JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 4, Oktober 2022, Page 2050-2055
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i4.4567

Sentiment Analysis of Telkom University as the Best BPU in Indonesia Using the Random Forest Method

Irfan Budi Prakoso*, Donni Richasdy, Mahendra Dwifebri Purbolaksono

Fakultas Informatika, Program Studi Informatika, Telkom University, Bandung, Indonesia
Email: [1,*]irfanbudi@students.telkomuniversity.ac.id, [2]donnir@telkomuniversity.ac.id, [3]mahendradp@telkomuniversity.ac.id
Email Penulis Korespondensi: irfanbudi@students.telkomuniversity.ac.id

Abstract−In this day and age, social media has become a necessity for every human being. By using social media networks, users can easily exchange information, especially on linkedin social media. Linkedin is a social media network that can search for information openly, mainly used for professional networking. It will be easier and more practical to connect with professionals worldwide. Like identity, LinkedIn is often used as a medium to introduce yourself or your business to potential colleagues or companies for various purposes. Social media networks are often used to deliver information in various institutions at State Universities (PTN) and Private Universities (PTS). For example, it conveys information about state and private universities' achievements (PTS) achievements. Telkom University uses Linkedin to convey the achievements that have been achieved. This triggers the public to see posts that are positive, negative, or neutral. This study aims to conduct a sentiment analysis about Telkom University which has become the best private university in Indonesia, based on opinions submitted on LinkedIn social media. The process carried out in this study is to process all opinion data about Telkom University, which is the best private university in Indonesia, from Linkedin and then classification using the Random Forest method based on the categories of positive, neutral, and negative sentiments. Sentiment analysis results that have been obtained using the Random Forest classification method are 92.85% accuracy, 83.33% precision, 91.67% recall, and 84.13% F1-score%.

Keywords: LinkedIn; Random Forest; Telkom University; Social Media; Sentiment Analysis

1. INTRODUCTION

In this day and age, social media has become a necessity for every human being. By using social media, users can easily share information. Linkedin is a social media network that can search for information openly. Mainly used for professional networking, connecting with professionals worldwide becomes more accessible and convenient. Like an identity, LinkedIn is often used to inform yourself or your line of business to potential partners or industries for various purposes. The social media network is often used in multiple PTN and PTS institutions as a medium for delivering information. For example, they were conveying information about the achievements of state universities (PTN) and private universities (PTS).

Telkom University is a private university supported by the Telkom Education Foundation (YPT), consisting of 7 faculties with 50 study programs, 800 lecturers, and 28,789 students. Telkom University is ranked as one of the best private universities in Indonesia, as reported by the Times Higher Education (THE) and World University Rankings (WUR) ranking sites[1]. This led to the slogan "Best Private Higher Education (PTS) Number 1" for the Telkom University brand. In addition to having a quality education, the brand aspect is one of Telkom University's focuses to be improved regularly.

In this study, sentiment analysis was used to assess the brand's positive, negative, and neutral aspects from the results of data collection on the LinkedIn social media network based on posts made by Telkom University. Based on this, the Random Forest method is used, a classification algorithm that decomposes data randomly into a Decision Tree. This method was chosen in this study because it can handle large amounts of information very accurately and is not affected by missing information. The author also uses the Term Frequency — Inverse Document Frequency (TF-IDF) extraction feature, which is this calculation is carried out for each word and gives each a weighted value. This method is also famous for being practical and has optimal results[2]. This research was conducted to find out the results of working sentiment analysis from the profile of Telkom University on the LinkedIn social network using the Random Forest method. Random Forest is an evolution of the Decision Tree method using multiple Decision Trees. Each Decision Tree is trained with an individual sample, and each attribute is split into a tree selected from a subset of random features.[3].

