Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Twitter Sentiment Analysis of Public Space Opinions using SVM and TF-IDF Methods

Ulya Ilhami Arsyah[1*], Mutiana Pratiwi[2], Abulwafa Muhammad[3]
ulya@pnp.ac.id, mutiana_pratiwi@upiyptk.ac.id, abulwafa_muhammad@upiyptk.ac.id
[1]Politeknik Negeri Padang
[2,3]Universitas Putra Indonesia YPTK Padang

Article Information

Submitted : 12 Dec 2023
Reviewed: 2 Jan 2024
Accepted : 15 Feb 2024

Keywords

Sentiment Analysis, Support Vector Machine, Comments, Public Spaces, Twitter

Abstract

Public space opinion reviews are currently a source of information for interested parties and decision-makers. Twitter is a social media that is a means of expressing themselves for people to express their opinions and criticize the current situation. This becomes information for readers. Information published on Twitter contains elements of commentary on a situation or object Sentiment analysis of public space opinion on Twitter using Machine Learning with the Support Vector Machine (SVM) method with the data weighting process using the Term Frequency-Inverse Document Frequency (TF-IDF) method. Dataset obtained by scraping using the Twitter API as much as 5000 data then labeled where the goal is to get accuracy on positive, negative, or neutral sentiment. The results of research conducted experiments on three Machine Learning algorithms with the extraction function "TF-IDF" obtained an accurate training model with good classification capabilities, especially SVM of 91,6% on data distribution 70: 30; SVM is 92.8% in the case of data distribution of 80: 20; the SVM is 91,8% in the case of 90:10 decomposition data.

A. Introduction

The development of technology has become a way to express opinions and thoughts about what is happening today. Social media is one of the most well-known information technologies. Nowadays, everyone utilizes social networks to get information and provide information in real-time. With social media users increasing daily over the past decade, millions of people have begun to express their opinions on various topics easily[1]. This is in line with the increase in data available on social media platforms such as Twitter (X). Determining people's opinions on current issues is crucial for policy-making.

Social networks provide opportunities for people to participate directly. Many people take advantage of this opportunity just to participate, some use social networks for information-sharing activities. However many people use social media to disseminate information and express themselves as a form of existence[2]. No matter if the person is upper or lower class, young or old, male or female, or even from buskers to the president, they all use social networks to share information about what they are doing. The development of social networks as a means of public space that currently replaces conventional media has given birth to many digital artists better known as celebrities. They show off their abilities like social media stars, almost all of their activities are displayed on social networks[3]. From waking up to going back to sleep, that's what their social media looks like. The presence of social media blurs the boundaries between the private and public spheres because people cannot distinguish between the two[4]. To anticipate this, of course, we must be careful in using social networks, we must be able to distinguish which social networks are private domains and which social networks are public open environments[5]. In addition, social networks have become a public space that can accommodate all ideas, thoughts, and opinions of the community. Many social media users utilize digital platforms to comment on anything, including government activities in the public sector[6].

Twitter can be used as a data source to analyze sentiment towards government policies. By classifying tweets into positive, negative, or neutral categories, this analysis can provide insight into the public's view of the policy. It is reported that there were 556 million Twitter users worldwide in January 2023, and Indonesia ranked fifth in the number of Twitter users[7][8]. Collecting data manually from Twitter will take a lot of time, therefore other techniques such as web scraping are needed to retrieve data quickly and efficiently. The data generated from web scraping can then be used for analysis. One type of analysis that is often used in text data analysis is sentiment analysis, where data is classified based on patterns and functions that distinguish data into certain groups[9][10].

Sentiment analysis is a technique for measuring people's opinions on the level of agreement on a topic. Commonly used approaches are natural language processing and machine learning algorithms. One method often used in sentiment analysis is the Support Vector Machine (SVM), which is considered the best method in text classification[11][12]. SVM can separate classes linearly with a large margin and can handle infinite dimen-sional feature vectors. Previously, research was conducted on Sentiment Analysis related to the fuel price case using the SVM method. This study used 258 data from Twitter user com-ments. The

results show that there are more positive than negative comments related to the increase in fuel prices[13][14].

