Data Science : Journal Of Computing And Applied Informatics
Journal homepage: https://jocai.usu.ac.id

Analysis Sentiment Of Users Internet Service Providers In Indonesia On Social Media X Using Support Vector Machine

Fachrurrozy Nurqoulby[*], Amalia Anjani Arifiyanti and Dhian Satria Yudha Kartika

Information System, Faculty of Computer Science, Universitas Pembangunan Nasional "Veteran" Jawa Timur, 60294, Surabaya, Indonesia

[*]Corresponding Author: fachrurrozyn@gmail.com

ARTICLE INFO

Article history:
Received 03 May 2024
Revised 14 May 2024
Accepted 17 May 2024
Available online 31 July 2024

E-ISSN: 2580-829X
P-ISSN: 2580-6769

How to cite:
Fachrurrozy Nurqoulby, Amalia Anjani Arifiyanti and Dhian Satria Yudha Kartika, "Analysis Sentiment Of Users Internet Service Providers In Indonesia On Social Media X Using Support Vector Machine", Data Science : Journal Of Computing And Applied Informatics, vol.8, no.2, July 2024, doi: http://doi.org/10.32734/jocai.v8.i2-16317

ABSTRACT

Various internet service providers are starting to appear in Indonesia, they are competing to provide attractive offers to attract customers. Through social media, someone can find out opinions about whether internet service providers provide the services as offered. X, formerly known as Twitter, is a social media platform where people can give their opinions in the form of *posts*. Various opinions were expressed by the public, ranging from positive, neutral, to negative. This research aims to create a *post* classification model regarding users of internet service providers into three sentiment classes, namely positive, neutral, and negative. The model is created through several stages, such as data retrieval, data labeling, data preprocessing, data division, term weighting, and creating a classification model using the Support Vector Machine algorithm. The results of this research show that the SVM model with a Linear kernel obtained the highest accuracy of 83% compared to the RBF kernel SVM and Polynomial kernel SVM, with an F1-score of 90% for the negative class, 66% for the neutral class, and 65% for the positive class.

Keyword: Text Mining, Analysis Sentiment, X, Support Vector Machine, Classification Text

ABSTRAK

Kegiatan audit internal di salah satu perusahaan EPC telah menemukan tren meningkatnya stres kerja sebagai alasan pengunduran diri karyawan pada periode Q4 2021 – Q1 2023. Dalam implementasi ISO 45001:2015 ini harus dikendalikan karena cenderung menjadi penyakit psikologis kerja. Oleh karena itu, dilakukan survei stres kerja. Hasilnya ditinjau menggunakan Cross Industry Standard Process for Data Mining. (CRISP-DM). Analisis deskriptif menemukan dua faktor yang mempengaruhi tingkat stres kerja, yaitu tuntutan untuk kualitas kerja dan tanggung jawab atas hasil kerja orang lain. Lebih khusus lagi, tingkat stres kerja adalah karena karyawan diminta untuk melebihi kemampuan mereka tetapi pada saat yang sama harus membantu orang lain memecahkan masalah. Selain itu, berdasarkan analisis Cluster, 2 cluster optimal ditemukan. Cluster-1 memiliki centroid stres moderat untuk faktor stres kerja secara keseluruhan. Cluster-2 memiliki centroid stres ringan untuk faktor stres kerja secara keseluruhan. Rekomendasi untuk mengendalikan stres kerja di cluster-1 adalah untuk menyiapkan program untuk meningkatkan kompetensi karyawan dan meningkatkan sistem pengukuran kinerja. Pengendalian stres kerja cluster-2 adalah pemantauan tahunan melalui survei stres kerja karyawan.

Keyword: Text Mining, Analisis Sentimen, X, Support Vector Machine, Klasifikasi Teks

1 Introduction

In the digital era, which continues to develop every day, the internet has become an inseparable part of everyday life. With the internet, it has made it easier for users to carry out various activities, such as interacting with friends and family, searching for information, watching television shows or films, and even shopping online. To be able to support these various activities, people need internet service providers who can access the internet smoothly and quickly according to their needs. Internet service provider companies are competing to increase the number of users of their services by providing many promotions, such as free installation fees, cheap subscription prices, and bonus cable television channels.

