Jurnal TAM (Technology Acceptance Model)
Jurnal TAM, Volume 13, Nomor 1, Juli 2022
E ISSN: 2579-4221; P ISSN: 2339-1103, pp.51-56
Accredited SINTA 4 Nomor 23/E/KPT/2019
https://ojs.stmikpringsewu.ac.id/index.php/JurnalTam/index

SENTIMENT ANALYSIS FOR EXTRACTING STUDENT OPINION DATA ON HIGHER EDUCATION SERVICES USING THE NAIVE BAYES CLASSIFIER AND SUPPORT VECTOR MACHINE METHODS (CASE STUDY AKPRIND INSTITUTE OF SCIENCE AND TECHNOLOGY YOGYAKARTA)

Uning Lestari[1], Tri Romadhani[2], Suraya[3], Erfanti Fatkhiyah[4]
Informatics Departmen Institut Sains & Teknologi AKPRIND Yogyakarta
[1,2,3,4]Balapan Street No.28, Klitren, Gondokusuman, Yogyakarta City,
Special Region of Yogyakarta, Indonesia
E-Mail: uning@akprind.ac.id, romadhantri@gmail.com, suraya@akprind.ac.id

Article history:
Received: October xx, 2021
Revised: November xx, 2021
Accepted: November xx, 2021

Corresponding authors
*uning@akprind.ac.id

Keywords:
Sentiment Analysis;
Naive Bayes Classifier;
Support Vector Machine;
Information Gain;
Text Processing.

Abstract

Opinions are ideas, opinions, or the results of someone's subjective thoughts in explaining or addressing something. IST AKPRIND Yogyakarta provides comment and suggestion box facilities in the learning evaluation questionnaire. Opinions that have been collected can be used to determine the sentiment of the campus community. This sentiment information can be used in future campus development. The development of a system that can analyze sentiment automatically is designed by comparing the Naive Bayes Classifier (NBC) method and the support vector machine (SVM) optimized by selecting the Information Gain (IG) feature. Prior opinion data needs to be prepared before being analyzed. Preprocessing (text preprocessing) used includes: cleanning, text folding, normalization, stemming, stopword removal, convert negation, and tokenization. The results of this study show that the SVM method produces higher accuracy than NBC. The accuracy test shows the highest accuracy of SVM reaches 99.09% while NBC is 96.56%. The application of IG did not significantly affect the accuracy of the analysis. GI greatly influenced the analysis duration of the SVM method, which could shorten the time by 195.71%.

I. INTRODUCTION

Currently opinion mining or sentiment analysis has become a research topic that is in great demand in the field of text mining. Sentiment analysis aims to create automated tools that can extract subjective information from texts that are natural language such as opinions and sentiments, so as to create structured knowledge that can be used in decision support systems or decision making [1]. Opinion mining is considered as a combination of text mining and natural language processing. Sentiment analysis is a classification process into two tendencies (binary classification), namely positive and negative [2]. The one method of text mining that can be used to solve opinion mining problems is the Naïve Bayes Classifier (NBC). NBC can be used to classify opinions into positive and negative opinions. NBC can work well as a text classifier method [3]. In addition to NBC, the Support Vector Machine (SVM) method is also used in text classification. Standard SVM takes a set of input data, and predicts, for any given input, the probability that the input is a member of one of the classes, so SVM is also a binary linear nonprobabilistic classifier. [4]. The collaboration of the NBC and SVM methods will further improve the accuracy of the text classification results [5]

AKPRIND Institute of Science & Technology Yogyakarta in an effort to improve its services to students always conducts a survey of service assessments at the end of each semester related to public services, facilities and the learning process from lecturers. Collecting student opinions is given through a questionnaire form at the end of each semester. This questionnaire is filled out by students and contains positive, negative, or neutral opinions. The results of this questionnaire can be used as an indicator for assessing the quality of services and facilities at IST AKPRIND Yogyakarta. Students as one of the important aspects of activities on campus, their opinions can have an effect on improving quality.

