JSDS: JOURNAL OF STATISTICS AND DATA SCIENCE
VOLUME 2, No 2, Oct 2023
e-ISSN: 2828-9986
https://ejournal.unib.ac.id/index.php/jsds/index

Sentiment Analysis of Twitter User's Perceptions of the Campus Merdeka Using Naïve Bayes Classifier and Support Vector Machine Methods

Intan Salsabilla [1*], Muhammad Arib Alwansyah [2], Sigit Nugroho [3], Winalia Agwil [4]
[1234] Statistics Study Program, Bengkulu University, Bengkulu

* Corresponding Author: intansalsabilla27@gmail.com

Article Info

Article History:
Received: October, 21 2023
Accepted: 08 23 2024
Available Online: November, 4 2024

Key Words:
Campus Merdeka
Sentiment Analysis
Naïve Bayes Classifier
Support Vector Machine

Abstract

The Campus Merdeka program is being implemented by the government to realize autonomous and flexible learning in tertiary institutions to create a learning culture that is innovative, not restrictive, and the needs of students. The Campus Merdeka provides added value and is attractive and provides various responses from the public both directly and on different social media platforms. One of the social media platforms is Twitter. Therefore, research was conducted on the community's response to the Campus Merdeka program on Twitter social media. Twitter documents in the form of community response tweets to the Campus Merdeka program are classified into two categories, namely positive responses and negative responses. The method used in this study is the Naïve Bayes Classifier (NBC) and Support Vector Machine (SVM) with a Polynomial Degree 2 kernel. The highest level of accuracy resulting from this research is 73.5% with a parameter value of λ of 0.5, a constant value γ is 0.5, with training data of 309 documents for training data and 132 documents for test data. The accuracy results obtained for the Naïve Bayes Classifier method are 65.9% and for the Support Vector Machine method, an accuracy is 73.5%.

1. INTRODUCTION

Education is all learning experiences that take place in all environments and throughout life. Education is all life situations that influence individual growth. Lifelong education means that education is part of one's life. Learning experiences can take place in all environments and throughout life. According to the Big Indonesian Dictionary, education is the process of changing the attitudes and behavior of a person or group to mature humans through teaching or training.

Based on the definition of education, education in a broad sense is an activity carried out in all situations and obtained throughout life where education can influence growth, provide life lessons, and change attitudes and behavior as a form of improvement or maturation [1]. A Campus Merdeka is a form of learning in higher education that is autonomous and flexible to create a learning culture that is innovative, not restrictive, and to student needs [2].

The Campus Merdeka provides added value and is interesting and provides various responses from the community both directly and on various social media platforms. One social media platform is Twitter. Twitter users' opinions are useful for providing criticism, praise, or suggestions regarding the Campus Merdeka program [3]. Support vector machine is a maximum margin classifier that uses hypotheses in the form of linear functions in a high-dimensional feature based on optimization theory.

This research is also supported by several previous studies regarding sentiment analysis using the naïve Bayes classifier method and the support vector machine method in the form of a journal entitled Sentiment Analysis on Twitter using the naïve Bayes classifier Method written by [4], the results obtained from four trials showed that the accuracy rate in the first trial was 62.98%, the second trial was 64.95%, the third trial was 66.36%, and the fourth trial was 66.79%. From the classification results, the percentage level of positive sentiment was 28%, negative sentiment was 20% and neutral sentiment was 52%. Meanwhile, in research [5] in a journal entitled Comparison of naïve Bayes and support vector machine methods in Twitter sentiment analysis. The results obtained from comparing the two methods show that Naïve Bayes obtained better accuracy results than the support vector machine method with an accuracy of 73.65%.

1.1 Sentiment Analysis

According to the Big Indonesian Dictionary, sentiment is an opinion or view that is raised on excessive feelings about something. Sentiment analysis or what can also be called opinion mining is the process of understanding, extracting, and processing textual data automatically to obtain sentiment information contained in a sentence, whether the opinion is positive or negative [6]. Sentiment analysis is used to determine the responses and attitudes of a group or individual toward a contextual topic of discussion throughout the document. These responses and attitudes can take the form of opinions, judgments, evaluations, affective states, or emotional communication [7].

1.2 Text Mining

Text mining is an activity of classifying documents, clustering, information extraction, sentiment analysis, and information retrieval using a technique where variations of data mining try to find interesting patterns from a set of textual data [8].

According to [9] text mining has the aim of finding words that can represent the contents of the document which can later analyze the relationship between documents defined as data in the form of text which is usually the source of data obtained from documents, so natural language processing for the analysis of unstructured text is generally widely used. Text mining in general has the following stages:

1.2.1Text Preprocessing

The preprocessing stage is the selection of data to be processed where the process in text mining is to convert raw data into structured data. Sentences will be broken down into smaller parts so that they have a narrower meaning. Stages in the preprocessing process are case folding, data cleansing, tokenizing, stop words, and stemming.

