International Journal of Technology and Education Research

https://e-journal.citakonsultindo.or.id/index.php/IJETER

E-ISSN: 2986-0733

Utilization of Natural Language Processing (NLP) For The Analysis and Classification of Career Concern Topics Based on Tiktok Comments Using Support Vector Machine (SVM) With a Web-Based System

Natasya Maharani[1], Yohanni Syahra[2]

[1]Program Studi Sistem Informasi Fakultas Ilmu Komputer Universitas Muhammadiyah Sumatera Utara Medan

[2]Program Studi Teknologi Informasi Fakultas Ilmu Komputer Universitas Muhammadiyah Sumatera Utara Medan

Corresponding Author: natasyamaharani47@gmail.com

Abstract

The rapid transformation of the labor market has increased the expression of career concerns through social media, including TikTok. This study aims to classify career concern topics in Indonesian-language TikTok comments using Natural Language Processing (NLP), Term Frequency-Inverse Document Frequency (TF-IDF), and Support Vector Machine (SVM), and to implement the resulting model in a web-based system. This quantitative study employed text mining and machine learning methods. The population comprised Indonesian-language TikTok comments concerning employment and career issues, while the sample consisted of 305 manually labeled comments categorized as Workload, Work Environment, and Work Compensation. Data were processed through cleansing, case folding, normalization, tokenization, stopword removal, and stemming, followed by TF-IDF feature extraction. An SVM model with a linear kernel and C = 1.0 was trained using an 80:20 data split and evaluated using a confusion matrix, accuracy, precision, recall, and F1-score. The results showed an accuracy of 96.72% and macro precision, recall, and F1-score of 96.67%. The model correctly classified 59 of 61 testing comments. In conclusion, the NLP-TF-IDF-SVM approach effectively classified career concern topics and was successfully implemented in a functional web-based classification system.

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

International Journal of Technology and Education Research

https://e-journal.citakonsultindo.or.id/index.php/IJETER

E-ISSN: 2986-0733

Keywords: Career Concerns, Natural Language Processing, Social Media, Support Vector Machine, TikTok

INTRODUCTION

The development of digital technology has influenced changes in the world of work. Advances in technology, automation, and business process digitalization have resulted in changes in the skills required of workers. Workers are not only expected to complete tasks according to their respective fields, but also need to adapt to technological developments, new work patterns, and increasingly dynamic organizational demands. These changes also affect job stability and career certainty. The World Economic Forum report indicates that technological developments and automation contribute to changes in the types of jobs and competencies required in the future (Bahram, 2023). Increasingly competitive labor market conditions may create pressure for individuals, particularly job seekers, young workers, and fresh graduates entering the workforce. Mismatches between expectations and realities in the workplace, such as difficulties in obtaining employment, high workloads, and inadequate compensation, may generate concerns about future careers. Under certain conditions, such pressure may also be associated with work exhaustion, stress, and feelings of insecurity regarding employment continuity (Luwei & Huimin, 2024).

Career concerns represent individuals' worries about employment conditions, employment opportunities, workload, and well-being related to their careers. Such concerns do not only emerge through direct conversations but are also widely expressed through social media. Social media has become a space where users can openly express opinions, experiences, and complaints (Kurniawan et al., 2021). The practice of sharing personal stories or "curhat" on social media also indicates that users utilize digital platforms to obtain social support, validation, and responses from other users (Wicaksana & Rudy, 2021). One of the widely used platforms is TikTok. Indonesia is recorded as one of the countries with the largest number of TikTok users, reaching 99.1 million users (Indriyani et al., 2023). In addition to being used for entertainment, TikTok has also become a space for discussions on various social issues, including employment-related problems. Through the comment section, users can express opinions, personal experiences, and complaints related to work and careers (Mardianto, 2023).

The research problem arises because TikTok comments related to career concerns are characterized as unstructured textual data. User comments often contain informal language, abbreviations, mixed languages, symbols, emoticons, and informal writing styles. These characteristics make manual analysis less effective, particularly when a large amount of data

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

needs to be analyzed (Abdurrazik & Wirawana, 2025). Nevertheless, such comments may contain information regarding various issues related to career concerns, including inadequate compensation, excessive workloads, unsupportive work environments, and concerns about employment conditions. Therefore, a processing approach is required to transform unstructured textual data into information that can be analyzed systematically.

Another research problem concerns the processing and classification of comment data. Natural Language Processing (NLP) is required to process textual data through stages such as *case folding*, tokenization, *stopword removal*, normalization, and stemming so that the data become cleaner and ready for analysis. These pre-processing stages are used in text mining to prepare textual data before further processing (Rifaldi et al., 2023). Following this process, textual data need to be represented numerically using Term Frequency–Inverse Document Frequency (TF-IDF), a weighting method that assigns weights to important words in textual data (Wibowo et al., 2024). Subsequently, a classification algorithm is required to group comments according to career concern topics. Support Vector Machine (SVM) works by identifying the optimal *hyperplane* to distinguish data into specific classes and has the ability to handle high-dimensional data, such as textual data represented using TF-IDF (Mutawalli et al., 2019). SVM has also been applied in text classification research and has demonstrated good performance on social media data (Fide et al., 2021). Meanwhile, Naïve Bayes assumes that features are independent of one another, whereas K-Nearest Neighbor (KNN) is sensitive to outliers and depends on the selection of the *k* parameter. Therefore, SVM is used in this study because it is considered appropriate for classifying TikTok comments into several career concern topics (Agustian & Bisri, 2019; Indrayuni et al., 2021).

Another problem can be identified from the research gap in previous studies. Text analysis on social media has generally focused on sentiment analysis, which categorizes opinions into positive, negative, and neutral classes. For example, Hilmi & Irwiensyah (2024) applied sentiment analysis to social media data to classify user opinions. Although this approach can indicate users' emotional tendencies, it does not explain the specific topics or forms of problems being discussed. In addition, previous studies have frequently used Twitter, application reviews, or surveys as research data sources. Herlinda et al. (2024), for example, used Twitter as a research data source, while other studies have demonstrated that SVM with TF-IDF has been applied to TikTok application reviews on Google Play with good sentiment classification performance. Topic classification has also been applied to Twitter tweets, but the utilization of TikTok comments to classify career concern topics remains relatively limited. Therefore, an

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

approach is needed that does not merely examine sentiment polarity but is able to identify specific topics representing forms of career concerns based on the content of comments.