In this study, a student opinion data processing system was made using Naive Bayes Classifier and Information Gain and Support Vector Machine (SVM). By combining these methods, it is hoped that sentiment analysis can be carried out more quickly, easily, and with a fairly high level of accuracy and effectiveness. The result will be able to know the tendency of students to be positive, negative or neutral. So the results can be used for service improvement and even better performance.

II. Related Work

A. *Text Mining*

Sentiment analysis is fundamentally used to express one's unique opinion. The most recent cutting-edge in conclusion divided classes into two categories: positive and negative. This section describes the literature review on the sentimental analysis, as well as the techniques used on user reviews.

Text mining is a technology used to analyze unstructured data in the form of text data. In text mining analysis there are two main phases, namely (1) Preprocessing and integration of unstructured data, (2) Statistical analysis of data that has been preprocessed to extract content from that contained in the text. [6]. Text mining is a transformation of text data into numeric data so that it is able to convert unstructured data into structured data [7].

Sentiment analysis is a very common field in text classification. Sentiment analysis is a process that analyzes and detects the sentiment of a text input having a positive, negative or neutral sentiment. However, until now, the sentiments that can be detected have become more diverse and not limited to only positive and negative, which can detect happiness, sadness, anger, fear, disgusted and surprised [8]. Sentiment analysis can be used, one of which is to monitor the quality or performance of an institution's products and services so that further conclusions can be drawn whether the service is accepted or not. Research in the field of sentiment analysis using Indonesian text has been carried out for various purposes, for example for service assessment. , prediction, facility assessment and others [9]–[11]. The methods used are varied, ranging from SVM, Naïve Bayes, KNN, to Deep Learning-based methods, such as Convolutional Neural Network (CNN).

B. *Naive Bayes Classifier (NBC)*

The NBC algorithm is often used for text classification problems. As an illustration, for example, training data is categorized into several k categories $C_j=\{C_1,C_2,C_{3,...,}C_k\}$ and prior probability for each category is $p(C_1)$, where j = 12,3,...,k. Data collection is symbolized $d_i=(w_1,...,w_2,...,w_m)$, and words or features that are in the document d_i, made by calculating the probability value of all documents (*posterior probability*). Posterior probability of a document in a category can be calculated by the equation :

$$p(C_j|D_i) = \frac{p(d_i|C_i)p(c_j)}{p(d_i)} \quad (1)$$

In naive bayes classification opinion is represented in attributes $(a_{1,}\ a_{2\,,}\ a_{3\,,}\ ...\ a_n)$, a_1 is the first word, a_2 is the second word, and so on until the last word. V is the set of classes. At the time for classification this method will look for V_{MAP} (category / class with the highest probability value) by enter attibutes $(a_{1,}\ a_{2\,,}\ a_{3\,,}\ ...\ a_n)$ using equation (2)

$$V_{MAP} = {}^{argmax}_{v_j \in V} P(v_j \vee a_1, a_2, a_3, ...a_n) \quad (2)$$

By applying the Bayes method, equation (2) can be written as in equation (3).

$$V_{MAP} = \begin{matrix} argmax \\ v_j \in V \end{matrix} \frac{P(a_1, a_2, a_3, \ldots a_n \vee v_j)P(v_j)}{P(a_1, a_2, a_3, \ldots a_n)} \quad (3)$$

With value $P(a_{1,}\ a_{2\,,}\ a_{3\,,}\ ...\ a_n)$ is constant for each v_j so that equation (3) can be written as equation (4).

$$V_{MAP} = {}^{argmax}_{v_j \in V} P(a_1, a_2, a_3, \ldots a_n \vee v_j)P(v_j) \quad (4)$$

The Naive Bayes classifier simplifies it by assuming that within each category, each attribute is conditionally independent of one another. So it becomes equation (5). $P(v_j)$ and the probabilities of the word ai for each category are calculated during training using formula (5) and formula (6).

$$(v_j) = \frac{docs_j \vee}{training \vee} \quad (5)$$

$$P(a_i|v_j) = \frac{n_i + 1}{n + kosakata} \quad (6)$$

Where *docs*$_j$ is the number of documents in category j and training is the number of documents used in the training process. While ni is the number of occurrences of the word ai in the v_j category. Where n_i is the number of words that appear in the v_j category and vocabulary is the number of unique words in all training data [12], [13].

