

Multidimensional Sentiment Analysis of Tourism Object in DKI Jakarta, Banten, East Java, Central Java and West Java using Support Vector Machine Algorithm

1st Anggia Arfilinia

Faculty Of Industrial Engineering
Telkom University
Bandung, Indonesia

anggiarfilinia@student.telkomuniversit
y.ac.id

2nd Rachmadita Andreswari, S.Kom.,
M.Kom

Faculty Of Industrial Engineering
Telkom University
Bandung, Indonesia

andreswari@telkomuniversity.ac.id

3rd Faqih Hamami, S.Kom, M.T

Faculty Of Industrial Engineering
Telkom University
Bandung, Indonesia

faqihhamami@telkomuniversity.ac.id

4th José Manuel Ferreira Machado

Centro ALGORITMI - Department of
Informatics

University of Minho
Portugal

jmac@di.uminho.pt

Abstract— Information technology provides several conveniences, one of which is facilitating tourists in searching for information about tourist attractions. One of the services that can be used is Google My Business. The comments or reviews from tourists are numerous, thus requiring a significant amount of time to review them one by one. Therefore, a method is needed to address this issue, which is sentiment analysis. In this study, a multidimensional sentiment analysis was conducted on tourist attractions in the provinces of DKI Jakarta, Banten, East Java, Central Java, and West Java. Data was collected between January and March 2023 using the Data Miner tool. Two labeling techniques, namely Transformer and Textblob, were compared for labeling the data. The confusion matrix was employed as the evaluation tool, and the Support Vector Machine (SVM) technique was used to implement sentiment analysis. Labeling using transformers obtains an accuracy of 0.8169 or 81.6%, then the average precision, recall and f1-score are 68%, 54%, and 58%. While labeling using textblob, obtained an accuracy of 0.9257 or 92.5%, the average precision is 88%, recall and f1-score results are 87%. The accuracy, precision, recall, and f1-score results indicate that labeling using Textblob outperforms the labeling using Transformers.

Keywords— Multidimensional Sentiment Analysis, Sentiment Analysis, Support Vector Machine, Transformers, Textblob, Confusion Matrix

I. INTRODUCTION

Information technology provides several conveniences, one of which is making it easy for tourists to find information about a tourist attraction. One of the services that can be utilized is Google My Business, a service provided by Google and integrated with Google Maps. This service is used to assist entrepreneurs or managers of tourist attractions in managing their businesses. In addition to providing information about interesting places, Google My Business also offers features for tourists to leave comments or reviews.

There are so many comments or reviews from tourists, so it takes a long time to review them one by one (Herlawati et al., 2021). Therefore, a method is needed to collect comments and analyze large-scale data. One method that can be used to solve this problem is sentiment analysis. In this study, sentiment analysis was carried out from a multidimensional perspective on tourist objects. Multidimensional is used to observe data from various perspectives or dimensions, in this study discussing the dimensions of time and area.

II. EASE OF USE

A. Tourism

The tourism industry, according to UU No.10 of 2009, is a group of interconnected firms that create goods and/or services to cater to the demands of tourists while planning travel. The government must seriously manage tourism in Indonesia to improve and develop existing tourist objects to increase the number of tourist visits (Mardhiyah, 2018). The most crucial component of tourism activity is the presence of tourist attractions and objects. This is due to the fact that the potential and allure of these tourist artifacts are what draws travelers to visit tourist sites (Febrianti, 2020).

B. Social Media

Social media is a place where users can share information, create content, and more. Google My Business is one of the social media platforms offered by Google, which can be utilized by organizations or companies to promote their businesses, brands, or products quickly, easily, and free of charge. This platform is directly integrated with Google Maps, making it convenient for tourists to discover and access tourist destinations both nationally and internationally. Moreover, it facilitates interaction between tourist attraction managers and tourists (Manu & Fallo, 2019).