This study aims to develop a career concern topic classification model based on TikTok comments using the Natural Language Processing (NLP) approach and Support Vector Machine (SVM) algorithm, determine the performance of SVM in classifying career concern topics based on the evaluation metrics used, and design and implement the model into a web-based system. The urgency of this study lies in the need to process unstructured TikTok comments into information that can provide an overview of career concern patterns emerging in the digital community. The novelty of this study lies in its focus on classifying career concern topics based on public TikTok comments into three categories, namely Workload, Work Environment, and Work Compensation, using NLP, TF-IDF, and SVM, followed by the implementation of the classification model into a web-based system so that the classification results can be presented in a more structured and understandable manner.

RESEARCH METHODOLOGY

This study employs a quantitative approach using *text mining* and *machine learning* algorithms to process and classify TikTok user comments related to career concern topics. The quantitative approach was selected because the study focuses on systematically processing textual data and objectively measuring classification model performance based on evaluation metrics. *Text mining* is used to extract information from TikTok comments related to career concerns, while the *machine learning* algorithm applied in this study is *Support Vector Machine* (SVM), which classifies comments into three categories, namely Workload, Work Environment, and Work Compensation. The selection of SVM is based on its ability to handle high-dimensional data and its application in text and social media classification (Fide et al., 2021; Mutawalli et al., 2019). The textual data are first processed using the *Natural Language Processing* (NLP) approach and then represented numerically using *Term Frequency–Inverse Document Frequency* (TF-IDF) before being used in the classification process.

The research instruments consist of a dataset of Indonesian-language TikTok comments relevant to career and workplace issues and data processing tools used in the classification process. The comments were collected from TikTok videos discussing work experiences, workload, work environment, salaries, compensation, and career concerns, and were subsequently selected and manually labeled according to the categories of Workload, Work Environment, and Work Compensation. Data analysis was conducted through *pre-processing* stages consisting of *cleansing*, *case folding*, normalization, *tokenizing*, *stopword removal*, and

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

stemming to clean and standardize the textual data. The pre-processed data were then transformed into numerical representations using TF-IDF with the assistance of Python and the *scikit-learn* library. The dataset was divided into training and testing data at an 80:20 ratio. The training data were used to develop an SVM model with a *linear kernel* and a regularization parameter of C = 1.0, while the testing data were used to assess the model's classification capability. Model performance was analyzed using a *confusion matrix* and the evaluation metrics of *accuracy*, *precision*, *recall*, and *F1-score* to determine the model's ability to classify comments into the specified categories (Putra et al., 2022).

The population of this study consists of Indonesian-language TikTok user comments related to career and workplace issues, particularly comments found in videos discussing employment, work experiences, workload, work environment, salaries, compensation, and career concerns. From the overall comments relevant to the research context, the data were selected to obtain comments that corresponded to the focus of career concern topic classification. The research sample consisted of 305 Indonesian-language TikTok comments that had undergone selection and manual labeling. Each comment was assigned one primary label based on its dominant topic, with three categories consisting of Workload, Work Environment, and Work Compensation. The resulting data were then used as the dataset for training and testing the classification model using SVM, with 80% of the data allocated as training data and 20% as testing data.

The research procedure was conducted systematically, beginning with problem identification and continuing through the implementation of the model into a web-based system. The first stage involved identifying the phenomenon of career concerns expressed by TikTok users through the comment section, followed by a literature review concerning career concerns, TikTok social media, text classification, NLP, TF-IDF, SVM, model evaluation, and web-based systems. The next stage involved collecting comments from several TikTok videos based on keywords related to work experiences, workload, work environment, salaries, compensation, and career concerns, after which the data were selected and manually labeled into the categories of Workload, Work Environment, and Work Compensation. The labeled data then underwent *pre-processing* consisting of *cleansing*, *case folding*, normalization, *tokenizing*, *stopword removal*, and *stemming*, followed by feature extraction using TF-IDF. The dataset was subsequently divided into training and testing data at an 80:20 ratio, after which the classification model was developed using SVM with a *linear kernel* and evaluated using a *confusion matrix*, *accuracy*, *precision*, *recall*, and *F1-score*. After model evaluation, the classification results were used as the basis for designing and implementing a web-based system

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

capable of managing the dataset, performing *pre-processing*, conducting classification, and displaying the classification results of new comments according to career concern topics.

RESULTS AND DISCUSSION

Data Attributes

Table 1. Data Attributes

Attribute	Explanation
No	An attribute used as a data sequence number. This attribute makes it easier to identify each comment in the dataset.
Comment	Attributes containing the text of TikTok user comments. These attributes are the primary data that will be analyzed and processed in the research.
Label	An attribute that contains the categories of comments based on the topic of career anxiety. This attribute is used as a classification target in the model building process.

Based on these attributes, the collected comment data becomes more structured and ready for use in the next stage. The comment attribute will be processed in the data preprocessing stage, while the label attribute will serve as a reference in the training and testing of the classification model.

Data Labeling Results

Table 2. Distribution of Dataset Labels

Label	Amount of Data
Workload	96
Work environment	102
Work Compensation	107
Total	305

Based on Table 2, 107 comments were labeled "Work Compensation," 102 comments were labeled "Work Environment," and 96 comments were labeled "Workload." The "Work Compensation" category had the largest number of entries, while the "Workload" category had the smallest number.

The difference in the amount of data in each category indicates that comments related to work compensation were more prevalent than in other categories. This indicates that issues

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

regarding salary, wages, benefits, and the mismatch between workload and rewards are topics frequently discussed by TikTok users in the context of career concerns. Meanwhile, the number of comments in the Workload and Work Environment categories also did not differ significantly, indicating a fairly balanced distribution of data across the three categories. This situation supports the model training process because each category has sufficient data for the Support Vector Machine algorithm to learn from.

Data Pre-processing Results

1. Cleaning

In the cleaning stage, the system processes comments by removing unnecessary text. Comments that originally contained punctuation, symbols, emojis, links, specific numbers, or special characters are cleaned to leave only the words needed for analysis.

For example, comments containing exclamation marks, emoticons, or excessive punctuation will be converted to cleaner text. This helps the system reduce distractions from characters not directly related to the comment's topic.

2. Folding Case

After comments are cleared, the system converts all letters in the comments to lowercase. This process ensures consistent spelling.

At this stage, words previously written in capital letters or mixed letters will be rendered the same. For example, the words "Work," "WORK," and "work" will be processed as "work." This way, the system doesn't read the same word as a different form simply because of the difference in capitalization and lowercase letters.

3. Normalization

The next stage is normalization. In this stage, the system replaces non-standard words, abbreviations, or inappropriate spellings with more standardized forms. This process is carried out because comments on TikTok generally use informal language and contain many abbreviations.