C. *Support Vector Machine*

SVM is used to find the best hyperplane by maximizing the distance between classes. Hyperplane is a function that is used as a data object separator based on its class. The distance between the hyperplane and the data objects varies. The outermost data object closest to the hyperplae is called a support vector. Support vectors are the most difficult to classify because of their almost overlapping positions with other classes. Given its critical nature, only this support vector is taken into account to find the most optimal hyperplane by SVM [14].

SVM receives input results from feature extraction in numerical form and patterns that will be used in the labeling process. The output of the SVM method is actually a line (hyperplane) that separates positive labeled opinions from negative opinions. From the hyperplane that has been formed, it becomes the basis for labeling new opinions using the kernel function $K(x_i,x_d)$ [15].

In this study, one of the Polynomial kernel equations shown in equation (7) and the Gaussian Radial Basic Function kernel equation shown in equation (7) will be used. (8).

$$K(x_i,x_d) = \left(X_i^T X_{J+1}\right)^d, \gamma > 0 \qquad (7)$$

$$K(x_i,x_j) = \exp\left(-\frac{\left\|x_i,x_j\right\|^2}{2\sigma^2}\right) \qquad (8)$$

The training process uses a sequence learning algorithm with the following steps:

- Calculate the hessian matrix using equation (9) :

$$D_{ij} = y_i y_j \left(K(x_i x_j)\right)^2 + \lambda^2 \qquad (9)$$

- To do the following 3 calculations until the iteration limit:

$$E_i = \sum_{j=1}^{i} \alpha_j D_{ij} \qquad (10)$$

- Will get support vector = (a_j> *thresholdSV*), followed by calculating the value of the bias with equation (11).

$$b = -\frac{1}{2}\left(\sum_{i=1}^{N} a_i y_i K(x_i,x^-) + \sum_{i=1}^{N} a_i y_i K(x_i,x^+)\right) \qquad (11)$$

- Calculate function with equation (12).

$$f(x) = \sum_{i=1}^{m} \alpha_i y_i K(x_i,x) + b \qquad (12)$$

III. METHOD

A. System Overview

The sentiment analysis process was taken from the student opinion questionnaire dataset on the services of the IST AKPRIND campus which was then carried outpre-processing to the dataset. The classification analysis will result in the orientation of positive opinions and negative opinions of the Naïve Bayes Classifier and Support Vector Machine. Additional featuresfeature extraction and selection in classification as a comparison of model performance. The process of this study is illustrated in figure 1:

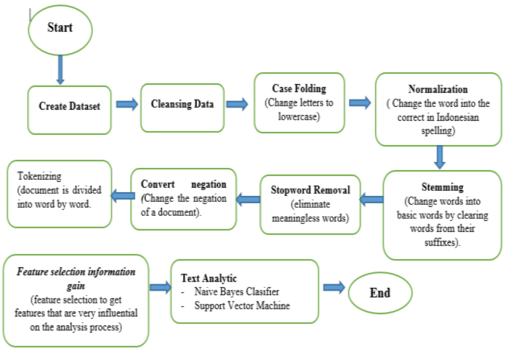

Start
Create Dataset
Cleansing Data
Case Folding
(Change letters to lowercase)
Normalization
(Change the word into the correct in Indonesian spelling)
Stemming
(Change words into basic words by clearing words from their suffixes).
Stopword Removal
(eliminate meaningless words)
Convert negation
(Change the negation of a document).
Tokenizing
(document is divided into word by word.
Feature selection information gain
(feature selection to get features that are very influential on the analysis process)
Text Analytic
- Naive Bayes Clasifier
- Support Vector Machine
End

Figure 1. The Sentiment Analysis Process

B. Dataset

The dataset used in this study was 29,759 opinion text data on the learning evaluation of IST AKPRIND Yogyakarta in the 2014/2015, 2015/2016, and 2016/2017 academic years. The tools used in this research include flask as a framework that provides various libraries needed to create a python website. The programming languages used are Python and HTML.