1. Case Folding.
   Case folding is because only letters 'a' to 'z' are accepted so all letters in the document must be converted to lowercase.
2. Data Cleansing.
   Data cleansing is used to remove numbers and punctuation marks in sentences because they are considered to not affect text mining, so they need to be removed in each review.
3. Tokenizing.
   Tokenizing cuts words in white space or spaces where tokenizing cuts character sequences in the form of paragraphs into sentences and words (tokens).
4. Stopword
   Stopwords are words that are less important and not descriptive of a document so that word removal can be done. Examples are "which", "in", "from", "and", "I", "you" and so on [10].
5. Stemming
   Stemming is a process to get the root/stem or base word of a word in a sentence by separating each word from the base word and its affixes both prefixes and suffixes [11].

1.2.2 Feature Selection

Feature selection is a stage of the process that is especially useful in reducing data dimensionality, removing irrelevant data, and improving accuracy results. Feature selection has two main goals. First, it makes the training data used for classification more efficient by reducing the size of the effective vocabulary. Secondly, feature selection can usually improve classification accuracy by removing noise features.

Feature selection itself is generally divided into two methods, namely unsupervised feature selection and supervised feature selection.

a. Unsupervised feature selection is a feature selection method that does not use class information in the training data when selecting features for the classifier. Examples of unsupervised feature selection are term frequency and inverse document frequency.
b. Supervised feature selection is a feature selection method that uses class information in the training data, so to use this feature selection a pre-classified set must be available. Examples of supervised feature selection are mutual information and NGram.

Feature selection is used to characterize the data. Feature selection is one of the most widely conducted research in various fields such as pattern recognition, process identification, and time series modeling.

1.3 TF-IDF Weighting

Weighting is a method to convert input data into a feature vector. A commonly used weighting method is bag-of-feature. Each word has a different level of importance in the document, so each word is given an indicator, namely term weight [12].

Weighting uses the calculation of term frequency $tf_{t,d}$ where t is a term in document d which is useful for showing the number of occurrences of term t in document d so that the value of tf will be calculated using the weighting term frequency (W_{tf}) formula in the following equation [12].

$$W_{tf_{t,d}} = \begin{cases} 1 + log_{10} tf_{t,d}, & if \ \ tf_{t,d} > 0 \\ 0, & if \ \ tf_{t,d} = 0 \end{cases} \quad (1)$$

The number of words displayed in a document in tf weighting is usually the term frequency value. This causes weighting on non-essential words, therefore inverse document frequency weighting is used to avoid weighting non-essential words. The inverse document frequency calculation is shown using the following equation:

$$idf_t = log_{10} \frac{N}{df_t} \quad (2)$$

Where N is the number of whole documents in the collections while df_t is that of documents containing term.

The calculation of TF-IDF weighting is the multiplication of term frequency weighting with inverse document frequency. This is shown in the following equation formula:

$$W_{t,d} = W_{tf_{t,d}} \times idf_t \quad (3)$$

1.4 Classification

Classification is a job that assesses a data object to enter a certain class from several available classes [13].

1.4.1 Naïve Bayes Algorithm

Naïve Bayes is a simple probabilistic-based classification technique based on the application of Bayes' theorem (Bayes' rule) with the assumption of strong independence. In Bayes (Naïve Bayes theorem), the meaning of strong independence on features is that a feature on the data is not related to the presence or absence of other features in the same data [13].

Naïve Bayes classification method for classifying text data using the concept of opportunity in determining document classes. This method uses the assumption that in a document the appearance of a word does not affect the appearance of other words and the non-appearance of a word does not affect the non-appearance of other words [14].

1.4.2 Support Vector Machine Algorithm

The basic principle of SVM is a linear classifier and further developed to work on non-linear problems with the concept of kernel trick in high-dimensional workspace. This has increased research interest in the field of pattern recognition to determine the potential capabilities of SVM theoretically and application. Currently, SVMs have been successfully applied in real-world problems, and generally provide better solutions than conventional methods such as artificial neural networks. Support vector machine is also a selection method that compares standard parameters to a set of discrete values called a candidate set, and takes the one that has the best classification accuracy [15]. In its application to text mining, the support vector machine algorithm performs the classification process with binary classification, so that comments in the form of text are first converted to binary by assuming that if the feature is contained in the list of unique features in the training data, it will be worth 1 and vice versa if the feature does not exist it will be worth 0.

The concept of a support vector machine can be explained simply as an effort to find the best hyperplane that functions as a separator of two classes in the input space. The support vector machine process in finding the hyperplane can be seen in Figure 2.1 [16].