For example, the words "gk," "ga," or "gak" were standardized to "tidak." The word "yg" was changed to "yang," while overwritten words like "lemburr" were adjusted to "lembur." This process allows words with the same meaning but written in different forms to be standardized.

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

Improvements to the normalization dictionary also helped the system standardize the word variations that frequently appear in TikTok comments. This resulted in more consistent preprocessing results before being used in the TF-IDF weighting stage.

4. Stopword Removal

After the normalization process is performed, the system removes words that are less influential in determining the label. Common words that frequently appear in sentences are removed to focus the preprocessing results on words with significant meaning.

At this stage, words like "yang," "di," "ke," "dan," "itu," or other common words can be removed if they don't significantly impact the topic of the comment. After this stage, comments become more concise and contain words more relevant to the classification category, such as "overtime," "salary," "tired," "boss," "toxic," and other important words.

5. Stemming

Table 3. Data Preprocessing Results

No	Initial Comments	Data Preprocessing Results
1	The boss is nice but the work partner is toxic and makes us uncomfortable 😭	Good boss, toxic coworkers make you uncomfortable
2	full time work + physical exhaustion without pay☹	full time work, physically tiring, no salary
3	But my brain feels like it's going to burst under the pressure of constantly changing regulations, it's tiring	brain feels like it's breaking, keep pressing, adjust, change, tired
4	Small salary but lots of work demands	Small salary requires a lot of work
5	absolutely true, the workload is very heavy and never ending 😭	It's true that the heavy workload doesn't end

Table 3 shows that the initial comments, which still contained non-standard words, punctuation, and inconsistent word forms, were processed into simpler text. Although the preprocessing results did not always form complete sentences, key words related to the comment topic were retained. These preprocessing results were then used as input data in the word weighting stage using TF-IDF.

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

International Journal of Technology and Education Research

https://e-journal.citakonsultindo.or.id/index.php/IJETER

E-ISSN: 2986-0733

TF-IDF Weighting Results

After preprocessing, the cleaned comment data is processed using TF-IDF to convert the text into numeric form for processing by the Support Vector Machine (SVM) algorithm. This process is carried out automatically using TfidfVectorizer in Python by calculating weights based on Term Frequency (TF) and Inverse Document Frequency (IDF). The result is a feature matrix that is used as input in training the SVM model. As an illustration, the comment "good boss, toxic coworkers make me uncomfortable" is used, with each word appearing once, resulting in a TF value of 1.

Table 4. Term Frequency (TF) Calculation

Say	TF
superior	1
Good	1
colleague	1
Work	1
toxic	1
make	1
No	1
feel at home	1

Next, the Inverse Document Frequency (IDF) value is calculated. The IDF value is calculated based on the total number of documents (N = 305) and the number of documents containing a particular word (Document Frequency/DF). The IDF formula used is:

$$IDF = \log\left(\frac{N}{DF}\right)$$

Information:

N = total number of documents (305 comments)

DF = number of documents containing a particular word.

For example, the word betah appears in 4 documents, so the calculation is as follows.

$$IDF = \log\left(\frac{305}{4}\right)$$
$$= \log(76{,}25)$$
$$= 1{,}8822$$

Next, the TF-IDF value is calculated using the formula:

$$TF\text{-}IDF = TF \times IDF$$

Because the TF value of the word betah is 1, then:

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

International Journal of Technology and Education Research

https://e-journal.citakonsultindo.or.id/index.php/IJETER

E-ISSN: 2986-0733

$$TF\text{-}IDF = 1 \times 1{,}8822 = 1{,}8822$$

The same calculation was performed on all the words in the comment to obtain the results as in Table 5.

Table 5. Example of TF-IDF Calculation

Say	TF	DF	IDF	TF-IDF
superior	1	20	1,1833	1,1833
Good	1	11	1.4425	1.4425
colleague	1	7	1.6399	1.6399
Work	1	187	0.2123	0.2123
toxic	1	64	0.6782	0.6782
make	1	15	1,3082	1,3082
No	1	125	0.3874	0.3874
feel at home	1	4	1.8822	1.8822

Based on Table 5, it can be seen that each word has a different TF-IDF weight. The word "betah" has the highest weight because it appears in only a few documents, making it considered more specific in representing the content of the comment. Conversely, "verb" has a lower weight because it appears in many documents, making the level of differentiation between categories relatively smaller. The resulting TF-IDF weights are then used as a numerical representation of each comment before further processing by the Support Vector Machine (SVM) algorithm.

Results of Data Sharing and SVM Model Training

Table 6. Distribution of Training and Testing Data

Data Types	Amount of Data	Percentage
Training Data	244	80%
Testing Data	61	20%
Total	305	100%

Based on Table 6, the 80:20 data split yields 244 training datasets and 61 testing datasets. This composition provides sufficient data for the model to learn patterns in each category, while also providing sufficient testing data to evaluate the model's performance on previously unstudied data.

After the data was divided, the model training process was carried out using the training data that had gone through the preprocessing and TF-IDF weighting stages. At this stage, the

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

comment data, which was previously in text form, was converted into numeric values, so it could be used as input for the Support Vector Machine algorithm. The SVM model was trained to recognize comment patterns based on three label categories: Workload, Work Environment, and Work Compensation. The parameters used in the model training process can be seen in Table 7.

Table 7. SVM Model Training Parameters

Parameter	Mark
Algorithm	Support Vector Machine
Kernel	Linear
Grade C	1.0
Random State	42
Training Data	244
Testing Data	61

Based on Table 7, the classification model was built using the Support Vector Machine algorithm with a linear kernel. A C value of 1.0 was used to maintain a balance between the model's ability to separate data and reduce classification errors, while a random state of 42 was used to ensure the data separation process produces a consistent structure when training is repeated. The use of these parameters resulted in a model ready for use in the evaluation and classification stages of new comments.

The result of the training process is a classification model that can be used to predict new comment categories. The successfully trained model is then stored by the system for use in the comment evaluation and classification process.

Model Evaluation Results

Table 8. SVM Model Evaluation Results

Evaluation Metrics	Mark
Accuracy	96.72%
Precision Macro	96.67%
Recall Macro	96.67%
F1-score Macro	96.67%

Based on Table 8, the model achieved an accuracy value of 96.72%. This value indicates that the model was able to correctly classify most of the test data. The macro precision value of

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

International Journal of Technology and Education Research

https://e-journal.citakonsultindo.or.id/index.php/IJETER

E-ISSN: 2986-0733

96.67% indicates that the model's prediction results for each category have a good level of accuracy. The macro recall value of 96.67% indicates that the model was able to recognize most of the data according to its actual label. Meanwhile, the macro F1-score value of 96.67% indicates that the model has a good balance between precision and recall.