C. Preprocessing

Data preprocessing is the process of transforming low-quality data into high-quality data that is easier to process [6]. In this study, several data preprocessing techniques were used, including dataset dimension reduction, case folding, punctuation removal, stopword removal, lemmatization, and tokenization. Dimensional reduction refers to the selection of dimensions required for research. The dimensions used in this case are text review. Case folding is the process of converting all letters to lowercase. Only the letters 'a' through 'z' are permitted. Furthermore, the letter is regarded as a delimeter or word separator.

The process of removing punctuation from a sentence is known as remove punctuation. Tokenization is the process of dividing an input string into tokens based on each compiler word. The principle is used to separate every word in a document. The removal of numbers, punctuation, and characters in this process, because the character is considered a word separator and has no effect on text processing.

The process of removing less important words that frequently appear on documents is known as stopword removal. It can eliminate stop words like "which," "the," and "and" to shorten the classification process.

For each tokenized word, lemmatization is the process of converting it into a word or root word. Each word affixed will be removed and converted into a basic word during the lemmatization process, allowing it to further optimize when text processing is completed. Lemmatization is used to convert the words "applied," "words," and "saw" to "apply," "word," and "see." Exsample preprocessing of the research dataset is shown in Table 1.

Table 1. Exsample preprocessing process

Pre processing	Before Pre Processing	After Pre Processing
Cleansing	wifi hidupkan !!!, keramik remuk	wifi hidupkan keramik remuk
Case Folding	Tambah sarana wifi diruang kelas	tambah sarana wifi diruang kelas
Normalization	Tambah sarana wifi diruang kelas	aku cinta akprind
Stemming	Kenyamanan	Nyaman
Stopword removal	kalau bisa bangun ruang kelas gedung lagi untuk perkuliahan karena sudah terlalu banyak mahasiswa tapi ruang kelas dan gedung kurang memadai	ruang kelas gedung perkuliahan mahasiswa ruang kelas gedung kurang memadai
Convert Negation	peningkatan kualitas pada kampus bukan biaya	peningkatan kualitas pada kampus bukanbiaya
Tokenizing	lengkapi sarana prasarana kampus	“lengkapi” “sarana” “prasarana” “kampus”

D. *Feature Selection Information Gain*

Feature selection is the process of selecting features to get features that have a big influence on the analysis process. By using this process, it is hoped that the analysis process will be efficient and the results of the analysis will be accurate.

Table 2 . Example Feature 15 documents

Document	Feature				Sentiment
	Room	Clean	Good	WIFI	
D1	-	-	V	-	Positive
D2	-	-	V	-	Positive
D3	V	-	V	V	Positive
D4	-	-	V	-	Positive
D5	-	V	-	-	Negative
D6	V	V	-	-	Negative
D7	V	-	-	V	Negative
D8	-	V	-	-	Negative
D9	V	V	-	-	Negative
D10	V	-	-	V	Negative
D11	V	-	-	V	Positive
D12	-	-	V	-	Positive
D13	-	-	V	V	Negative
D14	-	-	V	V	Negative
D15	-	V	V	V	Negative

For example, the weight of the information gain will be calculated from the "Bagus" feature. In table 2, out of 15 “Bagus” feature documents, 5 documents with positive sentiments and 3 documents with negative sentiments appear. There are 6 documents with positive sentiments and 5 of them contain “Bagus” features. Then there are 9 documents with negative sentiments and 3 of them contain “Bagus” features. The entropy value can be calculated::

$$Entropy(S) = \left|\left(\frac{8}{15}log_2\frac{8}{15}\right) + \left(\frac{7}{15}log_2\frac{7}{15}\right)\right| = 0{,}992706$$

$$Entropy(S_{positif}) = \left|\left(\frac{5}{6}log_2\frac{5}{6}\right) + \left(\frac{1}{6}log_2\frac{1}{6}\right)\right| = 0{,}650022$$

$$Entropy(S_{negatif}) = \left|\left(\frac{3}{9}log_2\frac{3}{9}\right) + \left(\frac{6}{9}log_2\frac{6}{9}\right)\right| = 0{,}918296$$