In a study by Boma B. B et al. in 2021[4]. This research used random forest and TF-IDF methods. This study aims to analyze guest sentiment at the Purwolkerto hotel by applying the random forest method and changing guest input data from the textual form into quantitative form, inverse document frequency (TF-IDF method). The dataset used in this study includes reviews of hotel guests in Purwokerto downloaded from the TripAdvisor.co.id website. A total of 1166 reviews from various hotels were successfully uploaded. This study's results indicate that the model's accuracy reaches 87.23%. However, if the rooting process is not carried out, the model's accuracy is only 87.01%.

Another study was conducted by Ragil D. et al. In 2021[5]. This study was conducted to determine the sentiment of public tweets on the official Twitter account of the DKI Jakarta Provincial Government during the COVID-19 pandemic. The data used in this study were obtained from the Twitter social network. 1028 data streams containing questions on tweets containing particular words or mentioning the handle @dkijakarta are separated into three classes based on sentiment: negative, neutral, and positive. This is done by using different classifiers,

Submitted: **27/07/2022**; Accepted: **31/08/2022**; Published: **25/10/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 4, Oktober 2022, Page 2050-2055
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i4.4567

namely Random Forest with an accuracy of 75.81%, Naive Bayes with an accuracy of 75.22%, and SVM with an accuracy of 77.58%. Dynamic analysis was carried out on tweets whose results were 8.8%, 83.6%, and 7.6% for negative, neutral, and positive, respectively.

Subsequent research was conducted by Adrian R et al. In 2021[6]. This study aims to analyze public sentiment related to PSBB using the Twitter social media platform by analyzing 466 tweet data. The data is separated into seven parts for training and three positions for testing, with a ratio of 7 to 3. The data is then processed through 2 different classification algorithm methods for comparison: the SVM classification method and the Random Forest.

The following research was conducted by Ahmad S et al. In 2022[7]. This research aims to find out public opinion on electric cars. Whether the opinion is more positive or negative and to determine the accuracy value, the AUC is from using the Support Vector Machine method and the Particle Swarm Optimization feature selection in RapidMiner Studio Software. In this study, it can be seen that 94.25% of Twitter users agree, and 5.75% of Twitter users disagree with the presence of electric cars. The Particle Swarm Optimization feature selection on the support vector machine method to analyze public sentiment about electric cars can increase the accuracy and AUC values. Where the accuracy value was originally 82.51% to 86.07%, there was an increase of 3.56%. While the AUC value was originally 0.844 to 0.862, there was an increase of 2.13%.

Further research was conducted by Yusril A et al. In 2022 [8]. This study aims to analyze sentiment towards the community [there is a vaccination program using the Sinovac Vaccine. This study used 1500 tweets with data divided into two categories, namely positive and negative. The data processing used in this research is by using the TF-IDF Algorithm and balancing the data using SMOTE. The model created will be trained with the Random Forest Classifier Algorithm and validated using K-fold Cross Validation and Confusion Matrix. The results of this study are public sentiment toward Sinovac Vaccination is positive, and the model can predict the sentiment of a tweet with an accuracy of 79% and a Precision value of 85%, a Recall of 90% and an F1 score of 88%.

2. RESEARCH METHODOLOGY

2.1 Research Stages

The following is a system design built, as shown in Figure 1 to analyze sentiment using the Random Forest method and the TF-IDF extraction feature.

