the Support Vector Machine performs better in terms of accuracy, precision, recall, and F1 Score.

Support Vector Machine using TF-IDF as feature extraction increases the accuracy of the resulting model compared to BoW. By all counts, with proven results, the Support Vector Machine was considered a polynomial kernel with good machine learning for text classifiers in sentiment analysis. Support Vector Machine can also classify public opinion into ternary sentiment analysis with high accuracy. Therefore, through the experimental result, this study demonstrated that most people agreed to eliminate the national exam in Indonesia. The government uses this study to develop other policies related to changing the approach to the national exam. Future studies will be developed by conducting further experiments using deep learning methods in a broader dataset. There will also be an experiment to integrate the proposed method with other classifiers as a new solution to improve Support Vector Machine performance on sentiment classification.

References

[1] L. T. Dzulfikar, “Commentary: Indonesia seeks to abolish national exams but could end up creating a new rat race,” *Channel NewsAsia*, Jakarta, 27-Jan-2020.

[2] P. G. Bhwana, “Nadiem Makarim Announces No More National Exam per 2021,” *TEMPO.CO*, Jakarta, Dec-2019.

[3] L. Afifa, “Jusuf Kalla: National Exam Removal May Harm Education Quality,” *TEMPO.CO*, Jakarta, Mar-2019.

[4] L. T. Dzulfikar, “PISA-inspired tests will replace Indonesia’s national exams in 2021: how should they be implemented?,” *The Conversation*, Jakarta, 17-Jan-2020.

[5] B. Liu, *Sentiment analysis: Mining opinions, sentiments, and emotions*. New York, USA: Cambridge University Press, 2015.

[6] V. A. Kharde and S. S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques," *International Journal of Computer Applications*, vol. 139, no. 11, pp. 5–15, 2016.

[7] A. Mittal and S. Patidar, "Sentiment analysis on twitter data: A survey," in *ACM International Conference Proceeding Series*, 2019, pp. 91–95.

[8] A. Ligthart, C. Catal, and B. Tekinerdogan, "Systematic reviews in sentiment analysis: a tertiary study," *Artificial Intelligence Review*, 2021.

[9] R. Primartha, B. Adhi Tama, A. Arliansyah, and K. Januar Miraswan, "Decision tree combined with pso-based feature selection for sentiment analysis," *Journal of Physics: Conference Series*, vol. 1196, no. 1, 2019.

[10] A. R. Isnain, N. S. Marga, and D. Alita, "Sentiment Analysis Of Government Policy On Corona Case Using Naive Bayes Algorithm," *IJCCS Indonesian Journal of Computing and Cybernetics Systems*, vol. 15, no. 1, p. 55-64, 2021.

[11] C. C. Le, P. W. C. Prasad, A. Alsadoon, L. Pham, and A. Elchouemi, "Text classification: Naïve bayes classifier with sentiment Lexicon," *IAENG International Journal of Computer Science*, vol. 46, no. 2, pp. 141–148, 2019.

[12] V. A. Fitri, R. Andreswari, and M. A. Hasibuan, "Sentiment analysis of social media Twitter with case of Anti-LGBT campaign in Indonesia using Naïve Bayes, decision tree, and random forest algorithm," *Procedia Computer Science*, vol. 161, pp. 765–772, 2019.

[13] A. C. Lorena *et al.*, "Comparing machine learning classifiers in potential distribution modeling," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5268–5275, 2011.

[14] M. Ghosh and G. Sanyal, "Performance Assessment of Multiple Classifiers Based on Ensemble Feature Selection Scheme for Sentiment Analysis," *Applied Computational Intelligence and Soft Computing*, vol. 2018, 2018.

[15] P. H. Prastyo, A. S. Sumi, A. W. Dian, and A. E. Permanasari, "Tweets Responding to the Indonesian Government's Handling of COVID-19: Sentiment Analysis Using SVM with Normalized Poly Kernel," *Journal of Information Systems Engineering and Business Intelligence*, vol. 6, no. 2, p. 112, 2020.