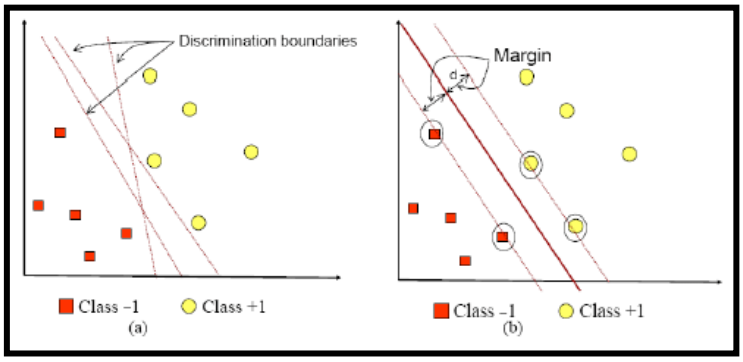

Figure 1. SVM Process in Finding Hyperplane

Figure 1 shows several patterns that are members of two classes (+1 and –1). Patterns belonging to class –1 are symbolized in red (squares), while patterns in class +1 are symbolized in yellow (circles). The classification problem can be interpreted by trying to find the hyperplane (line) that separates the two groups. Various alternative dividing lines (discrimination boundaries) are shown in Figure 1.

The best-separating hyperplane between the two classes can be found by measuring the margin of the hyperplane and finding the maximum point. Margin is the distance between the hyperplane and the closest pattern from each class. This closest pattern is called the support vector.

1.5 Model Validation

Confusion matrix is one of the important tools in the evaluation method used in machine learning which usually contains two or more categories [17]. The classification process is carried out by creating a confusion matrix to determine the level of accuracy. This matrix is used to evaluate the performance of the model formed by each classification algorithm. In evaluating the model, five dataset trials were conducted to get the best accuracy value [18]. To calculate the accuracy value, use the equation below [19].

$$Accuracy = \frac{TP + FN}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + FN} \tag{4}$$

The result of binary classification on a dataset can be represented with a 2 x 2 matrix called a confusion matrix.

Table 1. Confusion matrix

Class	Class 1 (predictions)	Class 2 (predictions)
Class 1 (Actual)	*TP (True Positive)*	*FN (False Negative)*
Class 2 (Actual)	*FP (False Positive)*	*TN (True Negative)*

1.6 Visualization

The visualization stage is used to find out the most words from each positive, negative, and neutral comment. A system called word cloud is used to visualize words using the frequency of words used in positive, negative, and neutral comments. Visualization using the word cloud used in each sentiment class is used to obtain information that can provide solutions to problems from each sentiment class.

1.7 Campus Merdeka

According to the official website of the Ministry of Education, culture, research, and Technology, a Campus Merdeka is a policy of the Minister of Education and Culture, which aims to encourage students to master various sciences that are useful for entering the world of work. This Campus Merdeka policy is a continuation of the concept of independent learning. Its implementation is most likely to be carried out immediately, only changing ministerial regulations, not changing Government Regulations or Laws.

1.8 Twitter

Twitter is a website owned and operated by Twitter, Inc. that offers a social network in the form of a microblog. It is called a microblog because the site allows users to send and read blog-like messages but is limited to 140 characters displayed on the user's profile page. Twitter has unique characteristics and writing formats with special symbols or rules. Messages on Twitter are known as posts. Twitter is also an example of social media. Social media is defined as online media where users can utilize online applications to convey messages in the form of writings, images, and videos.

1.9 Data Crawling

Data crawling is a stage in research that aims to collect or download data from the database. Data collection from this research is data downloaded from the Twitter server in the form of users and posts along with their attributes. Registration as a Twitter application developer to use the Twitter Application Programming Interface can be done at https://dev.twitter.com. After registering the developer will get a consumer key, access consumer, access token, and access token secret that will be used as authentication requirements for the application we are building. Crawling is very useful in the era of big data. Data from social media has a large volume, high speed, and diverse variations, making it easier for users to access the data using the crawling method [20].

2. METHOD COLLECTING DATA

The method used to collect data from this research is web crawling from Twitter user review data on the Campus Merdeka program provided by Twitter users using the RapidMiner application program. The RapidMiner application program provides special tools that can be used to collect Twitter user review data the steps for doing web crawling are in the RapidMiner application program using the web crawling method carried out in the following way:

1. Open and run the RapidMiner application program.
2. Create a new worksheet by selecting "Blank".
3. In the operators panel click and select "Search Twitter" and drag it to the center of the process page.
4. Click the "Search Twitter" panel to bring up the process page then in the parameters section click the Twitter symbol.
5. Click "add connection".
6. Create a name and click the "create" button. Select the name that has been created and click the box icon on the process page.
7. The box icon serves to get a token which is permission from Twitter to retrieve data from Twitter. Then, in OAuth, click the "Request Access Token" button.
8. login with a Twitter account, then press the "Authorize App" button, you will get a code copy of the code that has been given paste it on the "copy code" form then click "Complete" then click Test to see if it works or not, if there is a green checklist, then click "Save All Changes".
9. The Query column is the topic that will be crawled, namely the Campus Merdeka, and then the limit content is the limit that will be searched with the number 10000. Language is the language you want to filter and fill with ID.
10. Operator column, type "write csv" to save the file in csv format.
11. Connect all operators and run. Then the crawling result file has been obtained.