Because many new internet service providers are starting to emerge, people are using social media as a means to find the latest information regarding internet service providers that are currently being used or are planning to start subscribing or making service complaints. Twitter, which has now changed its name to X, is a media site where users can share and discuss everything, including news, jokes, their opinions about events, and even their mood. With a simple interface where only 280 character messages can be posted, called *tweets*, Twitter is increasingly becoming a system for getting information in real-time [1]. However, sometimes some information cannot be conveyed properly due to the use of language structuring, which may be difficult to understand due to the abbreviation of words or the use of non-standard words.

Sentiment analysis is a part of text mining that can identify people's opinions, sentiments, and emotions through their entities and attributes expressed in text form [2]. There are many other terms regarding sentiment analysis, including opinion mining, opinion extraction, subjectivity analysis, and emotion analysis. This sentiment analysis focuses on opinions that have positive or negative sentiment [3]. Sentiment analysis can be useful for analyzing someone's opinion that they express via social media, which is then translated into information. By using sentiment analysis, it is hoped that it can help to determine the sentiment of users' responses to an internet service provider, which can later be useful for companies to fix issues or disorders that occur or to improve the services.

To be able to carry out sentiment analysis, a machine learning approach with Support Vector Machines (SVM) is used, which is included in the linear classification category. The SVM method is carried out by looking for a hyperplane or decision boundary that separates one class from another class. SVM can solve linear problems and has been developed to work on non-linear problems by incorporating the concept of kernels in high-dimensional workspaces. The kernel function is used to map the initial dimensions of a data set to higher dimensions [4]. In the classification process using SVM, the use of kernel tricks can help in classifying a sentiment so that the results obtained are more optimal [5].

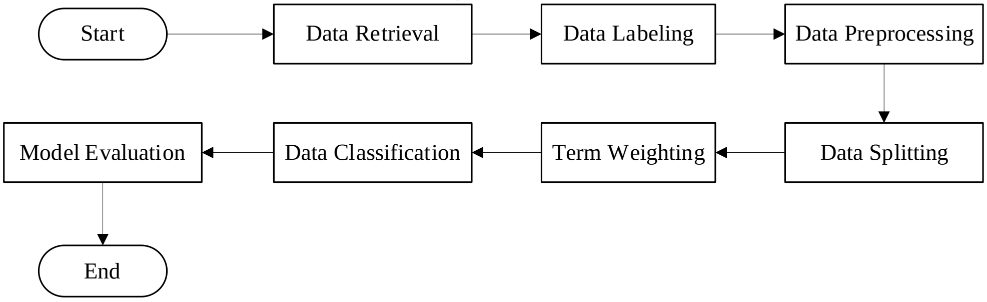
In research conducted by Shivaprasad and Shetty [6], they compared the Support Vector Machine, Naïve Bayes (NB), and Maximum Entropy (ME) methods to carry out sentiment analysis on review data on online shopping sites. The SVM method obtained the highest average accuracy value compared to other classification methods. Then, in research conducted by Ullah et al. [7] with the title "An algorithm and method for sentiment analysis using text and emoticons", it was found that classification using text and emoticons had slightly better results than just text alone. By using machine learning, the SVM algorithm has better results compared to Random Forest, Naïve Bayes, and Logistic Regression. The SVM algorithm with text and emoticons classification has an accuracy value of 78%, precision of 73%, recall of 71%, and F1-Score of 78%. Meanwhile, the SVM algorithm with text only has an accuracy value of 78%, a precision value of 74%, a recall of 69%, and an F1-Score of 71%.

In other research conducted by Yonatha and Eka [8] regarding the effect of using kernels on sentiment analysis, it was found that linear kernels had the best results compared to the other two types of kernels. The kernels used in this research include the linear kernel, the RBF kernel, and the polynomial kernel.

2 Methods

There are several stages in the research method, here is the flowchart that shows how the flow of the research is done.

Figure 1 Research Flow



2.1 Data Retrieval

The data used in this research was obtained from scraping data on X using the *snscrape* library. The data collected is post data containing several internet service providers in Indonesia (First Media, Indihome, and Biznet).

2.2 Data Labeling

The data that has been obtained is then given a label. Labeling of data is done manually which is divided into 3 classes, namely positive, neutral, and negative. This stage is necessary because classification is included in supervised learning, which means labels or classes are needed to be used in training the classification model created [9]. By having labels in the model, machine learning can learn them better so that it can make more accurate predictions.