C. Sentiment Analysis

Sentiment analysis is the technique of mechanically deciphering and analyzing text data to get sentiment information from a sentence or text that takes the form of an opinion. In order to evaluate if a tendency is viewed as good or negative, this analysis seeks to identify the perspectives or opinions that are conveyed in a text about a topic or item. To identify sentiments in documents, sentiment analysis employs computational linguistics, text analysis, and natural language processing (Imron, 2019).

D. Support Vector Machine (SVM)

One of the techniques used in supervised machine learning for classification is support vector machine (SVM). This technique seeks to locate a hyperplane that can divide data into two distinct groups in a multidimensional space (Pawangfg, 2021). The hyperplane serves as the dividing line between classes, in contrast to the margin, which shows the separation between each class's nearest data points and the hyperplane. Support vectors in each class are the data points that are most closely spaced from the hyperplane. While the hyperplane serves as the boundary between classes, the margin is the separation between the hyperplane and the nearest data points in each class (Satrya et al., 2022).

E. Confusion Matrix

The Confusion Matrix technique is used to assess the categorization model and determine if an object is correct or incorrect. A matrix of predictions will be compared with the original class which contains actual information and predicted classification values. After the system has successfully performed a classification, a measure is needed to determine how valid or appropriate the classification has been made by the system. The following is a confusion matrix table (Pravina et al., 2019).

Figure II- 1 Confusion Matrix

Classification	Predicted Positives	Predicted Negatives
Actual Positive Cases	Number of TP Cases	Number of FN Cases
Actual Negative Cases	Number of FP Cases	Number of TN Cases

F. Recall, Precision and Accuracy

Precision, recall, and accuracy are the three metrics used to assess the categorization system's performance. The precision value is the difference between the information the system provides and its sensitivity value or system accuracy value. This value indicates negative class data or positive class. Recall value is a measurement of how successfully or precisely information may be located. This information relates to data that is either positive class or negative class (Azhari et al., 2021). The accuracy value is used to assess how accurately the system can anticipate something.

G. TF-IDF (Term Frequency-Inverse Document Frequency)

The TF-IDF determines how pertinent a word in a corpus or series is to a text (Riturajsaha, 2023). The following are some of the TF-IDF methods (Satrya et al., 2022):

1. Term frequency (TF) is the frequency of word occurrences in a text document. TF (tf,d) is defined as the number of

occurrences of term t in document d. The following is the TF equation:

$$TF(t, d) = f(t, d)$$

Information :

TF = Frequency of occurrence of a word in a text document

f = Number of words in a document

2. Inverse Document Frequency (IDF) is the frequency of occurrence of terms in all text documents. Terms that rarely appear in all text documents have an Inverse Document Frequency value greater than terms that often appear. The following is the IDF equation:

$$IDF = \log \frac{N}{DFt}$$

Information :

N : Number of documents

DFt : TF Value

3. Term Frequency – Inverse Document Frequency (TF-IDF) is a calculation for weighting. After TF and DF are obtained by equations (1) and (2) then look for TF – IDF. The following is the TF – IDF equation:

$$Wt = TF \times IDF$$

Information :