3. RESULTS AND DISCUSSION

3.1 Description of Research Data

The research data used is data on the public perception of Twitter users of the Campus Merdeka program from July 01, 2022, to August 31, 2022, with the crawling process and using the keyword Campus Merdeka by Twitter users and obtained 12,322 reviews. The components retrieved are tweet creation time, username, user ID which is a special ID owned by a Twitter user that is not owned by other users, To-User or an account that reposts a tweet

(retweet), retweeted user ID, language used, source, text, tweet sending location information, number of sentences retweeted and id in the form of a location code in the region where the review was sent.

	Created-At	From-User	From-User-Id	To-User	To-User-Id	Language	Source	Text	Geo-Local	Geo-Location-Lo	Retweet-Count	Id
2	07.07.2022	COLLE	1046084008742801408		-1	in	<a href="h	[cm] haii, mau nanya soal kampus merdeka. Biasanya dari			,0	1545059784407531520
3	07.07.2022	Mii	1221080191159562240		-1	in	<a href="h	RT @UGM_FESS: ugm_fess [Review Magang Kampus Merd			28,0	1545058320884502531
4	07.07.2022	hiidin	1215960266006969	erstevn	855382858954	in	<a href="h	@erstevn @collegemenfess Kalo magang kampus merdek			,0	1545056680118890497
5	07.07.2022	COLLE	1046084008742801408		-1	in	<a href="h	[cm] gais ini kampus merdeka Studi Independent sm Maga			,0	1545056379194740736
6	07.07.2022	˚₊	1048135715798601	peachyfru	1452131767693	in	<a href="h	@peachyfruiitz pertukaran mahasiswa merdeka kakk, salal			,0	1545055689025802240
7	07.07.2022	COLLE	1046084008742801408		-1	in	<a href="h	[cm] kampus merdeka semua magang min. semester 5?			,0	1545053843578580992
8	07.07.2022	Unpadfess --SF	1209457724599095296		-1	in	<a href="h	- kalo prodinya ga masuk di list web kampus merdeka tuh g			,0	1545053685121576961
9	07.07.2022	mirza • looking	1329959205147676	collegeme	1046084008742	in	<a href="h	@collegemenfess sejauh ini masih nonton series sama filr			,0	1545051817238704136
10	07.07.2022	Little buddy™	1436744519450419	collegeme	1046084008742	in	<a href="h	@collegemenfess Tiap semester ada.Siap siap aja di bulan			,0	1545049485935788032
11	07.07.2022	COLLE	1046084008742801408		-1	in	<a href="h	[cm] ges program kampus merdeka tuh ada tiap semester			,0	1545048826729611264
12	07.07.2022	COLLE	1046084008742801408		-1	in	<a href="h	[cm] guys mau tanya, ini program kampus merdeka yg stud			,0	1545047552579358728
13	07.07.2022	UNS MENFESS	2340043838		-1	in	<a href="h	yg daftar mbkm kampus merdeka kalo udah daftar trus cek			,0	1545045574743117824
14	07.07.2022	sam priv??	1371144860770603011		-1	in	<a href="h	ya allah bismillahirrahmanirrahim semoga besok baim dap			,0	1545041940580040704
15	07.07.2022	Oong	1357189584782643	sbmptnfe	1205629250423	in	<a href="h	@sbmptnfess Kedua aja nderr, Kalo pertama Beratt Isiny U			,0	1545041017237569537
16	07.07.2022	yayas	784231927		-1	in	<a href="h	capek banget menjadi anak uin karna ga bisa ikut magang k			,0	1545040821665529856
17	07.07.2022	COLLE	1046084008742801408		-1	in	<a href="h	[cm] guys aku smt 5, kalo mau ikut studi/beasiswa kampus			,0	1545040037334876165
18	07.07.2022	COLLE	1046084008742801408		-1	in	<a href="h	[cm] aku otw smt 5, pengen ikut studi independen kampus			,0	1545039354329853956

Figure 2. Campus Merdeka Review Data

3.2 Classification

Before entering the classification process, review information that is still not well structured and not uniform needs to be preprocessed which will then be carried out word weighting using the TF-IDF method so that the sentiment analysis process can be carried out using the naïve Bayes classifier method and support vector machine.

3.3 Manual Calculation

Manual calculations in this study are used as an overview of the system design. This study uses the SVM method with a Polynomial Degree 2 kernel. The following is a manual of each process.

3.3.1 Data Set

The dataset that will be used in the manual calculation will be divided into two, namely training data and test data. The data set used in the manual process is 12 documents. The document consists of 10 documents used as training data with 5 positive sentiment documents and 5 negative sentiment documents while for test data 2 documents are used that do not yet have sentiment labels. Table 2 describes the training data that will be used in the manual process while table 3 describes the test data that will be used in the manual process. The dataset taken was selected to only use Indonesian.

