JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 1, Desember 2023, hlm. 59 - 66 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v10i1.2645
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

SENTIMENT ANALYSIS OF PUBLIC OPINIONS TOWARDS TELKOM UNIVERSITY POST PANDEMIC

Anindya Prameswari Putri Djakaria[1*], Oktariani Nurul Pratiwi [1], Hanif Fakhrurroja [1]
[1]Information System, Telkom University
[1]*email*: *anindyappdj@student.telkomuniversity.ac.id

Abstract: Twitter, as a social media platform, has rapidly grown as a means for people to express their opinions and thoughts on various topics, including education. The number of Twitter users surged to 10.645.000 in 2020, with a significant increase during the pandemic. Telkom University, as a private institution of higher education in Indonesia, has become one of the topics of discussion on Twitter. Users' opinions about Telkom University vary, ranging from positive to negative. To gain deeper insights into public view, sentiment analysis is essential. The analysis follows the Knowledge Discovery in Databases (KDD) process, utilizing the Naive Bayes classification algorithm. The evaluation results indicate the best accuracy achieved with an 80:20 data split, resulting in an accuracy rate of 82.05%, precision of 82.3%, recall of 82.05%, and F1-Score of 82.08%. The Naïve Bayes model demonstrates good performance for sentiment analysis of public views regarding Telkom University on Twitter.

Keywords: naïve bayes; sentiment analysis; twitter; telkom university.

Abstrak: Media sosial Twitter berkembang pesat sebagai sarana masyarakat berekspresi untuk menuangkan opini dan pikiran mereka mengenai topik apapun, termasuk pendidikan. Pengguna Twitter meningkat tajam hingga 10.645.00 pengguna pada tahun 2020 dan terus meningkat selama pandemi. Telkom University sebagai perguruan tinggi menjadi salah satu topik yang dibicarakan yang berkaitan dengan pendidikan. Pendapat mengenai Telkom University yang diungkapkan oleh pengguna Twitter beragam, baik positif maupun negatif. Analisis sentimen diperlukan untuk memahami pandangan publik lebih mendalam. Digunakan tahapan *Knowledge Discovery in Databases* dan algoritma klasifikasi Naïve Bayes dalam analisis ini. Hasil evaluasi menunjukkan akurasi paling baik dicapai dengan rasio data 80:20, dengan nilai akurasi sebesar 82.05%, nilai presisi sebesar 82.3%, nilai *recall* sebesar 82.05%, dan nilai *F1-Score* sebesar 82.08%. Model klasifikasi Naïve Bayes memiliki performa baik untuk analisis sentimen pandangan publik di Twitter mengenai Telkom University.

Kata kunci: analisis sentimen; naïve bayes; twitter; telkom university.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 1, Desember 2023, hlm. 59 - 66 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v10i1.2645
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

INTRODUCTION

Twitter has become the number one platform for people to express their feelings, opinions, views, and real-time events through Live Tweets. Twitter is a social media platform for computer-based online communication [1]. Various organizations and businesses are interested in Twitter data to determine different people's opinions about their products and events. Twitter is also used to understand different people's opinions about political events, movies, and more [2]. The number of Twitter users significantly increased to reach 10,645,000 users in 2020, which was the year the pandemic emerged. It then continued to rise, reaching 18.45 million users in 2022 [3].

According to Permatasari, Twitter users increased by 34% in the second quarter of 2020, with the platform becoming a means of expression for people regarding their activities during the pandemic in Indonesia, including those relate to education [4]. Education in Indonesia is divided into three types: academic education (bachelor, master, and doctoral degree), professional/specialist education, and vocational education (diploma/applied bachelor's degree) [5]. Academic education focuses on the mastery and development knowledge and technology. Generally, academic education is provided by universities and institutes [6].

Telkom University is one of the private higher education institutions in Indonesia, founded in 1994. Throughout its existence, Telkom University has achieved various outstanding achievements. In 2022. The university obtained 20 awards from various organizers. These achievements have become a topic of discussion among both the university's internal stakeholders and the general public. Social media, especially on Twitter, is no exception, with many users expressing their opinions about Telkom University.

With the discovery of various opinions, both positive and negative, expressed by Twitter users about Telkom University, the author finds it necessary to conduct sentiment analysis on the public's opinions about the university. Sentiment analysis is a process conducted to determine opinions, emotions, and attitudes reflected through text, usually classified into positive and negative opinions. This process is carried out to gather and examine public views on specific products or topics [7]. In order to enhance the precision of sentiment analysis, machine learning methods such as the Naïve Bayes classification algorithm are used, which can accelerate the automated evaluation of data [8]. To perform sentiment analysis, several stages are required based on the Knowledge Discovery in Database (KDD) methos. These stages begin with data selection, preprocessing, transformation, data mining, and evaluation [9].