[16] T. Hariguna, W. M. Baihaqi, and A. Nurwanti, "Sentiment Analysis of Product Reviews as A Customer Recommendation Using the Naive Bayes Classifier Algorithm," *IJIIS International Journal of Informatics and Information Systems*, vol. 2, no. 2, pp. 48–55, 2019.

[17] Stephenie, B. Warsito, and A. Prahutama, "Sentiment Analysis on Tokopedia Product Online Reviews Using Random Forest Method," *E3S Web of Conferences*, vol. 202, pp. 1–10, 2020.

[18] A. A. Lutfi, A. E. Permanasari, and S. Fauziati, "Sentiment Analysis in the Sales Review of Indonesian Marketplace by Utilizing Support Vector Machine," *Journal of Information Systems Engineering and Business Intelligence*, vol. 4, no. 1, pp. 57–64, 2018.

[19] X. Jiawei and T. Murata, "Stock market trend prediction with sentiment analysis based on LSTM neural network," *Lecture Notes in Engineering and Computer Science*, vol. 2239, pp. 475–479, 2019.

[20] M. S. Mubarok, A. Adiwijaya, and M. D. Aldhi, "Aspect-based sentiment analysis to review products using Naïve Bayes," in *International Conference on Mathematics: Pure, Applied and Computation AIP Conference Proceedings*, vol. 1867, August 2017.

[21] S. Wladislav, Z. Johannes, W. Christian, K. Andre, and F. Madjid, "Sentilyzer: Aspect-oriented sentiment analysis of product reviews," *Proc. - 2018 International Conference on Computational Science and Computational Intelligence CSCI 2018*, pp. 270–273, 2018.

[22] Mamtesh and S. Mehla, "Sentiment Analysis of Movie Reviews using Machine Learning Classifiers," *International Journal of Computer Applications*, vol. 182, no. 50, pp. 25–28, 2019.

[23] G. S. Brar and Ankit Sharma, "Sentiment Analysis of Movie Review Using Supervised Machine Learning Techniques," *International Journal of Engineering and Technology*, vol. 7, no. 13, pp. 12788–12791, 2018.

[24] Y. Nurdiansyah, S. Bukhori, and R. Hidayat, "Sentiment analysis system for movie review in Bahasa Indonesia using naive bayes classifier method," *Journal of Physics: Conference Series*, vol. 1008, no. 1, 2018.

[25] N. O. F. Daeli and Adiwijaya, "Sentiment Analysis on Movie Reviews Using Information Gain and K-Nearest Neighbor," in *Journal of Data Science and Its Applications*, vol. 3, no. 1, pp. 1–7, 2020.

[26] M. Ibrahim, O. Abdillah, A. F. Wicaksono, and M. Adriani, "Buzzer Detection and Sentiment Analysis for Predicting Presidential Election Results in a Twitter Nation," *Proc. - 15th IEEE International Conference on Data Mining Workshop ICDMW 2015*, pp. 1348–1353, 2016.

[27] W. Budiharto and M. Meiliana, "Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis," *Journal of Big Data*, vol. 5, no. 1, pp. 1–10, 2018.

[28] U. Yaqub, V. Atluri, S. A. Chun, and J. Vaidya, "Sentiment based Analysis of Tweets during the US Presidential Elections," *ACM International Conference Proceeding Series*, vol. Part F1282, pp. 1–10, 2017.

[29] N. B. Al Barghuthi and H. E. Said, "Sentiment Analysis on Predicting Presidential Election: Twitter Used Case," in *Communications in Computer and Information Science*, vol. 1187 CCIS, pp. 105–117, 2020.

[30] M. Khalid, I. Ashraf, A. Mehmood, S. Ullah, M. Ahmad, and G. S. Choi, "GBSVM: Sentiment classification from unstructured reviews using ensemble classifier," *Applied Sciences*, vol. 10, no. 8, pp. 1–20, 2020.

[31] G. Sun, Z. Wang, Z. Ding, and J. Zhao, "An Ensemble Classification Algorithm for Short Text Data Stream with Concept Drifts," *IAENG International Journal of Computer Science*, vol. 48, no. 4, pp. 1056–1061, 2021.