Table 2. Training Data

No	Training Data
1	merasakan langsung proses belajar mengajar https://t.co/VVz2Xmm3zf
2	Kampus merdeka yang sekarang bagus bagus jir
3	di kampus merdeka programnya seruu bahkan online kegiatannya variatif Pesta startup dari ekspert FGD workshop Seru seru deehh apalagi dapet uang saku juga dan konversi 20sks juga
4	Konversi nilai sulit, laporan sulit, uang saku sulit, program macam apa sih itu kampus merdeka
5	Kampus merdeka tapi mahasiswa nya wajib ikut dan ga merdeka nentuin pilihannya. Giliran ikut malah merugikan, konversinya ga pas.
6	Mata kuliah di kampus merdeka nilainya tidak realistis
7	alhamdulillah dapat bantuan dari kampus merdeka
8	Kampus Merdeka gak jelas.
9	Jujur aku sedih dan kesel banget se, grgr departemen ku awikwok banget Departemen lainnya, kalo ikut kampus merdeka dapet konversi. Departemenku aja yg engga Kek anjir Di hubungi malah ngegas, gjls Ewhhhhh
10	Enak kok kak aku udh magang di kampus Merdeka

Table 3. Test Data

No	Test Data
1	@fleurpurple program kampus merdeka si bantu banget
2	@collegemenfess susah si, nder. aku yang ikut program kampus merdeka yang lain aja keteteran waktunya, padahal online semua

3.3.2 Preprocessing

In this stage of writing, a pre-processing process is carried out which includes case folding, cleansing, tokenizing, stop word, and stemming processes which will be described as follows.

1. Case Folding.
   Case folding is useful for converting capital letters into lowercase letters. For example, changing the letter 'K' to 'k' in textual data.
2. Data Cleansing
   Data Cleansing is the process of cleaning unnecessary words to reduce noise. The words that are removed are URL, hashtag (#), username (@username), email, punctuation (!@#$%^&*()_+=.,<>?/|{}[]1234567890) and correcting words that are not standardized into standardized words.
3. Tokenizing.
   Tokenizing is cutting the sequence of characters in the form of paragraphs into sentences and words. Tokenizing is useful to simplify the calculation of the frequency of occurrence of words in documents.
4. Stop word Removal
   Stop words are words that are not unique characteristics (words) of a document. Examples of stop words are he, they, me, at, and so on. Before the stop word removal process is carried out, a stop word list must be made (stop list) this stop list contains common words, connecting words, personal pronouns, and not unique words. The stop list in this research is obtained directly from the Indonesian stop word document.
5. Stemming
   Stemming is the stage of eliminating affixed words into basic words (root) from each stop word result word by using certain rules. The rules that apply are that the stemming process in Indonesian text is different from the stemming process in English text. In English texts, the only process required is the process of removing suffixes while in Indonesian texts, in addition to suffixes, prefixes, and confides are also removed.

3.4 TF-IDF Manual Calculation

Manual calculations in this study are used as an overview of the system design. In this study using the NBC and SVM methods. The following is a manual of each process. After labeling and pre-processing the text, the next stage is the weighting stage using $tf - idf$, which at this stage will calculate the weight of the word. The $tf - idf$ weighting is only done on the training data. The following is the implementation of $tf - idf$ weighting where in this method the weight of each word in the document will be calculated using the formula:

$$W_{tf_{t,d}} = \begin{cases} 1 + log_{10} tf_{t,d}, & if \ \ tf_{t,d} > 0 \\ 0, & if \ \ tf_{t,d} = 0 \end{cases}$$

We will find the value of $W_{tf_{t,d}}$ and idf in each word. An example calculation to find out the value of $W_{tf_{t,d}}$ in the first word can be described as follows:

t (program) = 1
d (document) = 1 (d1)
$W_{tf_{t,d}}$ = 1 + log10 (1)
$W_{tf_{t,d}} = 1$

Meanwhile, to find out the value of idf in each word can be known as follows:

$$idf_t = log_{10}\frac{N}{Df}$$
$$idf_t = log_{10}\frac{10}{3}$$
$$= 0.5229$$

Based on the IDF value that has been obtained, the manual calculation of the $TF - IDF$ value on the first term can be explained as follows:

t (program) = 1
d (document) = 1 (d1)
$TF - IDF = W_{tf_{t,d}} \times idf_t$
$TF - IDF = 1 \times 0.5228$
$TF - IDF = 0.5228$

3.5 Sentiment Classification Manual Calculation

Data that has been labeled and weighted will then be predicted using the naïve Bayes classifier and support vector machine algorithms. Data that has previously been partitioned and labeled will be searched for accuracy values in each classification method to determine the proportion of correct sentiment predictions.