2.3 Data Preprocessing

After labeling, the data will then be processed further through the preprocessing stage. This stage functions so that the data obtained is more structured so that it can produce a better model [10]. There are several processes that need to be carried out, including:

A. *Case folding is a process that changes all the letters in a text into a uniform form, namely in lowercase.*

B. *Cleaning data is the process of cleaning characters in text that are not needed, such as emojis, usernames, numbers, hashtags, URLs, single letters, or punctuation marks*

C. *Tokenization is the process of breaking text or sentences into smaller units*

D. *Stopword removal is a process for removing words that do not have special meaning*

E. *Stemming is a natural language processing (NLP) which aims to remove affixes from words, leaving only the basic form of the word*

2.4 Data Splitting

The data splitting stage is carried out to divide the data into two parts, namely training data and test data. Train data is used as a place to train the model, while test data is used to see how well the model performs. The data sharing stage is carried out using the *k-fold cross-validation* data splitting type. The *cross-validation* algorithm method is carried out by randomly dividing data samples and grouping the data by *k* values. The *k* value that will be used is 10-fold.

2.5 Term Weighting

The term-weighting stage is carried out to give weight or value to each word contained in the text document. The more often the word appears, the greater its weight. This stage is necessary because it can help identify words that have a significant impact on understanding the document's content. The term weighting used in this research is Term Frequency-Inverse Document Frequency (TF-IDF). Term Frequency is used to measure how many times a term is present in a document. Inverse Document Frequency gives weight to words that appear, namely a lower weight for words that appear frequently, while those that rarely appear will have a higher weight [11].

The following is the formula for calculating TF-IDF [12]:

$$TF = \frac{number\ of\ word\ occurrences}{number\ of\ words\ in\ the\ document} \tag{1}$$

$$IDF = log\ \frac{number\ of\ documents}{the\ number\ of\ times\ a\ word\ appears} \tag{2}$$

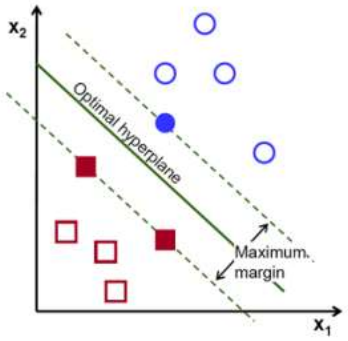
$$TFIDF\ =\ TF\ .IDF \tag{3}$$

2.6 Data Classification

The data classification stage is carried out with the aim of predicting the correct labels on the test data by studying the train data. The type of machine learning used at the classification stage is Support Vector Machines. SVM works by finding the best dividing line (hyperplane) that is used to separate the two classes. Margin is the distance between the closest hyperplane and the closest pattern from each class, and the pattern closest to the hyperplane is called the support vector [13]. The following is an illustration of SVM in Figure 2.

Figure 2 Illustration of Support Vector Machine Classification

In Support Vector Machines there are several types of kernels that can be used, including:



A. Linear

The linear kernel can also be called a soft margin, which will try to find a hyperplane that is a straight line but can tolerate one or more data classification errors. Even though it tolerates some errors, the linear kernel still tries to find the line that has the maximum margin and minimizes the error [14].

$$K\ (x, xi)\ =\ x.xi\ +\ C \tag{4}$$

B. Radial Basis Function (RBF)

This RBF kernel is also commonly called the Gaussian kernel. It is one of the most widely used kernel functions due to its good learning ability from all single kernel functions. This kernel is suitable for use when the data is uneven. When training a dataset using the RBF kernel, there are two parameters that need to be considered, namely C and gamma. The C parameter is useful for telling how many errors must be avoided in classifying training data. The greater the C value, the lower the classification error for the training data. The gamma parameter determines how much influence one sample of training data has. It can be interpreted that the smaller the gamma value, the greater the distance from the data points to be calculated [14].

$$K\ (x, xi)\ =\ exp(-\gamma||x.xi||2), \gamma\ >\ 0 \qquad (5)$$

C. Polynomials

Polynomial kernels are used when data cannot be separated by a straight line. Polynomial kernels can produce nonlinear decision boundaries. This kernel produces new features by applying a polynomial combination of existing features [14].

$$K\ (x, xi)\ =\ (\gamma\, x.xi\ +\ C)d, d\ >\ 0 \qquad (6)$$

2.7 Model Evaluation

The model evaluation stage is carried out to find out how accurate the model that has been created is. A confusion matrix is a tool used to understand the performance of a classification model by measuring how well the model predicts the correct classes. Table 1 shows a description of the confusion matrix, which stores four combinations of actual and predicted values [15].