Wt : TF-IDF

TF : TF Value

IDF : IDF Value

III. RESEARCH METHODOLOGY

A. Conceptual framework

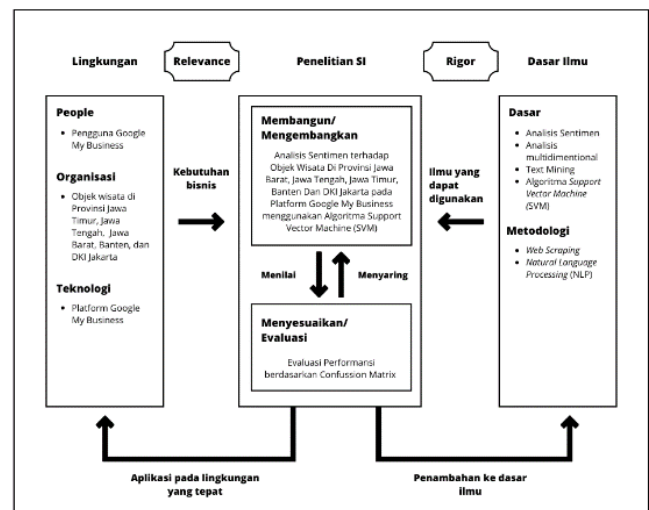


Figure III- 1 Conceptual framework

The framework implemented in this study was adopted from Design Science in IS Research (Hevner et al., 2004). This framework can be applied as a problem-solving strategy that aims to develop innovations based on theories, procedures, technical prowess, design analysis, implementation, management, and effective and efficient use of information systems (Andrian, 2020).

B. Problem Solving Systematics

The dataset used in this study will be examined using knowledge discovery in a database, where systematic problem-solving is employed to identify patterns in the dataset.

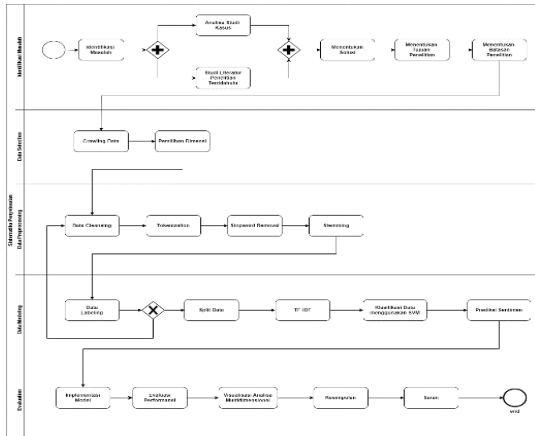


Figure III- 2 Problem Solving Systematics

C. Data Collection

At this stage, data collection is carried out through the Google Business platform. The collection of data used for this research is secondary data or data sourced from other parties. Secondary data includes reviews/comments from users on tourist objects. Crawling is carried out every month using keywords related to tourist objects in the Provinces of East Java, Central Java, West Java, Banten and DKI Jakarta.

D. Data Processing

Data processing is carried out according to the stages in Figure II-2. After the data is collected, data preprocessing will then be carried out. Data cleaning is done initially to make sure there is no noise data or information that is unrelated to the research topic before data analysis is done. Data that has passed the Data Cleaning process will be labeled to determine a positive or negative label. Data that has gone through the Data Preprocessing stage will be analyzed using the Support Vector Machine (SVM) algorithm. Word weighting is done at this stage to determine the TF-IDF (Term Frequency-Inverse Document Frequency). Then, the data will be split into training and test sets and further categorized using an algorithm to determine the accuracy of the predictions.

E. Evaluation method

The confusion matrix is the evaluation technique utilized in this study with the goal of determining if the classification model correctly or incorrectly classifies an object. Mapping the prediction results to the confusion matrix table to make it easier to calculate precision, recall and accuracy.

IV. ANALYSIS AND DESIGN

A. Data Collection

Based on the methodology in the previous chapter, the first step is data collection. The dataset used for the sentiment analysis process comes from a collection of tourist attraction reviews taken from Google My Business. The data crawling process is carried out using the Data Miner tool. The data taken is data in the form of visitor reviews about tourist attractions with a span of data collection starting from January 2023 to March 2023 and being crawled once a month. The data will be stored in excel format to facilitate the data

preprocessing process. In the process of crawling data, around 12,373 data were obtained.

B. Data Preprocessing

After the dataset has been collected using the Data Miner tools, then the raw data enters the data preprocessing stage. Data cleaning, tokenization, stopword removal, and stemming are performed during the data preprocessing stage. The following is the data preprocessing stage.