By using the popular classification method for sentiment analysis, namely Naïve Bayes classification. This method is used to categorize or assess opinions or tendencies towards a particular issue or object, determining whether it falls under the positive or negative category [10]. Several studies have been conducted related to sentiment analysis using Naïve Bayes because it is considered one of the methods that are

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 1, Desember 2023, hlm. 59 - 66 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v10i1.2645
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

easy to understand and still yield good accuracy. In a study, it was mentioned that this method is often used because it only requires a small amount of training data to determine the estimated parameters needed in the classification process [11]. With the issues outlined earlier, this forms the foundation for the author to conduct research with the aim of determining the public sentiment towards Telkom University through Twitter using the Naïve Bayes method.

METHOD

This research utilizes the text mining method. This method is employed because the data obtained is in the form of text. By using text mining, valuable and quality information can be derived from the data being used. This research uses the Naïve Bayes classification for sentiment analysis as it is considered as one of the popular, simple, and easily understandable classification algorithms that can produce reasonably accurate classifications [12].

$$P(R|S) = \frac{P(R)P(S|R)}{P(S)} \qquad (1)$$

Description:

R	:	Data with unknown class
S	:	Hypothesis on data R which is a special class
P(R|S)	:	The probability value of R based on codition S
P(R)	:	The probability value of R
P(S|R)	:	The probability value of S based on condition R
P(S)	:	The probability value of S

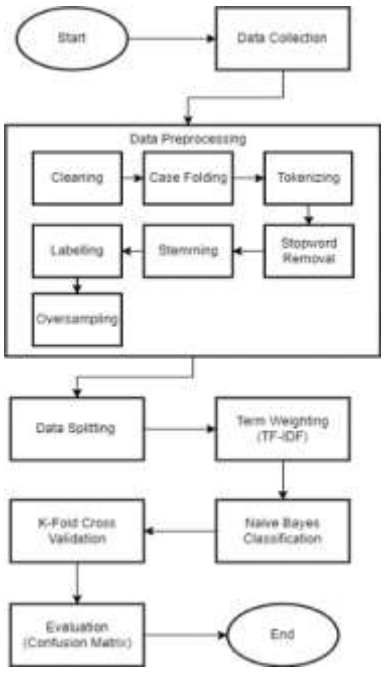

Image 1 Problem-Solving Systematics

Image 1 illustrates the stages of systematic problem-solving. There are several stages in the problem-solving systematics: data collection, text preprocessing, data splitting, term weighting, Naïve Bayes classification, K-Fold cross validation, and evaluation using confusion matrix. In data collectinig stage, data will be obtained by scraping tweets related to Telkom University on Twitter using predefined keywords.

Once the data is collected, preprocessing will be performed, involving data cleaning, case folding, tokenization, stopword removal, stemming, and manual data labeling. The data will be divided into two labels, positive and negative. After mapping the data according to the labels, oversampling will be conducted.

On the next step, data will be split into tranining and testing data and term weighting will be applied using Term Frequency-Inverse Document Frequency technique. Following those,

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 1, Desember 2023, hlm. 59 - 66 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v10i1.2645
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

the divided data will undergo data classification using the Naïve Bayes algorithm. The subsequent stage is ensuring the reliability and validity of the data division, K-Fold cross validation will be employed. Finally, the model will be evaluated using the confusion matrix to examine the performance results of the model.

RESULT AND DISCUSSION

Data scraping is performed on Twitter by searching for relevant tweets according to the research needs. In this study, relevant keywords like 'telkom university', 'telkom univ', 'universitas telkom', 'kuliah telkom', 'kuliah telyu', 'telyu', 'tel u', and 'tel-u' were used, with a date range limitation from January 1, 2022, to July 16, 2023. To perform data scraping, the TwitterSearchScraper library is used, utilizing the Python programming langauge. The data is extracted based on the specified keywords, resulting in a total of 3026 data. Subsequently, the data is manually filtered to remove irrelevant and neutral sentiment tweets. After this filtering process, 588 relevant data are obtained and ready for use.

After obtaining the data, manual labeling was done by matching the content of tweets with the appropriate sentiment. With this technique, out of the total 588 data, 390 data were labeled as positive sentiment and 198 data were labeled as negative sentiment.