This research focuses on senti-ment analysis using the SVM method to classify tweets related to public space. The researchers chose public space as a research topic because public space is currently a hotly discussed topic related to infrastructure development and the relocation of the National Capital. This research aims to collect opinions expressed by online communities on so-cial media related to these topics. This research is considered important because it can provide input to the government and stakeholders regarding the aspira-tions and views of the community so that it can help in making policies and plan-ning future work programs. In addition, the results of this research can be used as a reference for the development of sentiment analysis methods, especially in the context of social media. By using the SVM method, this research contributes to the development of algorithmic tech-niques for more accurate sentiment analysis.

B. Research Method

In this section, the methods and techniques used for the classification of reviews on public opinion spaces and the steps taken during the experiments are classified. Figure 1 illustrates the stages of this research starting from the online review dataset until each review is classified into positive, negative, and neutral.

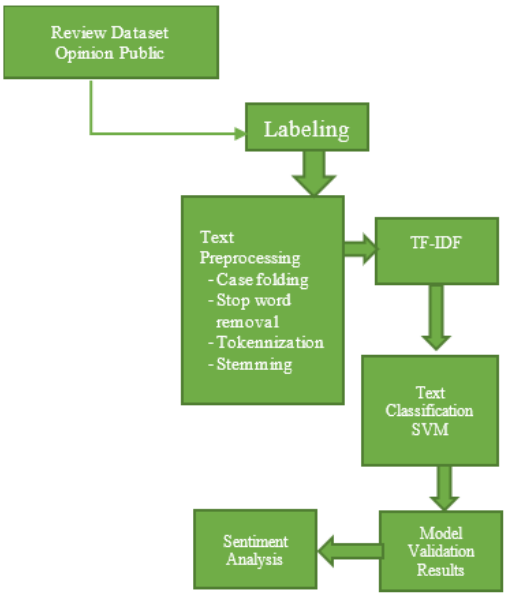

Figure 1. Research Framework

The first stage carried out in this research as shown in Figure 1 is the public opinion review dataset. The dataset is obtained by web scraping on Twitter. The second stage is to label the dataset, the labels given are positive, negative, and neutral sentiment. Datasets that have been labeled are then ready to be used for preprocessing. The preprocessing stage involves case-folding, stopword removal, tokenization, and stemming. The preprocessed dataset is continued with TF-IDF analysis (Term Frequency-Inverse Document Frequency, this is done to measure how important a word (term) is in a document or text corpus by giving weight to these words based on the two main factors of the frequency of occurrence of

words in documents (TF) and the presence of these words in the entire text corpus (IDF). In the next stage, the text classification is carried out using the SVM method and continued with the model validation process. The next step is to conduct an Intent Sentiment Analysis to find out the motivation for writing a review.

a. Research Data

The tweet data taken is Indonesian language data. The amount of data obtained is a sample of data from the results of scraping as much as 5000 data and the amount of data is preprocessed to produce data selection data. The following is an example of Twitter data samples in Table 1 taken using scraping techniques:

Table 1. Dataset

No	Tweet
1	Saya perlu mempertanyakan hak-hak saya sebagai manusia dan sebagai perempuan, di mana pada intinya perempuan memiliki hak untuk merasa aman di ruang publik tanpa dibebani rasa takut akan pelecehan dan kekerasan.
2	Gubernur Anies menggagas dan menyelenggarakan perayaan Christmas Carol di ruang publik di sepanjang Sudirman Thamrin DKI Jakarta, agar umat Kristiani juga merasakan Jakarta sebagai rumah sendiri.
3	RT @pinguingurun7o_: Kalo beropini di ruang publik tapi pake bahasa daerah mending lu diem aja tot bangsat emang, sampah [RE pinguingurun7o_]
4…	

b. Labeling

This research first conducted an initial data selection which aims to select the right data that falls into the category of comment data related to public space. After selection, comment data that is positive, negative, or neutral is obtained. Some examples of data selection results can be seen in Table 2.

Table 2. Data Labeling

No	Teks	Sentiment
1	Perlu dipertanyakan lagi hak saya sebagai seorang manusia dan sebagai seorang perempuan, dimana pada hakikatnya perempuan memiliki hak untuk merasa aman di ruang publik tanpa dibebani rasa takut akan pelecehan dan kekerasan.	Neutral
2	Gubernur Anies inisiasi dan selenggarakan perayaan Natal Christmas Carol di ruang publik di sepanjang Sudirman Thamrin DKI Jakarta, agar umat Kristen juga merasakan Jakarta sebagai rumah sendiri.	Positif
3	RT @pinguingurun7o_: Kalo beropini di ruang publik tapi pake bahasa daerah mending lu diem aja tot bangsat emang, sampah [RE pinguingurun7o_]	Negatives

C. Result and Discussion

This research was conducted in several stages, the first stage was data collection, the data collected was review data from tweets on Twitter. The data used to test this research is review data in Indonesian language public opinion in public spaces. The dataset used in this research is 5000 reviews of tweets in the public space. In this research, we do Pre-processing then we do TF IDF feature

extraction. Then we classify the review texts to find intent sentiment analysis with the ML model.