1. Data Cleaning

After going through the data cleaning process, the data will be more structured and later easier to analyze. The results of the data that have gone through the data cleaning process consist of deleting blank data, duplicate data, symbols and numbers, emoji and emoticons in the dataset.

2. Tokenization

The tokenization stage is the stage that forms a sentence which will be broken down word by word. This process is done with Python tools with the help of the Natural Language Toolkit (NLTK) library and importing several modules in it.

3. Stopword Removal

The stopword removal stage is when words like conjunctions that frequently appear in sentences but have no significant meaning are eliminated. This stage is carried out using the Natural Language Toolkit (NLTK) library and the Sastrawi library.

4. Stemming

Stemming is the process of modifying words using affixes to create fundamental words. This stage is carried out using the Natural Language Toolkit (NLTK) library. In addition, it also uses the Sastrawi library, which makes it possible to reduce inflected words in Indonesian to their standard form or according to standard dictionaries.

Tabel IV- 1 Output of Data Preprocessing

Before Preprocessing	After Preprocessing
Cape bener jalan dari pos 1 ke pos 6 🤔. Tiket masuk murah cuman 5k	cape bener jalan pos pos tiket murah cuman
Asli, keren banget sih ini curug.. 🤔❤️😊😊	asli keren banget sih curug
Membawa murid2 kesini sebagai tempat edukasi	bawa murid kesini edukasi
Sangat nyaman untuk berpiknik juga serta mengenal situs2nya	nyaman piknik kenal situs
cocok untuk mengisi liburan anak dan menambah pengetahuan sejarah gedung sate	cocok isi libur anak tambah tahu sejarah gedung sate

C. Data Modelling

1. Data Labelling

In this stage, the class labeling used is positive, negative, and neutral. The labeling process is carried out by comparing textblob and transformers, which are libraries for Natural Language Processing (NLP) tasks. The labeling results using transformers produced 8,340 data with 6,366 data labeled as positive, 1,157 data labeled as negative, and 818 data labeled as neutral. The following are some data and labels generated from the labeling process using transformers:

Tabel IV- 2 Labeling result using transformer and textblob

Labelling	Data	Label
Transformer	tempat sejuk ngopi angin sepoi mantap	positive
	ga rekom banget dr realita jalan akses jln rusak	negative
Textblob	a cool place to drink coffee with a steady breez	positive
	just unique	neutral
	banten mosque parking is difficult	negative

2. Split training data and testing data

After completing the labeling stage, the next step is to divide the data into training and testing sets. This stage involves comparing three ratios and selecting the one with the highest accuracy value. The three ratios being considered are 90:10, 80:20, and 70:30. Here are the details:

- For the 90:10 ratio, there are 7.760 training data and 863 testing data.
- For the 80:20 ratio, there are 6.908 training data and 1.725 testing data.
- For the 70:30 ratio, there are 6.036 training data and 2.587 testing data.

3. Calculation of TF-IDF

Calculation of Frequency-Inverse Document Frequency (TF-IDF) is the stage of weighting a word in the text mining process. The TF-IDF vectorizer replaces a collection of raw documents into a TF-IDF matrix. The TF-IDF calculation stages in this process use the scikit-learn (sklearn) library in python. The comparison of term weights provides an indication of which terms are more significant in representing a specific document. For example, the term "mahal" (expensive) has the highest value, indicating that this term is very important in a document.

4. Data classification using SVM

After calculating the Frequency-Inverse Document Frequency (TF-IDF), then carry out the data classification stages using the Support Vector Machine (SVM) algorithm. The following is the source code :

```
svm_model = SVC(kernel='rbf')

param_grid = [
    {'C': [0.1, 1, 10, 100],
     'gamma': [0.001, 0.01, 0.1, 1]}
]

model = GridSearchCV(svm_model, param_grid, cv=10, refit = True, verbose = 3)
# fitting the model for grid search
model.fit(x_train3, y_train3)
# print best parameter after tuning
print(model.best_params_)
```

Figure IV- 2 Source Code of Classification SVM

5. Sentiment Predictions

In this stage, the design of the Entity Relationship Diagram (ERD) is carried out as a guide in designing the visualization dashboard in Power BI. The ERD helps to understand and design the relationships between the entities involved, such as reviews, sentiment analysis results, and other information. By

designing an ERD, it can help visualize how data and information are related to each other.