Image 2 Data Sentiment Bar Chart

After preprocessing and data labeling, oversampling is performed to increase the number of samples from the minority class (in this case, the negative sentiment class) by creating new examples or repeating some existing examples to balance it with the majority class (positive sentiment class). The oversampling is done using the reandom oversampler method. As a result, the total usable data becomes 780, with 390 data each for positive and negative sentiment after oversampling.

After going through the stages of data collection, processing, data splitting, and term weighting, the data that was divided into training and testing data will be used for model creation using the Naïve Bayes algorithm. The training data is utilized to train the model with the weighting results and predefined labels given to the Naïve Bayes algorithm, while the testing data is used to measure the classifier's performance in classification with accurate predictions. Here is the accuracy comparison with three different ratios of training and testing data.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 1, Desember 2023, hlm. 59 - 66 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v10i1.2645
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

Table 1 Accuracy Comparisons

Ratio	Accuracy
60:40	75.9%
70:30	78.2%
80:20	807%

Based on Table 1, the Naïve Bayes classification model with 80:20 ratio has the highest performance with an accuracy of 80.7%. With this ratio, the number of training data used is 624 data and 156 data are used for testing data.

Next, K-Fold cross-validation is performed using the scikit-learn (sklearn) library with the modules 'GaussianNB' and 'cross_val_score'. Using 5-fold, the data is divided into 5 subsets (folds) and the model is trained and evaluated 5 times.

Table 2 K-Fold Cross Validation results with K=5

Fold	Accuracy		
	60:40	70:30	80:20
1	76.5%	72.7%	70.4%
2	72.3%	79.8%	82.4%
3	79.7%	73.3%	80%
4	83.8%	73.3%	74.4%
5	67.7%	70.6%	78.2%
Average	76%	73.9%	77%

Based on Table 2, the results from the five folds show that the 80:20 ratio produces the highest cross-validation score with an average of 75%. By using the confusion matrix method, the behavior of the classification model can be evaluated or visualized. The evaluation is performed using the testing data.

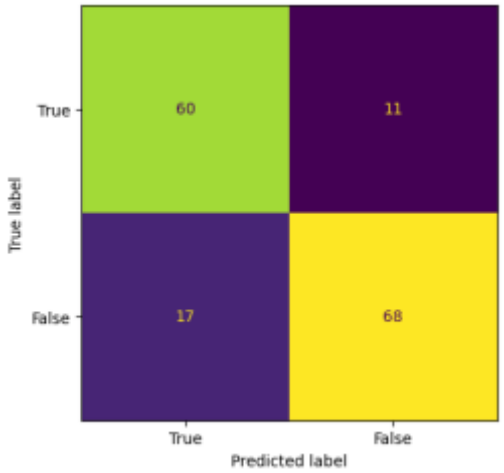

Image 3 Confusion Matrix for 80:20 ratio

Based on Image 3, the confusion matrix shows that out of the total 156 testing data, 63 data are classified as true positives, 11 data as false positives, 63 data as true negatives, and 19 data as false negatives. By using the confusion matrix, the average values of accuracy, precision, recall, and F1-score can be obtained.

Table 3 Evaluation Results

Ratio	60:40	70:30	80:20
Accuracy	74%	80.3%	82.05%
Precision	74%	81.1%	82.3%
Recall	74.8%	80.3%	82.05%
F1-Score	74%	80.2%	82.08%

Based on Table 3, the performance evaluation results with an 80:20 ratio show the highest accuracy value, indicating that the Naïve Bayes classification model with this ratio is considered good.

To display the results of the modeling implementation, WordCloud is used to show the list of words used in the collected tweets. The more frequently these words are used, the larger their size will be in WordCloud.

JURTEKSI (Jurnal Teknologi dan Sistem Informasi) ISSN 2407-1811 (Print)
Vol. X No 1, Desember 2023, hlm. 59 - 66 ISSN 2550-0201 (Online)
DOI: https://doi.org/10.33330/jurteksi.v10i1.2645
Available online at http://jurnal.stmikroyal.ac.id/index.php/jurteksi

Image 4 Positive Sentiments WordCloud

Based on Image 4, which is the result of the WordCloud for data labeled as positive sentiment, the words are found in tweets that provide information about scholarships offered by Telkom University, Twitter users' admissions in university selection, and their desire to study at Telkom University as a good private university.

Image 5 Negative Sentiments WordCloud

On the other hand, in Image 5, the result of the WordCloud for data labeled as negative sentiment include words that can be found in tweets discussing tuition fees, traffic congestion on the way to Telkom University, and the hot weather in the university area.