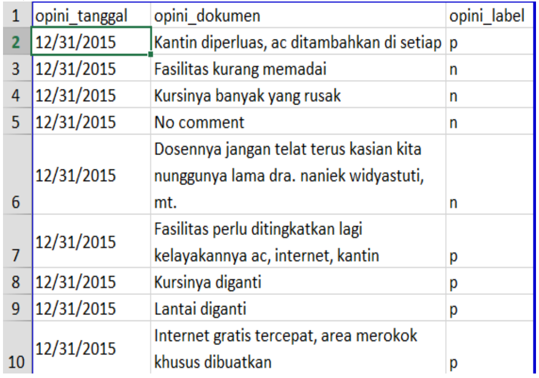
$$Entropy(S, bagus) = \frac{5}{15} \times 0{,}650022 + \frac{3}{15} \times 0{,}918296 = 0{,}400333$$

The last step is to calculate the information gain weight. The information gain weight is used to select features that do not have a major influence in the analysis process. The hope is to streamline the analysis process by using a few features that have a big impact.

$$Gain(bagus) = 0{,}992706 - 0{,}400333 = 0{,}592373$$

E. *Text Analytic using NBC and SVM*

The method used is naive bayes classifer and support vector machine. In this method the data used is divided into 2, namely: Training Data and Test Data. The training data contains a collection of data whose sentiment values are known as in Figure 2, and is used as a benchmark to obtain new sentiment data. In this study, the training data used were 3,999 randomly selected data. Test Data is a collection of data for which the sentiment value is unknown. The test data is filled in by the remaining data that has not been given a sentiment value of 25,760 data.



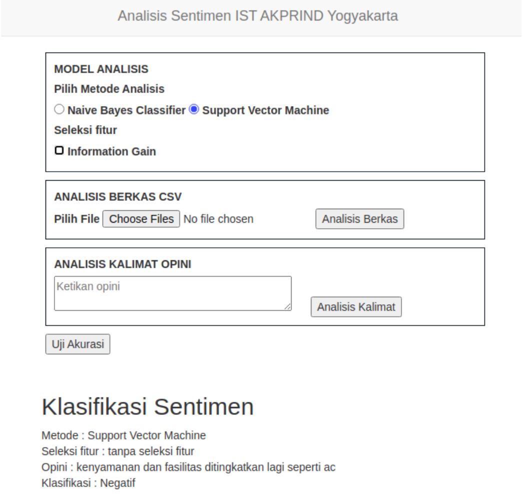
1	opini_tanggal	opini_dokumen	opini_label
2	12/31/2015	Kantin diperluas, ac ditambahkan di setiap	p
3	12/31/2015	Fasilitas kurang memadai	n
4	12/31/2015	Kursinya banyak yang rusak	n
5	12/31/2015	No comment	n
6	12/31/2015	Dosennya jangan telat terus kasian kita nunggunya lama dra. naniek widyastuti, mt.	n
7	12/31/2015	Fasilitas perlu ditingkatkan lagi kelayakannya ac, internet, kantin	p
8	12/31/2015	Kursinya diganti	p
9	12/31/2015	Lantai diganti	p
10	12/31/2015	Internet gratis tercepat, area merokok khusus dibuatkan	p

Figure 2. Training Data Sample

IV. RESULT AND DISCUSSION

In this research, the application of a web-based sentiment analysis system has been made. The application interface can be seen in Figure 3. The application uses four types of analysis models, namely: :

1. *Naive Bayes Clssifier*
2. *Naive Bayes Classifier* with *Information Gain*
3. *Support Vector Machine*
4. *Support Vector Machine* with *Information Gain*.