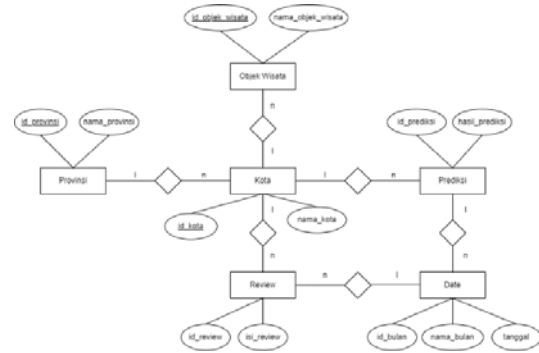


Figure IV- 1 ERD

6. Visualization of Prediction Results

Here is the mockup of the dashboard interface that will be designed using Power BI tools. The dashboard will display several components that provide various information, including:

1. Number of data
2. Pie Chart sentiment summary
3. Sentiment based graphs and also cities sorted by best sentiment provinces
4. Filter province, city, attractions and sentiments
5. Visitor review explanatory table

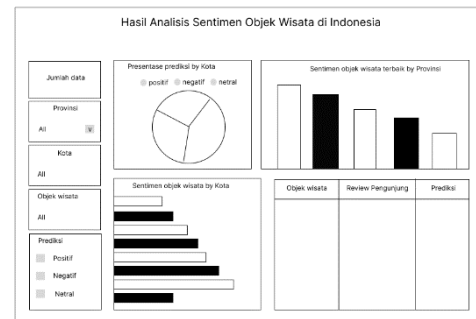


Figure IV- 3 Mockup of Visualization

V. RESULTS AND EVALUATION

D. Model Implementation

In the process of implementing the model, the author utilizes the Support Vector Machine (SVM) algorithm. The following is the configuration of the parameters used:

- Random_state = 1
- Test_size = 0.1, 0.2, 0.3
- Stratify = y

After conducting the Grid Search process on the SVC model, the output provides information about the best parameter combination. In the process of implementing the model Support Vector Machine (SVM) algorithm model with datasets that have been labeled using the transformers library. The results of the classification accuracy with the above parameters are as follows :

Tabel V- 1 Results accuracy per ratio

Rasio	Transformers	Textblob
90 : 10	0.8169	0.9192
80 : 20	0.8121	0.9257
70 : 30	0.8156	0.9165

Based on Table V-1, the accuracy results of labeling using a textblob show that the 80:20 ratio has the highest accuracy of 0.9257 or 92.5%. Meanwhile, in labeling using Transformers, the 90:10 ratio shows the highest accuracy of 0.8169 or 81.6%.

B. Performance Evaluation

During the performance evaluation stage, calculations of accuracy, precision, recall, and F1-score will be conducted based on the obtained confusion matrix. The following is the confusion matrix resulting from testing the Support Vector Machine (SVM) algorithm on tourism object review data from 5 provinces in Indonesia using transformer labeling. This testing was conducted with an 80:20 ratio as it obtained the highest accuracy among other ratios