After obtaining 588 relevant data representing public opinions on Telkom University from the Twitter social media platform, the public tends to hace a fairly positive opinion about the university. Positive opinions are frequently found in tweets discussing scholarships and the discussion of Telkom University as a reputable private higher education institution. However, negative opinions about this university are also present, as seen in tweets assuming that the employment prospects for Telkom University graduates are limited to being technicians fixing WiFi networks. Therefore, in order to mitigate unfavorable opinions such as assumptions about job prospects, Telkom University could publish information about its graduates having diverse employment opportunities. Additionally, Telkom University could periodically monitor Twitter to gauge public opinions for the purpose of institutional evaluation.

CONCLUSION

Based on the sentiment analysis results of public opinions on Twitter about Telkom University using Naïve Bayes classification, it was found that out of the total 588 data, 66.3% showed as positive sentiments, while only 33.6% showed as negative sentiments. This result indicates that on Twitter, the public's opinion about Telkom University tends to be positive. The process involved data selection through data scraping, followed by data preprocessing. Subsequently, data transformation was carried out. Followed by Naïve Bayes classification, and the evaluation phase was performed, which included K-Fold Cross Validation and the use of a confusion matrix to obtain evaluation results of the model's performance. The

evaluation results showed that the best accuracy was achieved with 80:20 data ratio after oversampling with the total of 780 data, 624 data for training and 156 data for testing. With that ratio, the evaluation results showed an accuracy of 82.05%, precision of 82.3%, recall of 82.05%, and F1-score of 82.08%. Based on these findings, the model's average performance with this ratio exceeds 80%, suggesting that the Naïve bayes classification algorithm demonstrates a strong performance.

BIBLIOGRAPHY

[1] A. Karami, M. Lundy, F. Webb and Y. K. Dwivedi, "Twitter and Research: A Systematic Literature Review Through Text Mining," *Social Science Research Grant Program*, vol. 8, pp. 67698-67717, 2020.

[2] S. Saha, J. Yadav and P. Ranjan, "Proposed Approach for Sarcasm Detection in Twitter," *Indian Journal of Science and Technology*, vol. 10, no. 25, 2017.

[3] D. Berniawan, Amri and Tinaliah, "Implementasi Algoritma Naïve Bayes Untuk Klasifikasi Sentimen Pengguna Twitter Terhadap KEMKOMINFO Di Indonesia," in *2nd MDP Student Conference (MSC) 2023*, 2023.

[4] N. Permatasari, R. Yosral and C. F. Annisa, "Twitter Analysis About Online Education During COVID-19 Pandemic In Indonesia," in *Seminar Nasional Official Statistics*, 2020.

[5] W. Sihaloho, R. U. Pratiwi, I. P. Sari, I. Q. Aini, Z. Yunita and T. Winanda , "Perkembangan Konsep Pendidikan dan Klasifikasi Pendidikan," *Jurnal Dirosah Islamiyah*, vol. 5, no. 3, pp. 754-762, 2023.

[6] R. Ambarwati and D. C. U. Lieharyani, "Visualisasi Data Tweet di Sektor Pendidikan Tinggi Pada Saat Masa Pandemi," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 1, pp. 116-123, 2022.

[7] M. Cindo, D. P. Rini and Ermatita, "Literatur Review: Metode Klasifikasi pada Sentimen Analisis," *Seminar Nasional Teknologi Komputer & Sains (SAINTEKS)*, pp. 66-70, 2019.

[8] J. Singh, G. Singh and R. Singh, "Optimization of sentiment analysis using machine learning classifiers," *Human-centric Computing and Information Sciences*, 2017.

[9] S. Z. Harahap and A. Nastuti, "Teknik Data Mining untuk Penentuan Paket Hemat Sembako dan Kebutuhan Harian dengan Menggunakan Algoritma FP-Growth (Studi Kasus di Ulfamart Lubuk Alung)," *Informatika: Jurnal Ilmiah Fakultas Sains dan Teknologi, Universitas Labuhanbatu*, vol. 7, no. 3, pp. 111-119, 2019.

[10] A. V. Sudiantoro and E. Zuliarso, "Analisis Sentimen Twitter Menggunakan Text Mining dengan Algoritma Naive Bayes Classifier," *Dinamika Informatika*, vol. 10, no. 2, pp. 69-73, 2018.

[11] A. Saleh, "Implementasi Metode Klasifikasi Naive Bayes dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga," *Citec Journal*, vol. 2, no.

3, pp. 207-217, 2015.

[12] L. A. Muhaimin, O. N. Pratiwi and R. Y. Fa'rifah, "Klasifikasi Soal Berdasarkan Kategori Topik Menggunakan Metode Algoritma Naive Bayes dan Algoritm C4.5," in *eProceedings of Engineering*, 2023.