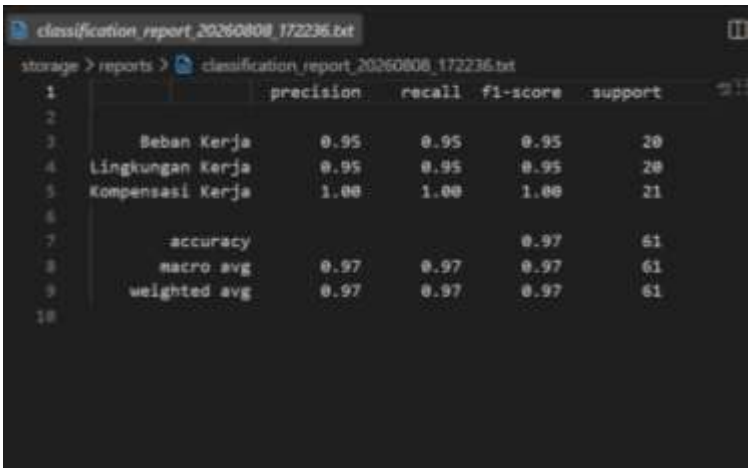


Figure 1. Classification Report Results Using Python

These values indicate that the model has a good ability to distinguish the three comment categories, so that the classification process can be carried out consistently on data that has never been studied before.

In addition to the overall evaluation metrics, the evaluation results are also displayed in the form of a confusion matrix. The confusion matrix is used to compare the actual labels with the model's predicted labels. The results of the confusion matrix can be seen in Table 4.9.

Table 9. Confusion Matrix SVM Model

Actual \ Prediction	Workload	Work environment	Work Compensation
Workload	19	1	0
Work environment	1	19	0
Work Compensation	0	0	21

Based on Table 9, the number of correct predictions is on the diagonal: 19 for Workload, 19 for Work Environment, and 21 for Work Compensation. Thus, 59 items were correctly classified out of a total of 61 test items. Meanwhile, two items were mispredicted.

In the Workload label, there are 20 actual data items. Of these, 19 were correctly predicted as Workload and 1 was predicted as Work Environment. In the Work Environment label, there are 20 actual data items. Of these, 19 were correctly predicted as Work Environment

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

International Journal of Technology and Education Research

https://e-journal.citakonsultindo.or.id/index.php/IJETER

E-ISSN: 2986-0733

and 1 was predicted as Work Load. Meanwhile, in the Work Compensation label, there are 21 actual data items, all of which were correctly predicted as Work Compensation.

Prediction errors only occurred between the Workload and Work Environment categories. This indicates that the two categories share some similar word characteristics, making it difficult for the model to distinguish between them in some comments. In contrast, the Work Compensation category was recognized very well, as there were no misclassifications within that category.



Figure 2. Confusion Matrix Results Using Python

The figure shows that the confusion matrix results generated by Python match the confusion matrix values presented in Table 9.

The evaluation metrics for each label can be seen in Table 10.

Table 10. Evaluation Metrics per Label

Label	Precision	Recall	F1- Score	Support
Workload	95.00%	95.00%	95.00%	20
Work environment	95.00%	95.00%	95.00%	20
Work Compensation	100.00%	100.00%	100.00%	21

Based on Table 10, the Work Compensation label achieved a precision, recall, and F1-score of 100.00%. This indicates that the model correctly recognized all test data in the Work

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

Compensation category. The Workload and Work Environment labels each achieved a precision, recall, and F1-score of 95.00%, indicating that the model successfully recognized both categories, although there were still small prediction errors.

The difference in evaluation scores between categories shows that comments in the Work Compensation category tend to have more specific word characteristics, such as 'salary', 'wage', or 'bonus', so they are easier for the model to distinguish than the other two categories.

Overall, the evaluation results showed that the Support Vector Machine model performed well in classifying TikTok comments based on career concerns. An accuracy value of 96.72% and a macro F1-score of 96.67% indicate that the model was able to produce accurate and balanced predictions across all three categories. With this performance, the model demonstrated good capability in supporting the comment classification process in the developed system.

Web System Implementation Results

1. Login Page

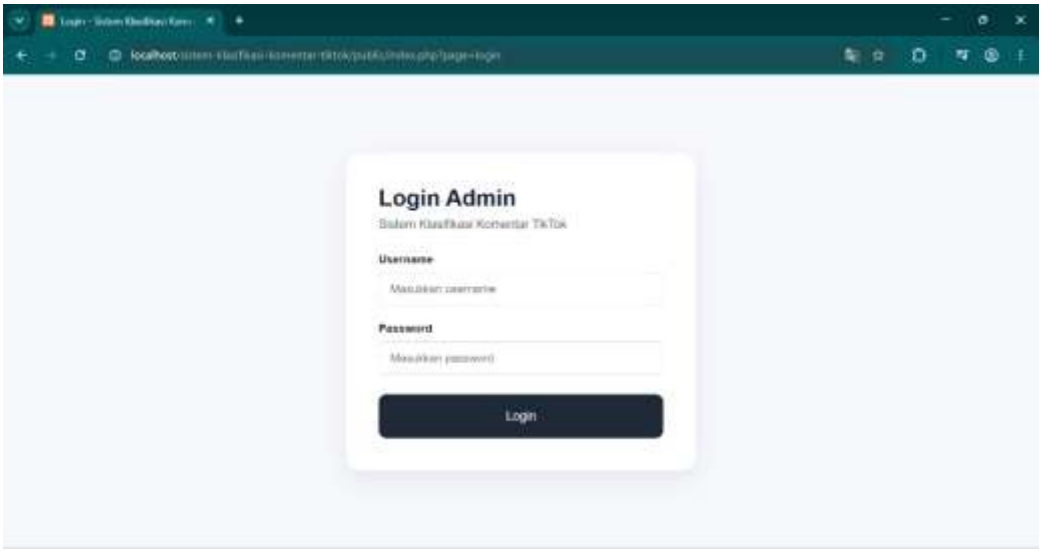


Figure 3. Login Page

The login page is the first page displayed when the system is run. This page serves as an authentication process to ensure that only admins with existing accounts can access the TikTok comment classification system. On this page, there are two input fields: username and password, which are used for user authentication. Once both fields are filled in, the admin can press the Login button to enter the system. If the username and password entered match the stored data, the system will direct the admin to the dashboard page. Conversely, if the entered

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

International Journal of Technology and Education Research

https://e-journal.citakonsultindo.or.id/index.php/IJETER

E-ISSN: 2986-0733

data is invalid, the system will display an error message so the admin can log in again with the correct information.