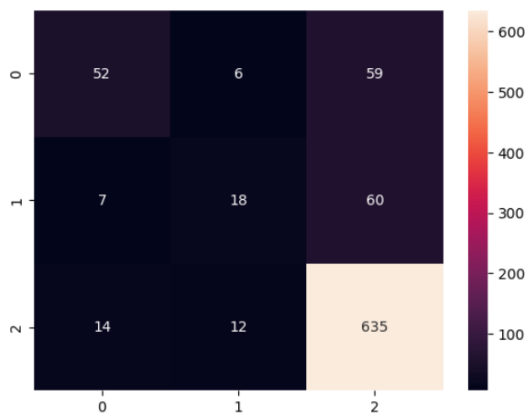


Figure V- 1 Confusion Matrix Transformer

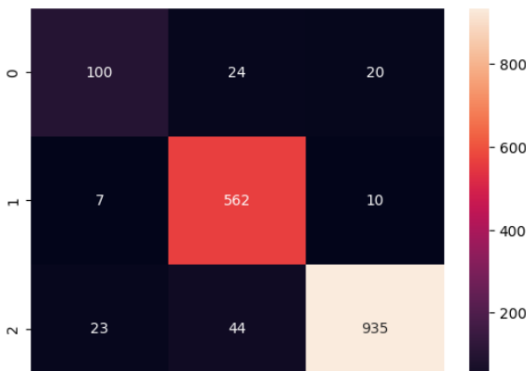


Figure V- 2 Confusion Matrix Textblob

Figure V-1 is the result of the visualization of the confusion matrix using transformer labeling. it is known that the average precision is 68%, recall is 54%, and f1-score is 58%. This indicates the model performs quite well in classification but needs to be improved especially on recall to identify more positive cases

The visualization results of the confusion matrix using labeling transformers show that the model has very good performance, with an average precision of 88%, recall of 87%, and F1 score of 87%, indicating a high ability to recognize

positive classes and a good balance between precision and recall.

C. Visualization Multidimensional Sentiment Analysis

The following is the result of visualization using Power BI tools. Based on the visualization of the results of the sentiment analysis that has been carried out, there are several decisions that can be proposed to the city government or regional government. Some of these recommendations are:

1. Identify negative reviews that contain visitor complaints and criticisms
2. Identification of the main causes of complaints or criticism
3. Carry out corrective actions to improve the quality of tourist objects, such as increasing accessibility, improving facilities, services, etc.
4. Doing marketing and promotions because positive reviews or reviews can attract the interest of other tourists.

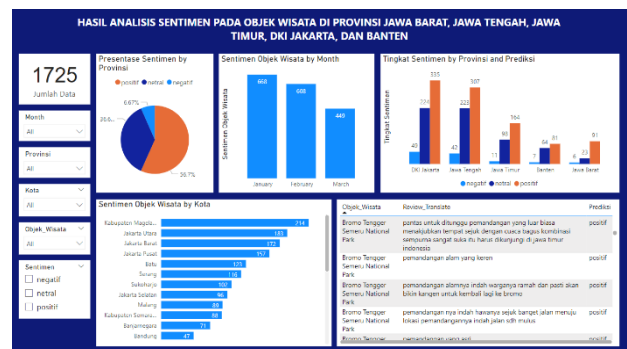


Figure V- 3 Visualization Multidimensional Sentiment Analysis

VI. CONCLUSIONS

A. Conclusions

1. Implementation of sentiment analysis is carried out using the Support Vector Machine (SVM) Algorithm. Data collection was carried out using the help of Data Miner tools on the Google My Business platform for a period from January to March 2023. Data labeling was done by comparing 2 labels, namely transformers and textblob. The labeling with the highest accuracy results is a textblob using a data splitting of 80:20 (80% training data and 20% testing data) and a confusion matrix as an evaluation method.

2. Comparison of positive, negative and neutral sentiments towards tourism objects generated using transformers for place (province) dimensions is 56.7% for positive sentiment, 6.67% for negative sentiment, and 36.6% for neutral sentiment. For the time dimension, in January the percentage was 57.93% for positive sentiment, 6.89% for negative sentiment, and 35.1% for neutral sentiment. In February, the percentages obtained were 56.41% for positive sentiment, 6.25% for negative sentiment, and 37.3% for neutral sentiment. For March, the percentages obtained were 55.23% for positive sentiment, 6.9% for negative sentiment, and 37.86% for neutral sentiment. Positive and neutral sentiment on tourism objects per province each month did not experience a significant decrease or increase. However, the highest percentage of negative sentiment was in March at 6.9%.