Analisis Sentimen IST AKPRIND Yogyakarta

MODEL ANALISIS

Pilih Metode Analisis

○ Naive Bayes Classifier ● Support Vector Machine

Seleksi fitur

☐ Information Gain

ANALISIS BERKAS CSV

Pilih File No file chosen

ANALISIS KALIMAT OPINI

Klasifikasi Sentimen

Metode : Support Vector Machine

Seleksi fitur : tanpa seleksi fitur

Opini : kenyamanan dan fasilitas ditingkatkan lagi seperti ac

Klasifikasi : Negatif

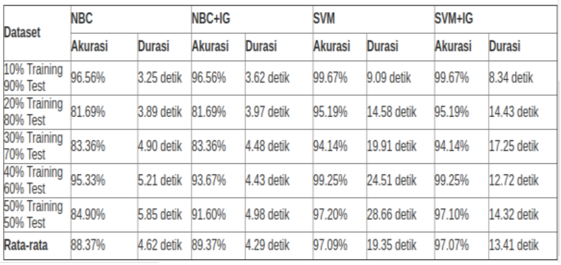
Figure 3. Sentiment analysis application interface

The results show that the accuracy of each model can reach up to more than 90%. The results of the analysis of 25,760 opinions show that there are more negative sentiments than positive sentiments. The model that uses information gain shows a faster analysis process than without information gain. Details of the results of the analysis of the four types of analysis models can be seen in Table 2 and the comparison of the accuracy of the models can be seen in Table 3.

Table 2. Analysis result

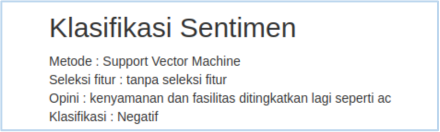
Result Model	NBC	NBC+ IG	SVM	SVM+ IG
Positive	11.038 (42,85%)	12.079 (46,89%)	12.285 (47,69%)	12.266 (47,62%)
Negative	14.722 (57,15%)	13.861 (53,11%)	13.475 (52,31%)	13.494 (52,38%)
Duration	14,47 second	14,6 second	472,49 second	159,78 second

Table 3. Accuracy test results



Dataset	NBC		NBC+IG		SVM		SVM+IG	
	Akurasi	Durasi	Akurasi	Durasi	Akurasi	Durasi	Akurasi	Durasi
10% Training 90% Test	96.56%	3.25 detik	96.56%	3.62 detik	99.67%	9.09 detik	99.67%	8.34 detik
20% Training 80% Test	81.69%	3.89 detik	81.69%	3.97 detik	95.19%	14.58 detik	95.19%	14.43 detik
30% Training 70% Test	83.36%	4.90 detik	83.36%	4.48 detik	94.14%	19.91 detik	94.14%	17.25 detik
40% Training 60% Test	95.33%	5.21 detik	93.67%	4.43 detik	99.25%	24.51 detik	99.25%	12.72 detik
50% Training 50% Test	84.90%	5.85 detik	91.60%	4.98 detik	97.20%	28.66 detik	97.10%	14.32 detik
Rata-rata	88.37%	4.62 detik	89.37%	4.29 detik	97.09%	19.35 detik	97.07%	13.41 detik

Based on the results of the average accuracy test, it can be concluded that the Support Vector Machine method is more accurate and more stable than the Naive Bayes Classifier method, with an average accuracy of 97.09% with the highest value of 99.67%. This application can also be used for sentence analysis per category through the File Analysis button. An example of the results of the analysis can be seen in Figure 4. As a result of the comparison of the results of the analysis of the 4 analysis models can be seen in table 4.



Klasifikasi Sentimen

Metode : Support Vector Machine

Seleksi fitur : tanpa seleksi fitur

Opini : kenyamanan dan fasilitas ditingkatkan lagi seperti ac

Klasifikasi : Negatif

Figure 4. Example of sentence analysis results

Tabel 4. Comparison of Sentence Analysis Results

Opinion	NBC	NBC+ IG	SVM	SVM+ IG
Comfort and facilities are improved again such as air conditioning	Positive	Positive	Negative	Negative
We need air conditioning	Negative	Negative	Negative	Negative
Less cctv parking lot less spacious	Negative	Negative	Negative	Negative

V. CONCLUSIONS

Based on the results of the research that has been carried out, it can be concluded that the combination of Naive Bayes Classifier with Information Gain and Support Vector Machine with Information Gain can analyze sentiment automatically. The results of trials using opinion data collected from 2014 to 2017 show that negative sentiment is more than positive sentiment. The accuracy of the analysis results reached 99.67% with an average of 97.09%.