Preprocessing Twitter comment data needs to be done before the classification process, to eliminate words unsuitable for research, equalize the shape of words or letters (lowercase), and reduce the number of words. By preprocessing the data, it can be ensured that the data used for training is better, cleaner, and more suitable for use in SVM analysis. After text preprocessing, it is necessary to weigh the words. Some sample results of the text preprocessing process can be presented in Table 2.

Table 3. Data Pre-Processing (Case Folding)

No	Text	Result
1	Perlu dipertanyakan lagi hak saya sebagai seorang manusia dan sebagai seorang perempuan, dimana pada hakikatnya perempuan memiliki hak untuk merasa aman di ruang publik tanpa dibebani rasa takut akan pelecehan dan kekerasan.	perlu tanya lagi hak saya sebagai seorang manusia dan sebagai seorang perempuan dimana pada hakikat perempuan memiliki hak untuk merasa aman di ruang publik tanpa beban rasa takut akan pelecehan dan kekerasan
2	Gubernur Anies inisiasi dan selenggarakan perayaan Natal Christmas Carol di ruang publik di sepanjang Sudirman Thamrin DKI Jakarta, agar umat Kristen juga merasakan Jakarta sebagai rumah sendiri.	gubernur anies inisiasi dan selenggara perayaan natal christmas carol di ruang publik di sepanjang sudirman thamrin dki jakarta agar umat kristen juga merasa jakarta sebagai rumah sendiri
3	RT @pinguingurun7o_: Kalo beropini di ruang publik tapi pake bahasa daerah mending lu diem aja tot bangsat emang, sampah [RE pinguingurun7o_]	opini di ruang public tetapi Bahasa daerah

Table 2 is an example of a tweet that has been labeled and processed to produce a Bar Chart of the number of public space sentiments based on the dataset presented in Figure 3.

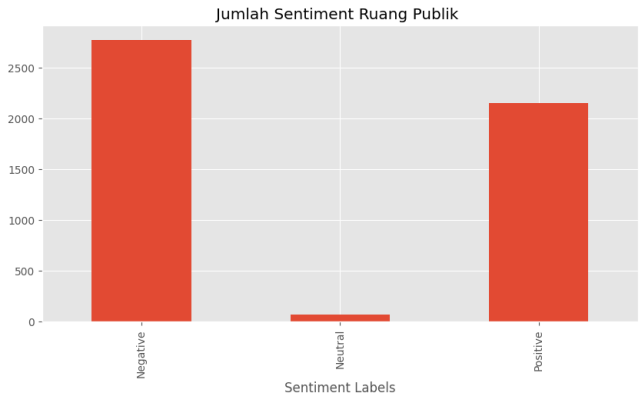

Figure 2. Sentiment Data Bar-Chart

After passing the preprocessing stage, a clean public space opinion dataset will be obtained. To determine sentiment analysis, the clean data is converted into a vector with TF-IDF word embedding. Each model in the algorithm is calculated for accuracy. Experiments conducted on the three Machine Learning algorithms with the extraction feature "TF-IDF', obtained training accuracy

models with good classification are SVM. Here is the accuracy comparison with three different ratios of training and testing data.

Table 4. Accuracy Comparison

Ratio	Accuracy
70:30	92.8%
80:20	91,6%
90:10	91,8%

The results of this accuracy are presented in Figure 10 confusion matrix of SVM.