2. Dashboard Page

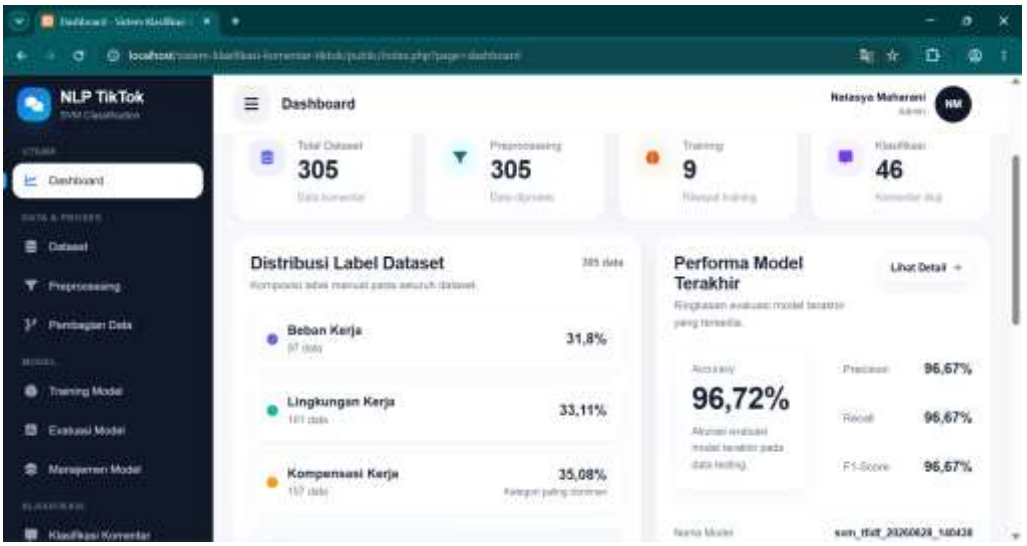


Figure 4. Dashboard page

The dashboard page is the main page displayed after the admin successfully logs into the system. This page provides general information about the state of TikTok's comment classification system in an easy-to-understand summary.

On the dashboard page, the system displays several information cards containing the number of datasets, pre-processing results, the models used, and a summary of system activity. Additionally, a label distribution visualization is displayed to show the distribution of data across each classification category.

The dashboard also displays the accuracy, precision, recall, and F1-score values of the active model, allowing admins to assess model performance without having to open the evaluation page. Based on the implementation results, the dashboard provides a concise presentation of important information about dataset conditions and model performance, making it easier for admins to monitor the classification process and manage the system more efficiently.

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

International Journal of Technology and Education Research

https://e-journal.citakonsultindo.or.id/index.php/IJETER

E-ISSN: 2986-0733

3. Dataset Page

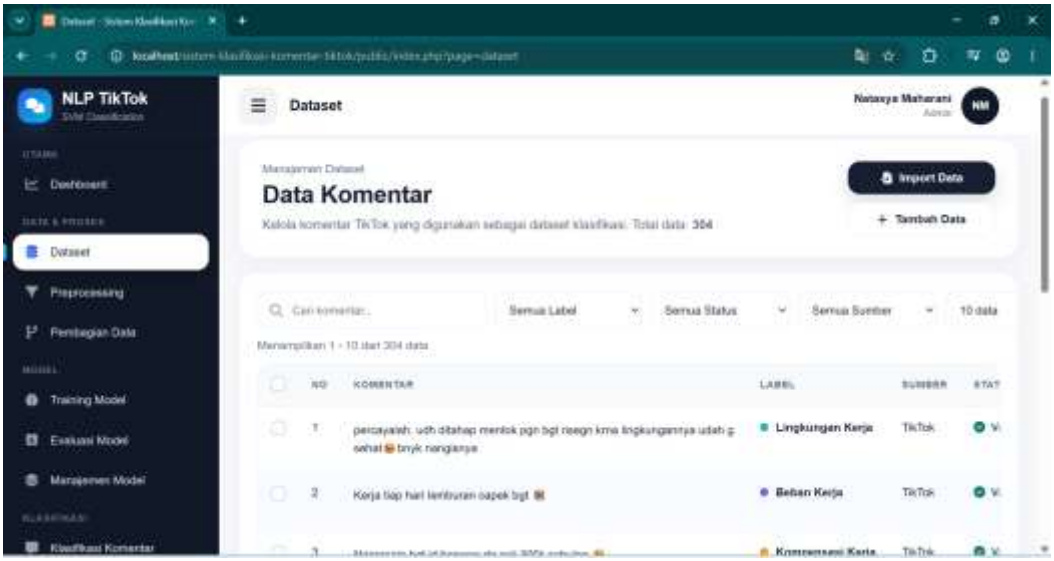


Figure 5. Dataset Page

The dataset page is used to manage TikTok comment data that will be used as a research dataset. All data is displayed in tabular form, making it easier for admins to view and manage the available information.

On this page, the system provides a search feature to make it easier for admins to find specific comments. Furthermore, there's a filtering feature based on label, status, and data source, making the search process faster and more targeted. The system also features settings for the amount of data displayed per page to keep the table display neat and easy to read.

The dataset page also has a data import feature for uploading datasets in CSV or XLSX format, as well as an add data feature for manually entering comments. Each record has an edit feature for updating comment information and a delete feature for removing unused data.

Through the dataset page, the admin can carry out the entire comment data management process before the data is processed to the next stage, namely preprocessing, data sharing, model training, and model evaluation.