Table 1 Confusion Matrix

		Predicted Class	
		Positive	Negative
Actual Class	Positive	TP	FP
	Negative	FN	TN

By using the confusion matrix, you can continue by calculating various model performance evaluation metrics, such as accuracy, precision, recall, and F1-score. Accuracy is the ratio obtained from the correct prediction values (positive and negative) from the total amount of data. Precision is the ratio of true positive predictions compared to all data that is true positive. Recall is the ratio of true positive predictions compared to all data that is true positive. Then F1-Score is a calculation of the average between precision and recall [16]. The following is the formula used to calculate accuracy, precision, recall, and F1-Score [17].

$$\text{Accuracy}\ = \frac{TP + TN}{TP + TN + FP + FN} \qquad (7)$$

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \tag{8}$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \tag{9}$$

$$\boldsymbol{F1 - Score} = \boldsymbol{2}\frac{\boldsymbol{Precision\,.Recall}}{\boldsymbol{Precision} + \boldsymbol{Recall}} \tag{10}$$

3 Results

The data obtained from scraping totaled 8529 rows of data taken from July to December 2022. The data contains the date and time the post was created, the name of the post creator, the text of the post, and the location of the post creator.

The data is then labeled one by one on each row and done manually by one person. The data is also filtered and the amount of data is equalized for each internet service provider being searched for. The amount of data then changed to 7500 lines of data, with details of 2500 lines of data for First Media, 2500 lines of data for Indihome, and 2500 lines of data for Biznet. Of the 7500 rows of data, negative labels totaled 5179 (69%), neutral labels totaled 1697 (23%), and finally positive labels totaled 624 (8%).

The model created has gone through several stages after labeling, such as data preprocessing, data splitting, and term weighting.

Table 2 Confusion Matrix of Linear Kernel Model

		Predicted Class		
		Negative	Neutral	Positive
Actual Class	Negative	492	36	1
	Neutral	56	104	3
	Positive	15	13	30

Table 2 shows that 492 data predicted to be negative correspond to the actual class, called True Negative (TN), and 37 data were incorrectly predicted, called False Negative (FN). Then 104 data are predicted to be neutral and match the actual class, called True Neutral (TN), and 59 data are incorrect predictions, called False Neutral (FN). Furthermore, 30 data are predicted to be positive and correspond to the actual class, called True Positive (TP), and 28 data are incorrect predictions, called False Positive (FP).

Table 3 Confusion Matrix of RBF Kernel Model

		Predicted Class		
		Negative	Neutral	Positive
Actual Class	Negative	498	31	0
	Neutral	71	90	2
	Positive	22	10	26

Table 3 shows that 498 data predicted to be negative correspond to the actual class, called True Negative (TN), and 31 data were incorrectly predicted, called False Negative (FN). Then 90 data are predicted to be neutral and match the actual class, called True Neutral (TN), and 73 data are incorrect predictions, called False Neutral (FN). Furthermore, 26 data are predicted to be positive and correspond to the actual class, called True Positive (TP), and 32 data are incorrect predictions, called False Positive (FP).

Table 4 Confusion Matrix of Polynomial Kernel Model

		Predicted Class		
		Negative	Neutral	Positive
Actual Class	Negative	514	15	0
	Neutral	129	33	1
	Positive	37	3	18

Table 4 shows that 514 data predicted to be negative correspond to the actual class, called True Negative (TN), and 15 data were incorrectly predicted, called False Negative (FN). Then 33 data are predicted to be neutral and match the actual class, called True Neutral (TN), and 130 data are incorrect predictions, called False Neutral (FN). Furthermore, 18 data are predicted to be positive and correspond to the actual class, called True Positive (TP), and 40 data are incorrect predictions, called False Positive (FP).

The following are the classification results obtained after carrying out a confusion matrix using a support vector machine with linear, RBF and polynomial kernels.

Table 5 Comparison of Classification Results

Kernel Type		Accuracy	Precision	Recall	F1-Score	Processing Time
Linear	Negative	83%	87%	93%	90%	2,7 second
	Neutral		68%	64%	66%	
	Positive		88%	52%	65%	
RBF	Negative	82%	84%	94%	89%	4,3 second
	Neutral		69%	55%	61%	
	Positive		93%	45%	60%	
Polynomial	Negative	75%	76%	97%	85%	6,7 second
	Neutral		65%	20%	31%	
	Positive		95%	31%	47%	

Based on the results from the table above, it can be said that the Support Vector Machine classification model using the linear kernel has the highest accuracy of 83% with an F1-score of 90% for the negative class, 66% for the neutral class, and 65% for the positive class. From this table, it can also be seen that the linear kernel has the fastest processing time with only 2.7 seconds. It can be concluded that SVM with a linear kernel is better than using the RBF kernel or polynomial kernel in this research.