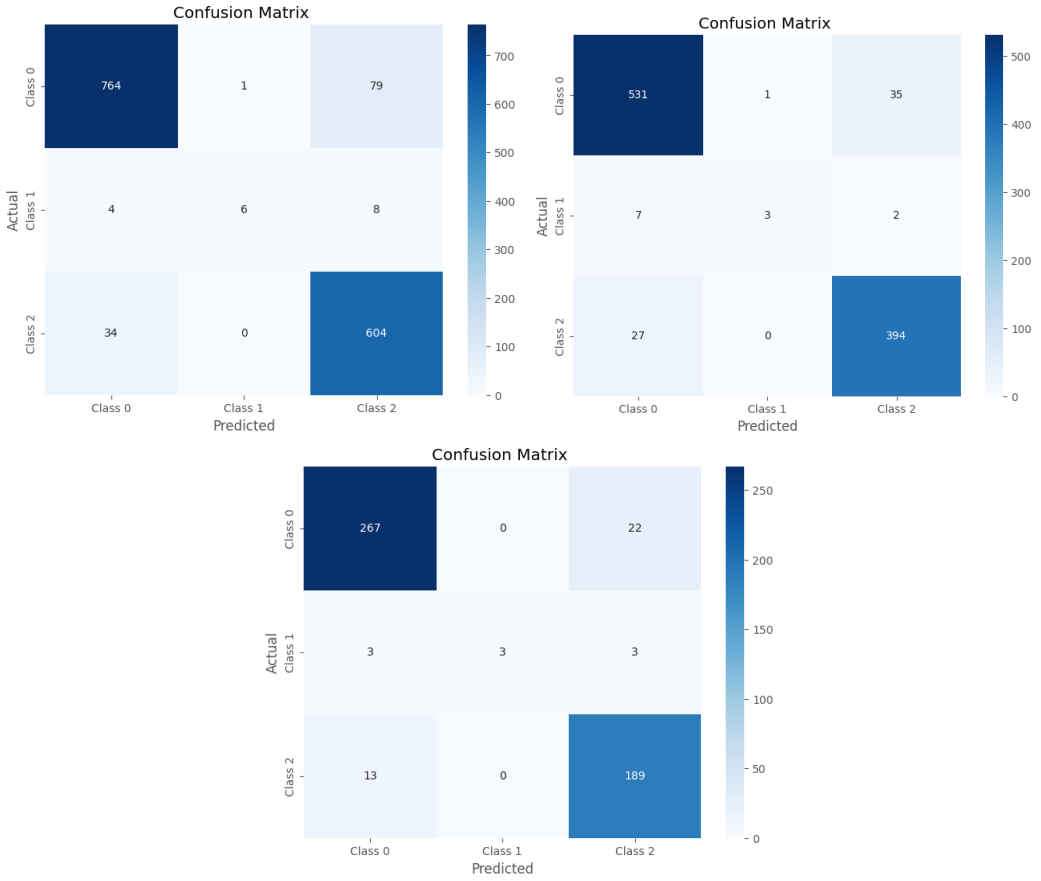

Figure 3. Confusion Matrix

D. Conclusion

Text classification accuracy using the SVM method and TF-IDF data weighting by testing the accuracy directly against the model created with 70:30 split data obtained an accuracy of 91.6% then at 80:20 split data obtained an accuracy of 92.8% and testing with 90:10 split data resulted in an accuracy of 91.8%. It is important to note that technical abbreviations are defined the first time they are used to ensure understanding. The language used throughout is objective, value neutral, and free from ornamental language or filler words. A formal structure is followed, with a logical progression between statements and a balanced argument that avoids bias. Appropriate vocabulary is used where relevant. The text follows grammatical correctness and consistent formatting features and citation style are used.

In this study, there is a difference between the accuracy tested directly on the model and the accuracy using the comparison chart. This can be an opportunity for future research. Further research can use other ML algorithms and methods and can also use deep learning methods with a larger number of datasets. In the future, intent sentiment analysis can be done using a larger number and variety of data sets. The methods used can also use various other Machine Learning methods.

D. References

[1] D. Darwis, E. S. Pratiwi, and A. F. O. Pasaribu, "Penerapan Algoritma Svm Untuk Analisis Sentimen Pada Data Twitter Komisi Pemberantasan Korupsi Republik Indonesia," *Edutic - Sci. J. Informatics Educ.*, vol. 7, no. 1, pp. 1–11, 2020, doi: 10.21107/edutic.v7i1.8779.

[2] S. Styawati, N. Hendrastuty, and A. R. Isnain, "Analisis Sentimen Masyarakat Terhadap Program Kartu Prakerja Pada Twitter Dengan Metode Support Vector Machine," *J. Inform. J. Pengemb. IT*, vol. 6, no. 3, pp. 150–155, 2021, doi: 10.30591/jpit.v6i3.2870.