3.5.1 Manual Calculation of Sentiment Classification Naïve Bayes Classifier

In the naïve Bayes classifier classification process, the entire dataset is divided into several stages of training data and test data. The following is an explanation of the steps along with calculation examples:

1. In the training data stage at the stage, each word that has a known TF-IDF weight value will be used as a reference to find the probability value for each class of training data to get a sentiment class. The following is the calculation:

a. First calculate the probability of each category (prior). In this study, there are two categories, namely positive, and negative categories.

$$P\ (\text{positive/negative}) = \frac{D(\text{positive/negative})}{|C|}$$
$$P\ (\text{Pos}) = \frac{d\ (\text{Pos})}{|C|} = \frac{5}{10} = \frac{1}{2} = 0.5$$
$$P\ (\text{Neg}) = \frac{d\ (\text{Neg})}{|C|} = \frac{5}{10} = \frac{1}{2} = 0.5$$

b. After calculating the probability of each category, then calculate the probability of each term from all documents. Many terms depend on the results of data preprocessing. The following is the calculation of the probability of each term:

$$P\ (w_k|\text{positive/negative}) = \frac{(n_k|\text{positive/negative}) + 1}{(n_k, |\text{pos/neg}) + |\text{vocabulary}|}$$

Table 4. Probability of each word

No	word	p(w|pos)	p(w|neg)
1	program	0,025572	0,017305
2	campus	0,0125	0,011364
3	independent	0,0125	0,011364
4	help	0,029974	0,011364
5	students	0,021237	0,019306
6	taste	0,025	0,011364
7	department	0,0125	0,028149

2. The testing stage at this stage is carried out by using test data in the model that has been formed at the training stage above. The process to be carried out is to calculate the probability value based on the value of each term.

The calculation of the probability value is done by multiplying the probability value of all words in each category by the probability of each term taken from all data.

Table 5. Probability value of test data 1

No	Test Data 1	Positive	Negative
1	program	0,025571969	0,01730544
2	campus	0,0125	0,011363636
3	independent	0,0125	0,011363636
4	Help	0,02997425	0,011363636
5	Probability of test data 1	$5{,}98829 \times 10^{-08}$	$1{,}26971 \times 10^{-08}$

Table 6. Probability value of test data 2

No	Test Data 2	Positive	Negative
1	difficult	0,0125	0,026148068
2	follow	0,0125	0,029640352
3	program	0,025571969	0,01730544
4	campus	0,0125	0,011363636
5	independent	0,0125	0,011363636
6	online	0,025	0,011363636
7	Probability of test data 2	$7{,}80395 \times 10^{-12}$	$9{,}84074 \times 10^{-12}$

The highest probability value on test data 1 is in the positive category with a value of $5{,}98829 \times 10^{-08}$, so it can be classified into the "positive" class. As for the second test data, the highest probability value is in the "negative" class, which is $9{,}84074 \times 10^{-12}$, so the comment can be classified into the "negative" class.

3.5.2 Support Vector Machine

The classification process of the support vector machine method is to find the boundaries between data classes (hyperplane) using sequential training SVM. To find the hyperplane, the only data needed is training data that has a known sentiment class. The stage of sequential learning by initializing the parameters used, such as $a_i, \lambda, \gamma, C, \varepsilon, i_{max}$ and d. Where for the value of $ai = 0,\ \lambda = 0.5,\ \gamma = 0{,}5,\ C = 1,\ \varepsilon = 0.0001,$ and $d = 2$.

1. Initialization $\alpha = 0$ to calculate the hessian matrix D_{ij} where the kernel is needed by using a polynomial kernel by taking into account the degree or d value of 2. So from the kernel multiplication performed with the formula $K(x_i, x_d) = (X_i^T X_j + C)^d$with an example of solving the calculation of the polynomial kernel in the 1st document can be described as follows:

$$(x_i, x_j) = (X_i^T X_j + C)^d$$
$$(x_1, x_1) = \big(((0{,}3802.\,0{,}3802) + \cdots + (0.0) + (0\ .0)\big) + 1)^2$$
$$(x_1, x_1) = 45{,}8476$$

The results of the calculation of the Hessian matrix on the 1st document can be described as follows:

$$D_{ij} = y_i y_j\ (K(x_i\ , x_i)\ +\ \lambda^2)$$
$$D_{11} = y_1 y_1\ (K(x_1\ , x_1)\ +\ \lambda^2)$$
$$D_{11} = 1.1(45{,}85 + 0.5^2)$$
$$D_{11} = 46{,}0976$$