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

International Journal of Technology and Education Research

https://e-journal.citakonsultindo.or.id/index.php/IJETER

E-ISSN: 2986-0733

4. Preprocessing Page

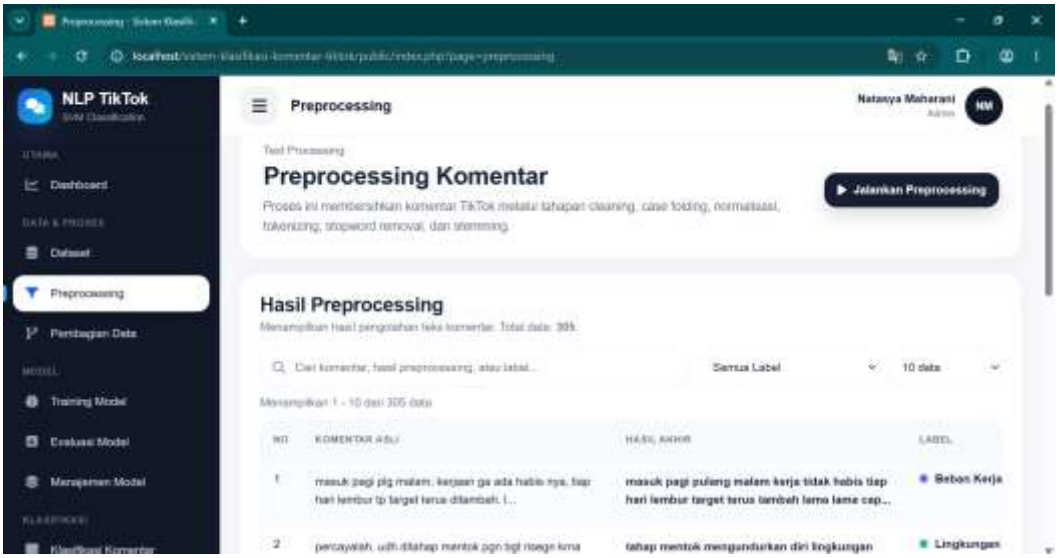


Figure 6. Preprocessing Page

The pre-processing page is used to process comment text before it is used in the TF-IDF weighting and model training stages. On this page, the system processes comments through several stages: cleaning, case folding, normalization, tokenizing, stopword removal, and stemming.

The pre-processing results are displayed in a table containing the original comment, the final pre-processing result, and the comment label. This display makes it easy for admins to see how comments have changed after each stage of text processing.

This page also features a search feature, label filters, and settings for the amount of data displayed in the table. This page allows admins to ensure that the data has been properly processed before being used in the data sharing, model training, and model evaluation stages.

5. Data Sharing Page

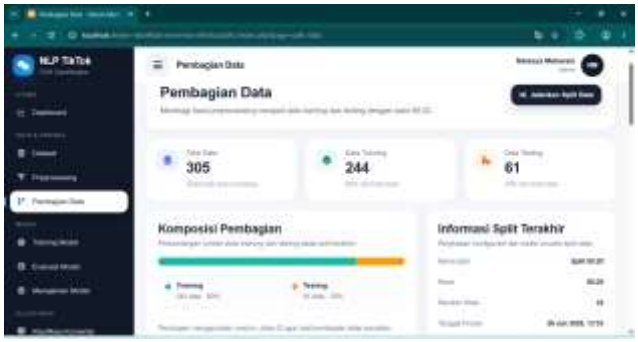


Figure 7. Data Sharing Page

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

International Journal of Technology and Education Research

https://e-journal.citakonsultindo.or.id/index.php/IJETER

E-ISSN: 2986-0733

The data splitting page is used to divide the pre-processed data into training and testing data. On this page, the admin can run the data splitting process using the Run Split Data button.

This page displays data split information, such as the split ratio, the composition of the training and testing data, and the final split information, including the random state and the splitting time. The system also displays the label distribution across the training and testing data to ensure each category remains represented after the splitting process.

At the bottom of the page, a detailed table displays the data split results, including data types, comments, preprocessing results, and labels. This page allows administrators to view the data split results before using them in model training and evaluation.

6. Model Training Page

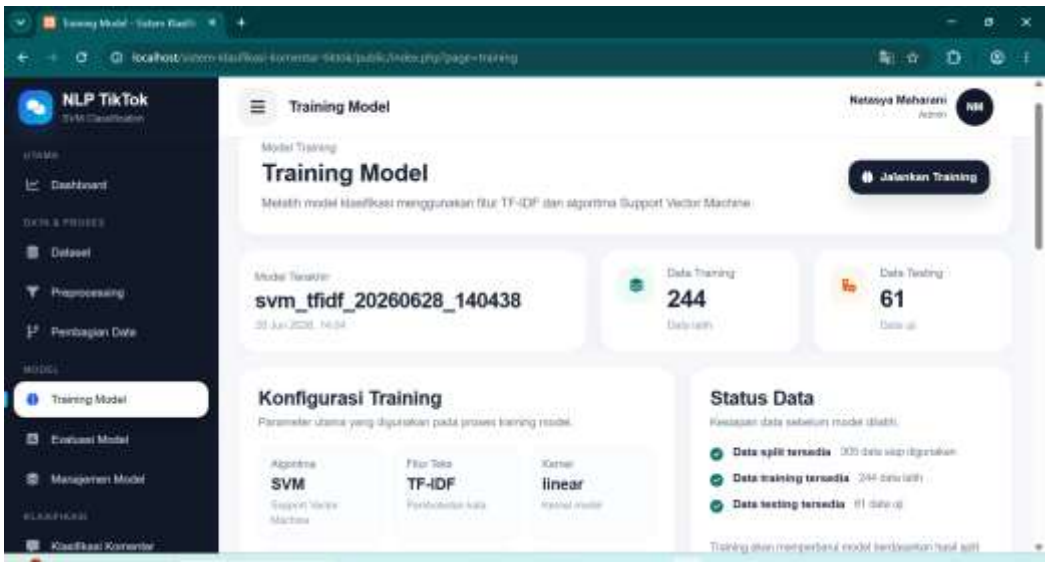


Figure 8. Model Training Page

The model training page is used to run the classification model training process using the Support Vector Machine algorithm. On this page, the system displays information about the model used, the amount of training data, the amount of testing data, and the applied training configuration.

Additionally, this page displays the data readiness status before the training process begins, allowing administrators to ensure that training and testing data are available. Once the training process is complete, the system saves the resulting model for use in the evaluation phase and comment classification.

Through this page, the admin can run the model training process directly through the system without having to run the program code manually.