[3] F. Sodik and I. Kharisudin, "Analisis Sentimen dengan SVM , NAIVE BAYES dan KNN untuk Studi Tanggapan Masyarakat Indonesia Terhadap Pandemi Covid-19 pada Media Sosial Twitter," *Prisma*, vol. 4, pp. 628–634, 2021.

[4] Adhitya Karel Maulaya and Junadhi, "Analisis Sentimen Menggunakan Support Vector Machine Masyarakat Indonesia Di Twitter Terkait Bjorka," *J. CoSciTech (Computer Sci. Inf. Technol.*, vol. 3, no. 3, pp. 495–500, 2022, doi: 10.37859/coscitech.v3i3.4358.

[5] M. R. Fahlevvi, "Analisis Sentimen Terhadap Ulasan Aplikasi Pejabat Pengelola Informasi Dan Dokumentasi Kementerian Dalam Negeri Republik Indonesia Di Google Playstore Menggunakan Metode Support Vector Machine," *J. Teknol. dan Komun. Pemerintah.*, vol. 4, no. 1, pp. 1–13, 2022, doi: 10.33701/jtkp.v4i1.2701.

[6] I. S. K. Idris, Y. A. Mustofa, and I. A. Salihi, "Analisis Sentimen Terhadap Penggunaan Aplikasi Shopee Mengunakan Algoritma Support Vector Machine (SVM)," *Jambura J. Electr. Electron. Eng.*, vol. 5, no. 1, pp. 32–35, 2023, doi: 10.37905/jjeee.v5i1.16830.

[7] N. Lutfianti, "Penerapan Sentimen Analisis Dengan Algoritma SVM Dalam Tanggapan Netizen Terhadap Berita Resesi 2023," *Sisfotenika*, vol. 13, no. 1, pp. 53–64, 2023.

[8] M. K. Anam, B. N. Pikir, M. B. Firdaus, S. Erlinda, and Agustin, "Penerapan Na¨ıve Bayes Classifier, K-Nearest Neighbor (KNN) dan.pdf," *Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, vol. 21, no. 1. p. 139~150, 2021.

[9] B. Edgar Maulana Thoriq,Rahayudi and D. E. Ratnawati, "Analisis Sentimen Opini Publik pada Media Sosial Twitter terhadap Vaksin Covid-19 menggunakan Algoritma Support Vector Machine dan Term Frequency-Inverse Document Frequency," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 12, pp. 5349–5355, 2021.

[10] R. Ramlan, N. Satyahadewi, and W. Andani, "Analisis Sentimen Pengguna Twitter Menggunakan Support Vector Machine Pada Kasus Kenaikan Harga

BBM," *Jambura J. Math.*, vol. 5, no. 2, pp. 431–445, 2023, doi: 10.34312/jjom.v5i2.20860.

[11] R. Rahmaddeni, M. K. Anam, Y. Irawan, S. Susanti, and M. Jamaris, "Comparison of Support Vector Machine and XGBSVM in Analyzing Public Opinion on Covid-19 Vaccination," *Ilk. J. Ilm.*, vol. 14, no. 1, pp. 32–38, 2022, doi: 10.33096/ilkom.v14i1.1090.32-38.

[12] P. Arsi and R. Waluyo, "Analisis Sentimen Wacana Pemindahan Ibu Kota Indonesia Menggunakan Algoritma Support Vector Machine (SVM)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 1, p. 147, 2021, doi: 10.25126/jtiik.0813944.

[13] E. S. R. Br.Situmorang, M. K. Anam, R. Rahmaddeni, and A. N. Ulfah, "Perbandingan Algoritma Svm Dan Nbc Dalam Analisa Sentimen Pilkada Pada Twitter," *CSRID (Computer Sci. Res. Its Dev. Journal)*, vol. 13, no. 3, p. 169, 2021, doi: 10.22303/csrid.13.3.2021.169-179.

[14] Z. Alhaq, A. Mustopa, S. Mulyatun, and J. D. Santoso, "Penerapan Metode Support Vector Machine Untuk Analisis Sentimen Pengguna Twitter," *J. Inf. Syst. Manag.*, vol. 3, no. 2, pp. 44–49, 2021, doi: 10.24076/joism.2021v3i2.558.