SVM method has higher accuracy than NBC. Support vector machine produces the highest accuracy reaching 99.67% and the lowest 94.17%. Meanwhile, the Naive Bayes classifier recorded the highest accuracy up to 96.56% and the lowest 81.69%. The application of information gain does not significantly affect the accuracy. However, it is very influential on the duration of the analysis, especially on the SVM method. In the test data analysis process, the application of information gain on SVM accelerates the duration of the analysis process by 195.71%.

REFERENCES

[1] F. A. Pozzi, E. Fersini, E. Messina, en B. Liu, *ANALYSIS IN SOCIAL NETWORKS :*, vol 1. Elsevier Inc., 2017.

[2] P. Gupta, R. Tiwari, en N. Robert,

"Sentiment analysis and text summarization of online reviews: A survey", in *International Conference on Communication and Signal Processing, ICCSP 2016*, 2016.

[3] J. Song, K. T. Kim, B. Lee, S. Kim, en H. Y. Youn, "A novel classification approach based on Naïve Bayes for Twitter sentiment analysis", *KSII Trans. Internet Inf. Syst.*, vol 11, no 6, bll 2996–3011, 2017.

[4] M. Ahmad, S. Aftab, M. S. Bashir, N. Hameed, I. Ali, en Z. Nawaz, "SVM optimization for sentiment analysis", *Int. J. Adv. Comput. Sci. Appl.*, 2018.

[5] Gintautas GARˇSVA, P. DAN˙ENAS, en Gintautas GARˇSVA, "SVM and Naıve Bayes Classification Ensemble.pdf", *Balt. J. Mod. Comput.*, vol 5, no 4, bll 398–409, 2017.

[6] S. M. Weiss, N. Indurkhya, en T. Zhang, *Fundamentals of Predictive Text Mining*. Springer International Publishing, 2015.

[7] C. Zucco, B. Calabrese, G. Agapito, P. H. Guzzi, en M. Cannataro, "Sentiment analysis for mining texts and social networks data: Methods and tools", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2020.

[8] P. Nandwani en R. Verma, "A review on sentiment analysis and emotion detection from text", *Social Network Analysis and Mining*. 2021.

[9] U. Lestari en D. Anugrahni, "Sentiment Analysis of Performance Effectiveness of Malioboro Pedestrian Using Sentistrength Method on Twitter", *J. TAM (Technology Accept. Model.*, vol 12, no 1, bll 75–79, 2021.

[10] G. A. Buntoro, "Analisis Sentimen Calon Gubernur DKI Jakarta 2017 Di Twitter", *INTEGER J. Inf. Technol.*, vol 1, no 1, bll 32–41, 2017.

[11] D. A. Muthia, "Analisis Sentimen Pada Review Restoran Dengan Teks Bahasa Indonesia Mengunakan Algoritma Naive Bayes", *Jurnalilmu Pengetah. Dan Teknol. Komput.*, vol 2, no 2, bll 39–45, 2017.

[12] M. Granik en V. Mesyura, "Fake news detection using naive Bayes classifier", *2017 IEEE 1st Ukr. Conf. Electr. Comput. Eng. UKRCON 2017 - Proc.*, bll 900–903, 2017.

[13] S. Dey, S. Wasif, D. S. Tonmoy, S. Sultana, J. Sarkar, en M. Dey, "A Comparative Study of Support Vector Machine and Naive Bayes Classifier for Sentiment Analysis on Amazon Product Reviews", *2020 Int. Conf. Contemp. Comput. Appl. IC3A 2020*, bll 217–220, 2020.

[14] J. Liu en E. Zio, "Integration of feature vector selection and support vector machine for classification of imbalanced data", *Appl. Soft Comput. J.*, vol 75, bll 702–711, 2019.

[15] C. Jian, J. Gao, en Y. Ao, "A new sampling method for classifying imbalanced data based on support vector machine ensemble", *Neurocomputing*, vol 193, bll 115–122, 2016.