Here are the steps in sequential learning to calculate the value of $E_i, \delta_{\alpha i}$ and α_i:

a. The first step is to calculate the error rate value, for example:

$$E_i = \sum_{j}^{i} a_j D_{ij}$$
$$E_i = (46{,}10 \times 0) + (1{,}25 \times 0) + \cdots + (1{,}25 \times 0)$$
$$E_1 = 0$$

b. The error rate value that has been known will be used in the calculation of the alpha delta value symbolized by $\delta_{\alpha i}$. The manual calculation of $\delta_{\alpha i}$ in the first iteration is explained as follows:

$$\delta_{\alpha i} = \min(\max[\gamma(1 - E_i), \alpha_i], C - \alpha_i)$$
$$\delta_{\alpha 1} = \min(\max[0{,}0001(1 - 0), 0], 1 - 0)$$
$$\delta_{\alpha 1} = \min(\max[0{,}0001{,}0], 1)$$
$$\delta_{\alpha i} = \min(0{,}0001, 1)$$
$$\delta_{\alpha i} = 0{,}0001$$

c. After obtaining the value of $\delta_{\alpha i}$, the value of α_i needs to be updated to be used in the next iteration. A manualized example of the calculation of α_i is described as follows:

$$\alpha_i = \alpha_i + \delta_{\alpha i}$$
$$\alpha_i = 0 + 1$$
$$\alpha_i = 1$$

d. The manual process is initialized the maximum iteration is 2. Then the results of the final calculation process of the value of E, δ_α, and α_i

e. At the sequential learning calculation stage, the final value α_i will be obtained which functions as the support vector (SV) value. Entering the classification stage, it is necessary to find the bias value to find the bias value, the value of $(x^-)$and (x^+)must be known. Both values are obtained from the maximum value of α_i from the positive class $(x^+)$and the maximum value of α_i from the negative class (x^-).

f. The final step in the data classification process with the support vector machine method is to insert the test data that you want to know the sentiment class to the hyperplane that has been found in the previous calculation. The positive or negative value shown in the final result of the calculation $sign\ (f(x))$ will determine the class of the test data. If the calculation value shows -1. Then the test data document is included as a negative class. But if the calculation result shows $+1$. Then the test data document is included as a positive class.

Table 7. Classification calculation results on test data

$\alpha_i . y_i K(x . j)$	u_1	u_2
d_1	0,000454232	0,000262
d_2	0,0002	0,0002
d_3	0,000262004	0,000613
d_4	-0,000262	-0,00141
d_5	-0,0002	-0,00043
d_6	-0,0002	-0,0002
d_7	0,000371268	0,0002
d_8	-0,0002	-0,0002
d_9	-0,0002	-0,00037
d_{10}	0,0002	0,0002
Total	0,0004255	-0,00114
Total+bias	0,000425556	-0,00114
Sentiment	positive	negative

The classification calculation results on test data 1 have a value of 0.000425556 so test data 1 shows the calculation results show +1 or positive value, so test data 1 can be categorized in positive sentiment. As for the second test data, the classification calculation results on test data 2 have a value of -0.00114 so test data 2 shows the calculation results show -1 or a negative value, so test data 2 can be categorized in negative sentiment.

The next stage is the evaluation of classification results using the NBC method which will be carried out against the classification evaluation results using the SVM method in the following confusion matrix values:

Table 8. Classification calculation results on test data

Algorithm	Actual	Prediction	
		Positive	Negative
NBC	Positive	43	21
	Negative	24	44
SVM	Positive	52	18
	Negative	17	45

$$Accuracy\ NBC = \frac{TP + TN}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{F}N} = \frac{43 + 44}{43 + 21 + 44 + 24} = 0{,}659 = 65{,}9\%$$

$$Accuracy\ SVM = \frac{TP + TN}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{F}N} = \frac{52 + 46}{51 + 18 + 17 + 24} = 0{,}735 = 73{,}5\%$$

Based on the accuracy results, it can be seen that the sentiment in the naïve Bayes classifier method is obtained as much as 43 predicted review data correctly classified as positive and 44 predicted data correctly classified as negative with a data prediction accuracy rate of 65.9%. While using the support vector machine as many as 52 data were predicted to be correctly classified as positive and 45 data were correctly predicted as negative with a data prediction accuracy rate of 73.5%. So that the method that works better in classifying is the support vector machine method because the support vector machine method can produce a data accuracy rate of 73.5%.

4. CONCLUSION

Based on the results and discussion that have been presented in the previous chapter, the following conclusions are obtained:

1. Based on sentiment using the labeling results of Twitter user reviews of the Campus Merdeka program obtained from July 01, 2022, to August 31, 2022, it is known that the number of reviews obtained is 12.322 reviews. The 12.322 review data that has been obtained from the crawling process is processed so that the review data is obtained as many as 441 with test data as many as 132 reviews.
2. Based on the results obtained, it can be seen that the sentiment in the naïve Bayes classifier method is obtained as much as 43 positive review data. 44 predicted data are correctly classified as negative with a data prediction accuracy rate of 65.9%. While using the support vector machine as much as 52 data is predicted to be true positive and 45 data is predicted to be negative with a data prediction accuracy rate of 73.5%.
3. The method that works better in classifying is the support vector machine method because the support vector machine method can produce a data accuracy rate of 73.5%.
4. In negative sentiment, problems are found which are then sought for problem-solving from the factors found. These problems include errors during the opening of the Campus Merdeka web which occurred for a long time so that obstacles such as failure to register occurred. There are several campuses where the Campus Merdeka program is not available. There are complaints about the assignments given that are too burdensome for students. There is a long delay in the payment of money and the difficulty of graduating from the Campus Merdeka program. There is unpreparedness in the implementation of the program so students become due to the cut off of student admission to the Campus Merdeka program.