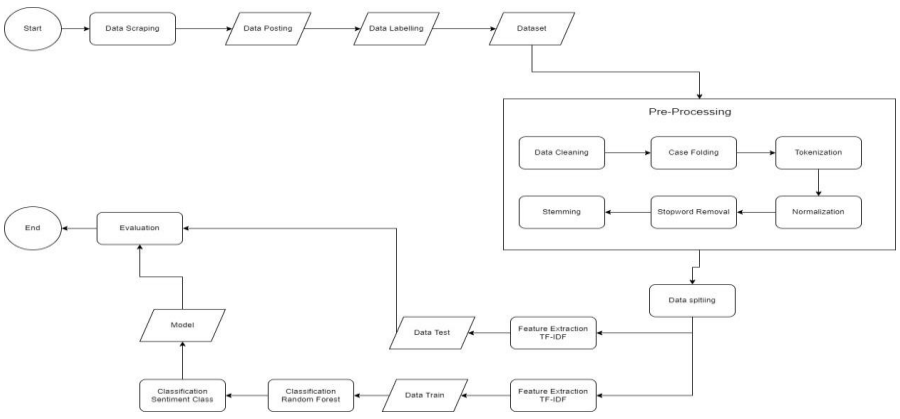

Figure1. Sentiment Analysis Flowchart

Based on Figure 1 there is a flow of the stages to be carried out. The first stage is to collect uploaded data from the Telkom University account on LinkedIn social media to be used as a dataset for the sentiment analysis model using the Random Forest method. The second stage is labelling the uploaded data based on 3 categories: positive, negative, and neutral. Then proceed with doing several pre-processing of the dataset that has been labelled, such as data cleaning, case folding, tokenization, normalization, stopwords removal, and stemming. The fourth stage performs data splitting to be used as train data and test data with data ratios of 80:20, 75:25, and 70:30. Then weighting each word in the dataset using the TF-IDF. Then do the model training based on the data training with the Random Random method. The last stage is to evaluate the performance of the model that has been made based on the calculation of the confusion matrix.

2.2 Data Collection

On Linkedin, there is a Telkom University account. The data used in this study was obtained from the upload of the Telkom University account. Data collection is obtained by manual processes such as taking a one-by-one description of each post in the Telkom University account. They were taking a one-by-one description of each post based on keywords related to Telkom University to become PTS No.1 or the best such as "best", "No.1", "PTS No.1", and "best PTS". This data will be used as a training model for sentiment analysis. Below is a word cloud showing some of the words that appear most often from the data obtained:

Submitted: **27/07/2022**; Accepted: **31/08/2022**; Published: **25/10/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 4, Oktober 2022, Page 2050-2055
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i4.4567

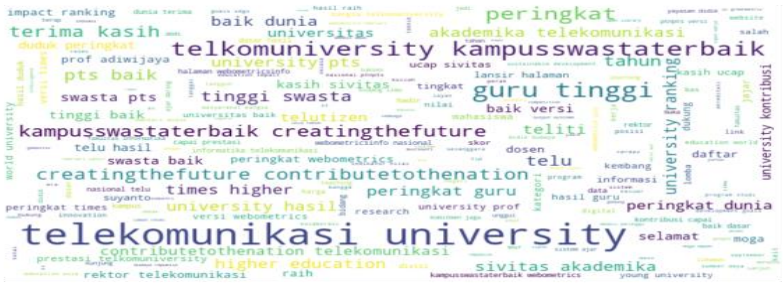

Figure 2. Wordcloud Most Frequently Occurring Words

2.3 Data Labeling

This process collects data which is then labelled with 3 labels. That is positive, negative, and neutral.

Table 1. Data Labeling

Post	Keyword
Lagi, Telkom University jadi PTS Terbaik No.1 di Indonesia tahun 2020 https://lnkd.in/gknuRFv	Positive
BRIN Kunjungi Fasilitas Laboratorium Elektronika dan Telekomunikasi Terbaik di Tel-U https://lnkd.in/dGgEjthj	Neutral
Selamat Hari Raya Idul Adha 1442 H Semoga Idul Adha kali ini mampu meningkatkan keimanan dan ketakwaan kita dalam beragama dan semakin bersabar dalam menghadapi musibah Covid-19 Semoga kita semua dalam lindungan Allah Swt #telkomuniversity #kampusswastaterbaik #creatingthefuture	Negative

2.4 Preprocessing

Before the data is processed for sentiment analysis, it will be preprocessed. Preprocessing is a step taken before the data is analyzed to see the words' sentiments. Many preprocessing steps can be done, including case folding, data cleaning, tokenization, normalization, stop keywords, stemming, and untokenizing data. This study performs several preprocessing steps, which are outlined below.:

1. Case Folding, Case folding is converting uppercase letters into lowercase letters (lowercase). Make all letters in the data be entered uniform [9].
2. Data Cleaning, Data Cleaning is the process of removing noise. This process removes Hashing, username, memorable characters, and numbers. The table below is an example of the implementation of data cleaning.
3. Tokenization, Tokenizing is a process of breaking a sentence into words, terms, or symbols [10].
4. Normalization is converting non-standard words into standard ones and converting abbreviated words into actual words.
5. Stopwords Removal, After tokenization, then Stopwords Removal is the process of removing non-topic words that are considered unimportant, in this case, the words included in the stoplist, one of which is conjunctions such as "and", "or", "which ", etc. Processes help reduce irrelevant features in the data[11].
6. Stemming is the process of separating the affixes, namely prefix, infix, suffix, and confix (a combination of prefix and suffix) derived words into essential words. With stemming, variations of words with the same root will be considered the same token (feature).[11].
7. Untokenziation, combining words that have been separated into one again.

Table 2. Pre-processing

Step	Post
Original Data	Lagi, Telkom University jadi PTS Terbaik No.1 di Indonesia tahun 2020 https://lnkd.in/gknuRFv
Data Cleaning	Lagi Telkom University jadi PTS Terbaik No di Indonesia tahun
Case Folding	lagi telkom university jadi pts terbaik no di indonesia tahun
Tokenization	‘lagi’, ‘telkom’, ‘university’, ‘jadi’, ‘pts’, ‘terbaik’, ‘no’, ‘di’, ‘indonesia’, ‘tahun’
Normalization	‘lagi’, ‘telkom’, ‘university’, ‘jadi’, ‘pts’, ‘baik’, ‘no’, ‘di’, ‘indonesia’, ‘tahun’
Stopword Removal	‘lagi’, ‘telkom’, ‘university’, ‘jadi’, ‘pts’, ‘baik’, ‘no’, ‘di’, ‘indonesia’, ‘tahun’
Stemming	‘lagi’, ‘telkom’, ‘university’, ‘jadi’, ‘pts’, ‘baik’, ‘no’, ‘di’, ‘indonesia’, ‘tahun’

2.5 Random Forest Method

The random forest method is self-predictive by calculating the majority vote to combine the results from many decision trees, each of which is made at random. Bootstrap templates are used to create multiple predictive trees.

Submitted: **27/07/2022**; Accepted: **31/08/2022**; Published: **25/10/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 4, Oktober 2022, Page 2050-2055
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i4.4567

The Random Forest method specifically uses bootstrap templates to create multiple trees that can be used to predict future data. The Random Forest method also uses Predictor-Correctors, a self-prediction way that calculates the average regression[12]. In addition to using decision trees, the Random Forest method also uses Predictor-Correctors and bootstrap templates. The Random Forest method produces the first decision tree to assess the impurity of an attribute value and the information obtained from that attribute. The formula used to calculate the entropy of an attribute value is given by equation 1, while the equation to calculate the information obtained can be found in equation[13].

$$Entropy\ (Y)\ =\ -\sum p\ (c|Y)\ i\ log2\ p\ (c|Y) \quad (1)$$

Y is the set of cases, and p(c|Y) is the ratio of Y values to class c.

$$Information\ Gain\ (Y,a) = Entropy(Y) - \sum_{VeValues} \frac{[Yv]}{[Ya]}\ \ Entropy(Yv) \quad (2)$$

Where the value of (a) is the value of all possible events in case a. Yv is a subclass of Y, and class v is correlated to a. Yes, all values correspond to a.

2.6 Term Frequency–Inverse Document Frequency (TF-IDF)

One way to extract features for sentiment analysis is through TF-IDF [14]. TF-IDF value is calculated by multiplying TF by IDF[15]. This calculation is performed for each word and assigns each a weighted value. The equation in (3) is used to obtain the weights, and the final result is used during the sentiment analysis process. A table shows the development of consequences using TF-IDF.

$$idf_j = log\left(\frac{D}{df_j}\right) \quad (3)$$

a calculation can be done TF-IDF to get the results. Equation (4) is the calculation formula for TF-IDF

$$w_{ij}\ =\ tf_{ij}\ \times\ idf_j \quad (4)$$

The following is a description of the formula that has been described.