4 Conclusions

From the results of the analysis and testing that have been carried out, it can be concluded that the Support Vector Machine model with a linear kernel has the highest accuracy of 83% and the fastest processing time with only 2.7 seconds compared to the RBF and polynomial kernels. In testing the model, it was found that there were quite large differences in F1-score values. This is because the comparison of the amount of data in each sentiment class is not balanced and has quite significant differences. The suggestion for further research is to use data between classes whose numbers do not differ significantly or use a data balancing method.

References

[1] Ivan, Y. A. Sari, and P. P. Adikara, "Klasifikasi Hate Speech Berbahasa Indonesia di Twitter Menggunakan Naive Bayes dan Seleksi Fitur Information Gain dengan Normalisasi Kata," *J. Pengemb. Teknol. Inf. Dan Ilmu Komput.*, vol. 3, no. 5, pp. 4914–4922, 2019.

[2] O. Oueslati, E. Cambria, M. Ben HajHmida, and H. Ounelli, "A review of sentiment analysis research in Arabic language," *Futur. Gener. Comput. Syst.*, vol. 112, pp. 408–430, 2020.

[3] B. Liu, *Sentiment Analysis and Opinion Mining*. Morgan & Claypool, 2012.

[4] E. Fitri, Y. Yuliani, S. Rosyida, and W. Gata, "Sentiment Analysis of the Ruangguru Application Using Naive Bayes, Random Forest and Support Vector Machine Algorithms," *J. Transform.*, vol. 18, no. 1, p. 71, 2020.

[5] P. Fremmuzar and A. Baita, "Uji Kernel SVM dalam Analisis Sentimen Terhadap Layanan Telkomsel di Media Sosial Twitter," *Komputika J. Sist. Komput.*, vol. 12, no. 2, pp. 57–66, 2023.

[6] T. K. Shivaprasad and J. Shetty, "Sentiment analysis of product reviews: A review," *Int. Conf. Inven. Commun. Comput. Technol.*, 2017.

[7] M. A. Ullah, S. M. Marium, S. A. Begum, and N. S. Dipa, "An algorithm and method for sentiment analysis using the text and emoticon," *ICT Express*, vol. 6, no. 4, pp. 357–360, 2020.

[8] K. D. W. Yonatha and A. Eka, "The Effects of Different Kernels in SVM Sentiment Analysis on Mass Social Distancing," vol. 9, no. 2, pp. 161–168, 2020.

[9] F. S. Darusman, A. A. Arifiyanti, and S. F. A. Wati, "Sentiment Analysis Pedulilindungi Tweet Using Support Vector Machine Method," *Appl. Technol. Comput. Sci. J.*, vol. 4, no. 2, pp. 113–118, 2022.

[10] M. Kamber and J. Han, *Data Mining: Concepts and Techniques : Concepts and Techniques*. 2018.

[11] S. Qaiser and R. Ali, "Text Mining : Use of TF-IDF to Examine the Relevance of Words to Documents," vol. 181, no. 1, pp. 25–29, 2018.

[12] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Unstructure Data*. New York: Cambridge University Press, 2007.

[13] P. Nomleni, "Sentiment Analysis Menggunakan Support Vector Machine (SVM)," Institut Teknologi Sepuluh November, 2015.

[14] A. F. Rochim, K. Widyaningrum, and D. Eridani, "Comparison of Kernels Function between of Linear, Radial Base and Polynomial of Support Vector Machine Method Towards COVID-19 Sentiment Analysis," pp. 224–228, 2021.

[15] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit.*, vol. 91, pp. 216–231, 2019.

[16] T. N. Wijaya, R. Indriati, and M. N. Muzaki, "Analisis Sentimen Opini Publik Tentang Undang-Undang Cipta Kerja Pada Twitter," *Jambura J. Electr. Electron. Eng.*, vol. 3, no. 2, pp. 78–83, 2021.

[17] Q. A. Memon and S. Ahmed Khoja, Data Science Theory, Analysis, and Applications, 1st ed. CRC Press, 2019.