REFERENCES

[1] Gazali, M. (2013). Optimalisasi Peran Lembaga Pendidikan Untuk Mencerdaskan Bangsa. Al-TA'DIB: Jurnal Kajian Ilmu Kependidikan, 6(1), 126-136.

[2] Kementrian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia. 2022. Kampus Merdeka. Jakarta, Kementrian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia. Diaksees pada tanggal 29 Januari 2022 https://kampusmerdeka.kemdikbud.go.id

[3] Araque, O. et al. 2017. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. Expert Systems with Applications. Elsevier Ltd, 77, p.

[4] Suryono, S., & Taufiq Luthfi, E. (2021). Analisis sentimen pada Twitter dengan menggunakan metode Naïve Bayes Classifier. Jnanaloka, 81–86.

[5] Fikri, Mujaddid Izzul, Trifebi Shina Sabrila, and Yufis Azhar. 2020. "Perbandingan Metode Naïve Bayes Dan Support Vector Machine Pada Analisis SentimenTwitter." 10: 71–76.

[6] Sudiantoro, A. V., & Zuliarso, E. 2018. Analisis Sentimen Twitter Menggunakan Text Mining Dengan. 398-401.

[7] Saraswati, N., 2011. Text Mining dengan Metode Naive Bayes Classifier dan Support Vector Machines untuk Sentiment Analysis. Skripsi. Yogyakarta: Program Studi Teknologi Informasi Fakultas Teknik Universitas Gadjah Mada.

[8] Hearst, M., 2003. What Is Text Mining?. SIMS UC Berkeley Publication. [Online] Available at: http://people.ischool.berkeley.edu/~hearst/text-mining.html [Accessed 05 January 2022].

[9] Harlian, M., 2006. Machine Learning Text Kategorization. Austin: University of Texas.

[10] Putri, D.U.K. 2016. Implementasi Inferensi Fuzzy Mamdani Untuk Keperluan Sistem Rekomendasi Berita Berbasis Konten. Skripsi. Program Studi Ilmu Komputer FMIPA UGM Yogyakarta.

[11] Wahyudi, D., Susyanto, T., & Nugroho, D. 2017. Implementasi dan analisis algoritma stemming nazief & adriani dan porter pada dokumen berbahasa indonesia. Jurnal Ilmiah SINUS, 15(2), 49-56.

[12] Akbari, M. I. H. A. D., Astri Novianty S.T., M. & Casi Setianingsih S.T., M., 2012. Analisis Sentimen Menggunakan Metode Learning Vector Quantization. Telkom University.

[13] Prasetyo, Eko. 2012. Data Mining Konsep dan Aplikasi menggunakan Matlab. Yogyakakara. Penerbiy: Andi.

[14] Rianto, Bagus.2016. Implementasi dan Perbandingan Metode Prapemprosesan Pada Analisis Sentimen Gubernur DKI Jakarta Menggunakan Metode Support Vector Machine dan Naïve Bayes. Skripsi. Program Studi Ilmu Komputer FMIPA UGM Yogyakarta.

[15] Dong, Y., Xia, Z., Tu, M., & Xing, G. 2007. An optimization method for selecting parameters in support vector machines. In Sixth International Conference on Machine Learning and Applications (ICMLA 2007) (pp. 1-6). IEEE.

[16] Hasanah, U., Resita M., L., Pratama, A., Cholissodin, I.,2016, Perbandingan Metode SVM, Fuzzy K-NN, dan BDT-SVM untuk Klasifikasi Detak Jantung Hasil Elektrokardiografi, JTIIK, 3(3): 201- 207.

[17] Manning, C. D., Prabhakar, R., dan Hinrich, S. 2009. An Introduction to Information Retrieval – Online Edition. Cambridge: Cambridge University Press.

[18] Putri, B. A. D., 2019. Analisis Sentimen Data Ulasan Pengguna Grab Menggunakan Metode Support Vector Machine dan Maximum Entropy. Skripsi. Program Studi Teknik Industri Fakultas Teknologi Industri UII Yogyakarta.

[19] hum Lim, S. Y., Song, M.H., & Lee, S.J. 2006. Ontology-based automatic classification of web documents. SPringer-Verlag, 690-700.

[20] Zuliarso, E., & Mustofa, K. (2009). Crawling Web berdasarkan Ontology. Dinamik, 14(2).