$tfij$: The number of occurrences of the term in the document
wij : The weight of the term on the document
D : Total of all documents
$idfj$: Distribution of terms on documents
dfj : Number of documents containing the term

2.6 Performance Measurement

Confusion Matrix is a helpful tool for analyzing classifiers to identify tuples of different classes. When measuring performance using the Confusion Matrix, four terms describe the results of the classification process[15]. The confusion matrix consists of 4 important terms such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN)[16]. In addition, there is a performance evaluation with calculations such as Accuracy, Precision, Recall, and F1-Score. The following is the formula for calculating the performance evaluation.

a. Accuracy
Accuracy is a calculation of how accurately the model can classify data. Accuracy is how close the predicted value is to the actual value. The unit of accuracy uses a percentage (%). Formula (5) is a precision formula.

$$accuracy = \frac{TP\ +TN}{TP\ +FP\ +TN\ +FN} \quad (5)$$

b. Precision
Precision is a calculation of the level of accuracy between the requested information and the estimated results generated by the model. Formula (6) is a precision formula—the exact unit of calculation using percentage (%).

$$precision = \frac{TP}{TP\ +FP} \quad (6)$$

c. Recall
Recall adalah perhitungan seberapa sukses model dalam mengambil informasi. Persamaan (7) adalah rumus untuk menghitung pemulihan perhitungan *Recall*.

$$recall = \frac{TP}{TP\ +FN} \quad (7)$$

d. F1-Score
F1-Score is a performance matrix that considers Recall and Precision. Equation (8) is the formula for calculating F1-Score.

$$F1\ Score = 2 \times \frac{precision\ \times recall}{precision\ +recall} \quad (8)$$

Submitted: **27/07/2022**; Accepted: **31/08/2022**; Published: **25/10/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 4, Oktober 2022, Page 2050-2055
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i4.4567

3. RESULT AND DISCUSSION

In the evaluation phase of this research, there are 3 test scenarios to evaluate the system that has been built. Scenario 1 is a baseline test. Scenario 2 is a comparison of unigram and bigram in sentiment analysis. Scenario 3 is testing the Hyperparameter Tuning GridSearchCV.

3.1 Test Result

3.1.1 Scenario 1 Baseline

The first scenario is to determine the baseline or initial data that will be used for testing the following scenario. Testing this scenario is carried out on post that is already available. The results of the first scenario can be seen in the table below.

Tabel 3. Baseline

Method	Data Ratio	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Baseline	80:20	89.29	75.00	89.91	74.34
	75:25	88.57	79.36	90.30	82.04
	70:30	85.71	75.83	81.90	76.61

The results of testing scenario 1 show that sentiment analysis with a data ratio of 80:20 gets the best results compared to other data ratios. The results were accuracy of 89.29%, the precision of 75.00%, recall of 89.91%, and F1 score of 74.34%. The results of testing scenario 1 will be used for testing scenario 2.

3.1.2 Scenario 2 The Effect of Unigrams and Bigrams on Sentiment Analysis

In scenario 2 testing, we will test the effect of unigram and bigram on the extraction of TF-IDF features on the performance model using the Random Forest method. The following are the results of a comparison between the use of unigrams and bigrams in the model.

Table 4. The Effect of Unigrams and Bigrams on Sentiment Analysis

N-Gram	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Unigram (80:20)	92.85	83.33	91.67	84.13
Bigram (80:20)	89.28	75.00	89.91	74.34

The results of testing scenario 2 show that sentiment analysis with unigram gets the best results compared to bigram. With 92.85% accuracy, 83.33% precision, 91.67% recall, and 84.13% F1-score. With that, the results of scenario 2 testing will be used for scenario 3 testing.

3.1.3 Scenario 3 *Hyperparameter Tuning GridSearchCV*

In testing scenario 3, we will perform a Hyperparameter Tuning GridSearchCV test scenario to compare the parameters of the best Random Forest method using the python library. The parameters to be compared are the Gini parameter and the Entropy parameter. The results of the comparison of the two parameters are as follows.