[32] M. Umer, I. Ashraf, A. Mehmood, S. Ullah, and G. S. Choi, "Predicting numeric ratings for Google apps using text features and ensemble learning," *ETRI Journal*, vol. 43, no. 1, pp. 95–108, 2021.

[33] J. R. Chang, L. S. Chen, and L. W. Lin, "A Novel Cluster based Over-sampling Approach for Classifying Imbalanced Sentiment Data," *IAENG International Journal of Computer Science*, vol. 48, no. 4, pp. 1118-1128, 2021.

[34] C. Colón-Ruiz and I. Segura-Bedmar, "Comparing deep learning architectures for sentiment analysis on drug reviews," *Journal of Biomedical Informatics*, vol. 110, pp. 1-11, 2020.

[35] K. Pasupa and T. S. N. Ayutthaya, "Thai sentiment analysis with deep learning techniques: A comparative study based on word embedding, POS-tag, and sentic features," *Sustainable Cities and Society*, vol. 50, pp. 1–14, 2019.

[36] M. Umer, I. Ashraf, A. Mehmood, S. Kumari, S. Ullah, and G. Sang Choi, "Sentiment analysis of tweets using a unified convolutional neural network-long short-term memory network model," *Computational Intelligence*, vol. 37, no. 1, pp. 409–434, 2021.

[37] C. Zuheros, E. Martínez-Cámara, E. Herrera-Viedma, and F. Herrera, "Sentiment Analysis based Multi-Person Multi-criteria Decision Making methodology using natural language processing and deep learning for smarter decision aid. Case study of restaurant choice using TripAdvisor reviews," *Information Fusion*, vol. 68, pp. 22–36, 2021.

[38] F. Huang, X. Zhang, Z. Zhao, J. Xu, and Z. Li, "Image–text sentiment analysis via deep multimodal attentive fusion," *Knowledge-Based Systems*, vol. 167, pp. 26–37, 2019.

[39] W. Sharif, N. A. Samsudin, M. M. Deris, and R. Naseem, "Effect of negation in sentiment analysis," in *The 6th International Conference on Innovative Computing Technology, INTECH 2016*, 2016, pp. 718–723.

[40] J. D. Kinyua, C. Mutigwe, D. J. Cushing, and M. Poggi, "An analysis of the impact of President Trump's tweets on the DJIA and S&P 500 using machine learning and sentiment analysis," *Journal of Behavioral and Experimental Finance*, vol. 29, pp. 1–14, 2021.

[41] P. Sharma and A. K. Sharma, "Experimental investigation of automated system for twitter sentiment analysis to predict the public emotions using machine learning algorithms," in *Materials Today: Proceedings*, pp. 1-9, 2020.

[42] F. Rustam, I. Ashraf, A. Mehmood, S. Ullah, and G. S. Choi, "Tweets classification on the base of sentiments for US airline companies," *Entropy*, vol. 21, no. 11, pp. 1–22, 2019.

[43] Erlin, U. Rio, and Rahmiati, "Text message categorization of collaborative learning skills in online discussion using support vector machine," in *Proceeding - 2013 International Conference on Computer, Control, Informatics and Its Applications: "Recent Challenges in Computer, Control and Informatics", IC3INA 2013*, 2013, pp. 295–300.

[44] Erlin, U. Rio, and Rahmiati, "Two Text Classifiers in Online Discussion : Support Vector Machine vs Back-Propagation Neural Network," *Telkomnika*, vol. 12, no. 1, pp. 189–200, 2014.

[45] Z. Liu and H. Xu, "Kernel parameter selection for support vector machine classification," *Journal of Algorithms and Computational Technology*, vol. 8, no. 2, pp. 163–177, 2014.

[46] L. Muflikhah, D. J. Haryanto, A. A. Soebroto, and E. Santoso, "High Performance of Polynomial Kernel at SVM Algorithm for Sentiment Analysis," *Journal of Information Technology and Computer Science*, vol. 3, no. 2, pp. 194–201, 2018.