3. The accuracy generated using the Support Vector Machine (SVM) algorithm with transformers labeling is 0.8169 or 81.6%, then the average precision, recall and f1-score are 68%, 54%, and 58%. While labeling using textblob, obtained an accuracy of 0.9257 or 92.5%, the average precision is 88%, recall and f1-score are 87%.

REFERENCES

- [1] Andrian, R. (2020). Sistem Informasi Tunjangan Kinerja Untuk Menentukan Tambahan Penghasilan Pegawai Negeri Sipil Menggunakan Metode Design Science Research (Performance Allowance Information System for Determining Additional Civil Servant Income Using the Design Science Res. *Jurnal Teknologi Informasi Dan Multimedia*, 2(1), 1–11.
- [2] Azhari, M., Situmorang, Z., & Rosnelly, R. (2021). Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes. *Jurnal Media Informatika Budidarma*, 5(2), 640. <https://doi.org/10.30865/mib.v5i2.2937>
- [3] Febrianti, A. V. (2020). *Analisis Sentimen Data Ulasan Pengunjung Objek Wisata Lawang Sewu Kota Semarang Pada Situs Tripadvisor*. <http://lib.unnes.ac.id/41832/1/4112317002.pdf>
- [4] Herlawati, H., Handayanto, R. T., Atika, P. D., Khasanah, F. N., Yusuf, A. Y. P., & Septia, D. Y. (2021). Analisis Sentimen Pada Situs Google Review dengan Naïve Bayes dan Support Vector Machine. *Jurnal Komtika (Komputasi Dan Informatika)*, 5(2), 153–163. <https://doi.org/10.31603/komtika.v5i2.6280>
- [5] Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information System Research. *Management Information Systems*, 28(1), 75–105. <https://doi.org/https://doi.org/10.2307/25148625>
- [6] Imron, A. (2019). Analisis Sentimen Terhadap Tempat Wisata di Kabupaten Rembang Menggunakan Metode Naive Bayes Classifier. *Teknik Informatika*, 10–13. <https://dspace.uir.ac.id/handle/123456789/14268>
- [7] Manu, G. A., & Fallo, D. (2019). Implementasi Google My Business (Gmb) Dalam Promosi Pariwisata Di Kota Kupang Dan Sekitarnya. *Jurnal Pendidikan Teknologi Informasi (JUKANTI)*, 2(2), 8–15. <https://doi.org/10.37792/jukanti.v2i2.69>
- [8] Mardiyah, A. (2018). Strategi Komunikasi Pemasaran Kementerian Pariwisata Indonesia dalam Pesona Indonesia melalui Youtube. *Jurnal Manajemen Komunikasi*, 02(2), 144–161.
- [9] Pawangfg. (2021). *Separating Hyperplanes in SVM*. Geeksforgeeks.Org. <https://www.geeksforgeeks.org/separating-hyperplanes-in-svm/>
- [10] Pravina, A. M., Cholissodin, I., & Adikara, P. P. (2019). Analisis Sentimen Tentang Opini Maskapai Penerbangan pada Dokumen Twitter Menggunakan Algoritme Support Vector Machine (SVM). *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, 3(3), 2789–2797. <http://j-ptiik.ub.ac.id>
- [11] Riturajsaha. (2023). *Understanding TF-IDF (Term Frequency-Inverse Document Frequency)*. <https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/>
- [12] Satrya, R. N., Pratiwi, O. N., & Abawajy, J. (2022). Cryptocurrency Sentiment Analysis on the Twitter Platform Using Support Vector Machine (SVM) Algorithm. *IEEE Access*, 10–29. <https://ieeexplore.ieee.org/abstract/document/10037413>