Tabel 5. Hyperparameter Tuning GridSearchCV

Parameter	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Gini	92.85	83.33	91.67	84.13
Entropy	89.28	81.48	80.56	80.00

The results of testing scenario 3 show that the Gini parameter gets the best results compared to the entropy parameter. With 92.85% accuracy, 83.33% precision, 91.67% recall, and 84.13% F1-score.

3.2 Analysis of Test Result

After conducting trials on 3 scenarios using the Random Forest classification, it can be concluded that each test can affect the results of the Random Forest model. Therefore, testing is carried out according to the dataset obtained along with the sentiment analysis model to get the best results. In testing scenario 1, a baseline scenario test was conducted that compared which proportion of the 80:20 ratio, 75:25 ratio, and 70:30 ratio had the best accuracy. The best results were obtained from the scenario test, namely the baseline with a ratio of 80:20 compared to other data ratios. With the results accuracy of 89.29%, precision of 75.00%, recall of 89.91%, and F1 score of 74.34%. Then in scenario 2 testing, a comparison of unigram and bigram scenarios with the TF-IDF extraction feature was carried out. From the scenario test, unigram gets better results than bigram because it uses unigram, which is a sentence-by-word decomposition, producing more information than what is found in the dataset. In scenario 3, the Hyperparameter Tuning GridSearchCV. That is comparing the Gini parameter with the entropy. The results of the

Submitted: **27/07/2022**; Accepted: **31/08/2022**; Published: **25/10/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 4, Oktober 2022, Page 2050-2055
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i4.4567

scenario 3 test proved that Gini gets the best results compared to the entropy with 92.85% accuracy, 83.33% precision, 91.67% recall, and 84.13% F1-score.

4. CONCLUSION

After conducting trials on 3 scenarios using the Random Forest classification, it can be concluded that each test can affect the results of the Random Forest model. Therefore, testing is carried out according to the dataset obtained along with the sentiment analysis model to get the best results. In testing scenario 1, a baseline scenario test was conducted that compared which proportion of the 80:20 ratio, 75:25 ratio, and 70:30 ratio had the best accuracy. The best results were obtained from the scenario test, namely the baseline with a ratio of 80:20 compared to other data ratios. The results were accuracy of 89.29%, the precision of 75.00%, recall of 89.91%, and F1 score of 74.34%. Then in scenario 2 testing, a comparison of unigram and bigram scenarios with the TF-IDF extraction feature was carried out. From the scenario test, unigram gets better results than bigram because it uses unigram, which is a sentence-by-word decomposition, producing more information than what is found in the dataset. In scenario 3, the Hyperparameter Tuning GridSearchCV scenario is tested. That is, comparing the Gini parameter with the entropy parameter. The results of the scenario 3 test proved that the Gini parameter gets the best results compared to the entropy parameter with 92.85% accuracy, 83.33% precision, 91.67% recall, and 84.13% F1-score. Suggestions for further research are to use a dataset with a more significant number of data than previous studies to find out the amount of data that can affect the model using Random Forest. In addition, further research can test various Random Forest parameters to find out how much influence they have on the sentiment analysis model.

REFERENCES

[1] Public Relations TelkomUniversity, "'Telkom University, PTS Terbaik di Indonesia,'" *telkomuniversity.ac.id*, 2021. https://telkomuniversity.ac.id/telkom-university-pts-terbaik-di-indonesia/ (accessed Nov. 30, 2021).

[2] A. A. Maarif, "Penerapan Algoritma TF-IDF untuk Pencarian Karya Ilmiah," *Dok. Karya Ilm. | Tugas Akhir | Progr. Stud. Tek. Inform. - S1 | Fak. Ilmu Komput. | Univ. Dian Nuswantoro Semarang*, no. 5, p. 4, 2015, [Online]. Available: mahasiswa.dinus.ac.id/docs/skripsi/jurnal/15309.pdf.