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

International Journal of Technology and Education Research

https://e-journal.citakonsultindo.or.id/index.php/IJETER

E-ISSN: 2986-0733

7. Model Evaluation Page

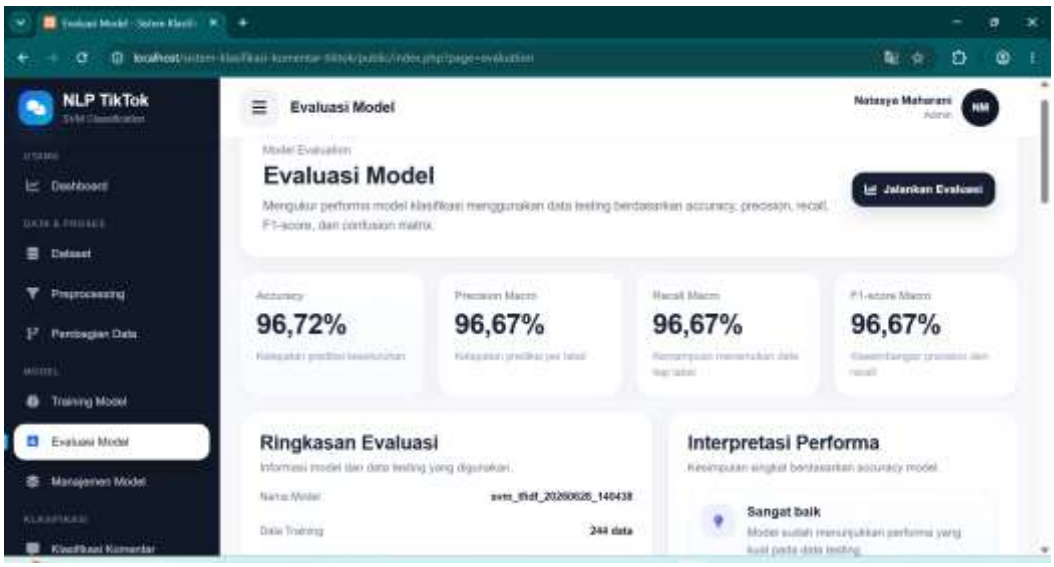


Figure 9. Model Evaluation Page

The model evaluation page displays the results of model testing against the test data. On this page, the administrator can run the evaluation process using the Run Evaluation button to determine the performance of the trained model.

This page displays a summary of the evaluation results, including accuracy, precision, recall, and F1-score values. The system also displays information about the evaluated model along with performance interpretations to provide an overview of the model's quality.

In addition to the evaluation summary, this page displays a confusion matrix and evaluation metrics for each category, allowing administrators to view the model's prediction results in greater detail. The system also provides an evaluation history, which lists previous model evaluation results to document the testing process.

8. Model Management Page

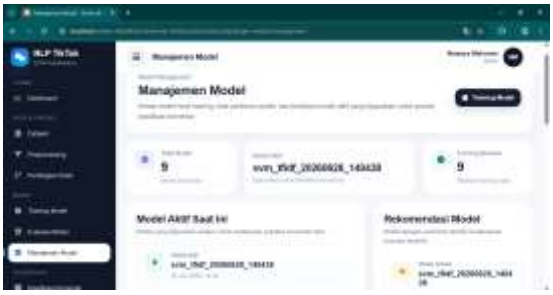


Figure 10. Model Management Page

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

International Journal of Technology and Education Research

https://e-journal.citakonsultindo.or.id/index.php/IJETER

E-ISSN: 2986-0733

The model management page is used to manage the training models stored in the system. On this page, administrators can view information about available models, active models, and the status of the model training results.

Additionally, the system displays recommendations for the best models based on the evaluation results, allowing administrators to determine which models to use in the comment classification process. This page also provides a model list containing configuration and performance information for each trained model.

Through the model management page, admins can activate the desired model for use in the classification process without the need for retraining or changing the program code manually.

9. Comment Classification Page

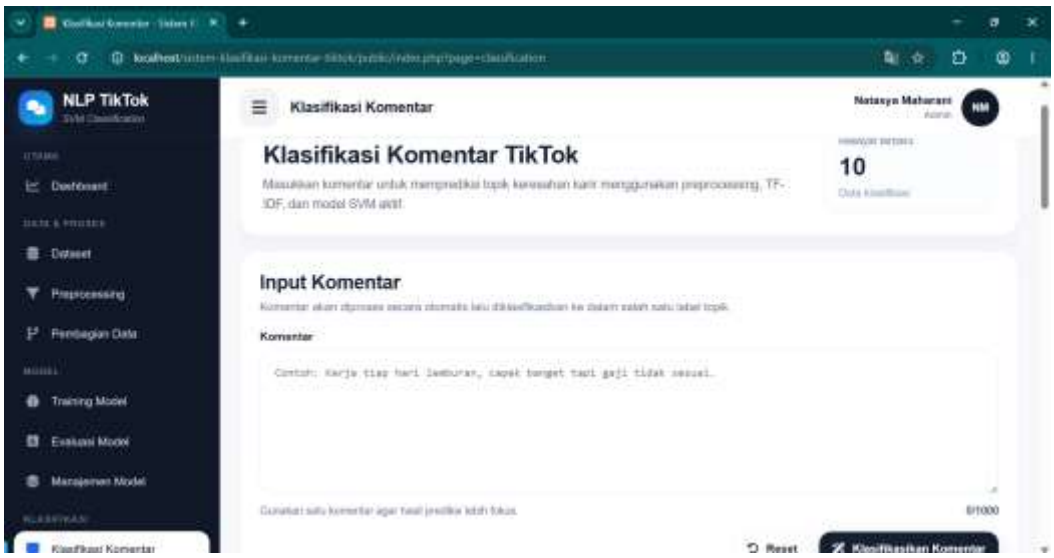


Figure 11. Comment Classification Page

The comment classification page is used to predict new comments submitted by the admin. This page features a comment input field and a Classify Comments button to initiate the classification process.

The submitted comments will undergo pre-processing, TF-IDF weighting, and classification using the active Support Vector Machine model. The classification results are displayed as comment categories based on the predictions.

This page also displays the most recent classification history, including processed comments, preprocessing results, prediction categories, the model used, and classification time. This page allows admins to test new comments and view their classification results live.

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

International Journal of Technology and Education Research

https://e-journal.citakonsultindo.or.id/index.php/IJETER

E-ISSN: 2986-0733

10. Classification History Page



Figure 12. Classification History Page

The comment history page displays all comments classified by the system. On th is page, admins can review previous classification results as documentation of the classification process.

This page features a search and filter feature based on both the label and the model used. The classification history is displayed in a table containing comments, pre-processing results, classification results, the model used, the classification date, and data details.

Through the comment history page, admins can monitor classification results, retrace comments that have been processed, and view more complete classification information.

System Test Results

Table 11. System Test Results

No	Tested Features	Testing Scenario	Expected results	Information
1	System Login	Admin enters username and password	The system successfully logged in and entered the dashboard.	In accordance
2	Dashboard	Displays a summary of system data	Statistics and information displayed	In accordance
3	Comments Dataset	Adding, changing, and deleting comment data	Data is stored and updated in the database	In accordance

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

4	Text Pre-processing	The system doespreprocessing of the dataset	Data becomes a process-ready form	In accordance
5	Data Sharing	Split the dataset into training and test data	Data is divided according to the specified ratio	In accordance
6	Training Model	Training an SVM model using training data	The model is well formed	In accordance
7	Model Evaluation	Testing model performance	Displays accuracy, precision, recall, f1-score	In accordance
8	Model Management	Saving, selecting, and activating models	Model is saved and can be used	In accordance
9	Comment Classification	Input comments for classification	The system displays the category results	In accordance
10	Comment History	Displays previous classification results	Complete history data is displayed	In accordance
11	User Management	Admin adds, changes, and deletes user data	User data is successfully saved, updated and deleted from the system.	In accordance

Based on the black-box testing results, it can be concluded that all system features are operating according to their designed functional requirements. All processes, from data management and model training to comment classification, functioned effectively.