[3] R. Supriyadi, W. Gata, N. Maulidah, and A. Fauzi, "Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah," *E-Bisnis J. Ilm. Ekon. dan Bisnis*, vol. 13, no. 2, pp. 67–75, 2020, doi: 10.51903/e-bisnis.v13i2.247.

[4] "View of Analisis Sentimen Pelanggan Hotel di Purwokerto Menggunakan Metode Random Forest dan TF-IDF (Studi Kasus_ Ulasan Pelanggan Pada Situs TRIPADVISOR).pdf." .

[5] R. D. Himawan and E. Eliyani, "Perbandingan Akurasi Analisis Sentimen Tweet terhadap Pemerintah Provinsi DKI Jakarta di Masa Pandemi," *J. Edukasi dan Penelit. Inform.*, vol. 7, no. 1, p. 58, 2021, doi: 10.26418/jp.v7i1.41728.

[6] M. R. Adrian, M. P. Putra, M. H. Rafialdy, and N. A. Rakhmawati, "Perbandingan Metode Klasifikasi Random Forest dan SVM Pada Analisis Sentimen PSBB," *J. Inform. Upgris*, vol. 7, no. 1, pp. 36–40, 2021, doi: 10.26877/jiu.v7i1.7099.

[7] A. Santoso, A. Nugroho, and A. S. Sunge, "Analisis Sentimen Tentang Mobil Listrik Dengan Metode Support Vector Machine Dan Feature Selection Particle Swarm Optimization," vol. 2, no. 1, pp. 24–31, 2022.

[8] N. A. S. N. Muhammad Yusril Aldean, Paradise, "Analisis Sentimen Masyarakat Terhadap Vaksinasi Covid-19 di Twitter Menggunakan Metode Random Forest Classifier (Studi Kasus: Vaksin Sinovac)," vol. 8106, pp. 64–72, 2022.

[9] "Tampilan Analisis Sentimen Tentang Opini Maskapai Penerbangan pada Dokumen Twitter Menggunakan Algoritme Support Vector Machine (SVM).pdf." .

[10] V. S and J. R, "Text Mining: open Source Tokenization Tools – An Analysis," *Adv. Comput. Intell. An Int. J.*, vol. 3, no. 1, pp. 37–47, 2016, doi: 10.5121/acii.2016.3104.

[11] E. B. Setiawan, D. H. Widyantoro, and K. Surendro, "Feature expansion using word embedding for tweet topic classification," *Proceeding 2016 10th Int. Conf. Telecommun. Syst. Serv. Appl. TSSA 2016 Spec. Issue Radar Technol.*, no. 2011, 2017, doi: 10.1109/TSSA.2016.7871085.

[12] A. Primajaya and B. N. Sari, "Random Forest Algorithm for Prediction of Precipitation," *Indones. J. Artif. Intell. Data Min.*, vol. 1, no. 1, p. 27, 2018, doi: 10.24014/ijaidm.v1i1.4903.

[13] K. Schouten, F. Frasincar, and R. Dekker, "An information gain-driven feature study for aspect-based sentiment analysis," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9612, pp. 48–59, 2016, doi: 10.1007/978-3-319-41754-7_5.

[14] R. Dzisevic and D. Sesok, "Text Classification using Different Feature Extraction Approaches," *2019 Open Conf. Electr. Electron. Inf. Sci. eStream 2019 - Proc.*, pp. 1–4, 2019, doi: 10.1109/eStream.2019.8732167.

[15] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, "Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking," *IEEE Access*, vol. 8, pp. 90847–90861, 2020, doi: 10.1109/ACCESS.2020.2994222.

[16] Karsito and S. Susanti, "Klasifikasi Kelayakan Peserta Pengajuan Kredit Rumah Dengan Algoritma Naïve Bayes Di Perumahan Azzura Residencia," *J. Teknol. Pelita Bangsa*, vol. 9, pp. 43–48, 2019.

Submitted: **27/07/2022**; Accepted: **31/08/2022**; Published: **25/10/2022**