The test results also showed that the integration between the web application, database, and SVM model ran smoothly without any significant issues. Therefore, the developed system can be used for the research objectives, namely to classify TikTok comments into Workload, Work Environment, and Work Compensation categories.

CONCLUSION

The study found that Natural Language Processing (NLP) combined with Term Frequency-Inverse Document Frequency (TF-IDF) and Support Vector Machine (SVM) can effectively classify Indonesian-language TikTok comments related to career concerns into three categories: Workload, Work Environment, and Work Compensation. Using 305 manually labeled comments and an 80:20 training-testing split, the linear SVM model with C = 1.0 achieved 96.72% accuracy, 96.67% macro precision, 96.67% macro recall, and 96.67% macro F1-score. The confusion matrix

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

showed that 59 of 61 test comments were correctly classified, while the two errors occurred between Workload and Work Environment. The model was also successfully implemented in a web-based system that supports data management, preprocessing, model training, evaluation, and classification of new comments.

However, the findings are limited by the relatively small dataset, the use of manually labeled comments, the restriction to Indonesian-language TikTok comments, and the classification of each comment into only one dominant category. Future studies should expand the dataset, include more diverse TikTok content and languages, and examine alternative classification approaches to improve model robustness and generalizability. Practically, the developed system can support more structured analysis of career concerns expressed through social media and provide an efficient tool for identifying recurring workplace-related topics.

BIBLIOGRAPHY

Abdurrazik, & Wirawana, I. M. W. (2025). Analisis klasifikasi tweet berdasarkan topik sosial menggunakan SVM. *JNATIA, 3*(4).

Agustian, A. A., & Bisri, A. (2019). Data mining optimization using sample bootstrapping and particle swarm optimization in the credit approval classification. *Indonesian Journal of Artificial Intelligence and Data Mining, 2*(1), 18–27. https://doi.org/10.24014/ijaidm.v2i1.6299

Bahram, M. (2023). Transformasi masyarakat di era digital: Menjaga kaidah hukum sebagai landasan utama. *2*(5), 1733–1746.

Fide, S., Suparti, & Sudarno. (2021). Analisis sentimen ulasan aplikasi TikTok di Google Play menggunakan metode Support Vector Machine (SVM) dan asosiasi. *Jurnal Gaussian, 10*(3), 346–358. https://ejournal3.undip.ac.id/index.php/gaussian/

Herlinda, V. A., Aditya, C. S. K., & Chandranegara, D. R. (2024). Analisis sentimen masyarakat terhadap Generasi Z dalam dunia kerja pada media sosial Twitter menggunakan metode Naïve Bayes. *REPOSITOR, 6*(4), 405–414.

Hilmi, N. F., & Irwiensyah, F. (2024). Analisis sentimen terhadap aplikasi TikTok dari ulasan pada Google Playstore menggunakan metode Naïve Bayes. *Smatika Jurnal, 14*(1), 146–156. https://doi.org/10.32664/smatika.v14i01.1210

Indrayuni, E., Nurhadi, A., & Kristiyanti, D. A. (2021). Implementasi algoritma Naive Bayes, Support Vector Machine, dan K-Nearest Neighbors untuk analisa sentimen aplikasi Halodoc. *Faktor Exacta, 14*(2), 64. https://doi.org/10.30998/faktorexacta.v14i2.9697

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.

Indriyani, F. A., Ahmad, F., & Sutan, F. (2023). Analisis sentimen aplikasi TikTok menggunakan algoritma Naïve Bayes dan Support Vector Machine. *TEKNOSAINS: Jurnal Sains, Teknologi dan Informatika, 10*(2), 176–184. https://doi.org/10.37373/tekno.v10i2.419

Kurniawan, E. P., S, I. B. H., & Irwansyah. (2021). Pengungkapan diri di situs media sosial YouTube. *Jurnal Lensa Mutiara Komunikasi, 5*(1), 42–51. https://doi.org/10.51544/jlmk.v5i1.1526

Luwei, W., & Huimin, M. (2024). From jobs to careers: Drivers and barriers to career development in emerging labor markets. *Frontiers in Sociology, 9*, Article 1486871. https://doi.org/10.3389/fsoc.2024.1486871

Mardianto, D. (2023). Komunikasi ekspresif penggunaan media sosial TikTok (Studi kasus Generasi Z usia 18–23 tahun). *IKOMIK: Jurnal Ilmu Komunikasi dan Informasi, 3*(2), 50–58. https://doi.org/10.33830/ikomik.v3i2.6481

Mutawalli, L., Zaen, M. T. A., & Bagye, W. (2019). Klasifikasi teks sosial media Twitter menggunakan Support Vector Machine (Studi kasus penusukan Wiranto). *Jurnal Informatika dan Rekayasa Elektronik, 2*(2), 43. https://doi.org/10.36595/jire.v2i2.117

Putra, T. I. Z. M., Suprapto, & Bukhori, A. F. (2022). Model klasifikasi berbasis multiclass classification dengan kombinasi IndoBERT embedding dan Long Short-Term Memory untuk tweet berbahasa Indonesia. *Jurnal Ilmu Siber dan Teknologi Digital, 1*(1), 1–28. https://doi.org/10.35912/jisted.v1i1.1509

Rifaldi, D., Fadlil, A., & Herman. (2023). Teknik preprocessing pada text mining menggunakan data tweet "mental health." *DECODE: Jurnal Pendidikan Teknologi Informasi, 3*(2), 161–171.

Wibowo, I. S., Witanti, A., & Susilawati, I. (2024). Keyword extraction judul berita online di Indonesia menggunakan metode TF-IDF. *Jurnal Teknik Informatika dan Sistem Informasi, 11*(1), 99–111.

Wicaksana, D. G. A., & Rudy, D. G. (2021). Perlindungan konsumen "curhat online" dalam platform media sosial. *Jurnal Magister Hukum Udayana (Udayana Master Law Journal), 10*(3), 644. https://doi.org/10.24843/jmhu.2021.v10.i03.p15

Copyright to the author of some rights reserved This work is licensed under aCreative Commons Attribution-ShareAlike 4.0 International License.