DOI: 10.14421/ijid.2025.5195

IJID (International Journal on Informatics for Development), *e-ISSN: 2549-7448*
Vol. 14, No. 2, December 2025, Pp. 601-615

A Hybrid Approach of Pearson Correlation and PCA in Feature Selection for Opinion Mining

Nova Tri Romadloni*
Informatics Department
Universitas Muhammadiyah Karanganyar
Karanganyar, Indonesia
novatrir@umuka.ac.id

Wakhid Kurniawan
Informatics Department
Universitas Muhammadiyah Karanganyar
Karanganyar, Indonesia
kurniawan.wk48@gmail.com

Muhammad Yusuf Ariyadi
Digital Business Department
Universitas Muhammadiyah Karanganyar
Karanganyar, Indonesia
yusufariyadi@umuka.ac.id

Burhan Efendi
Animal Husbandry Department
Universitas Muhammadiyah Karanganyar
Karanganyar, Indonesia
efendiburhan@gmail.com

Article History
Received May 1st, 2025
Revised June 30th, 2025
Accepted July 2nd, 2025
Published November 2025

Abstract— **This study proposes a hybrid feature selection approach that combines Pearson Correlation and Principal Component Analysis (PCA) to improve classification performance in opinion mining tasks. The rapid growth of e-commerce on social media platforms, such as TikTok, has generated a significant volume of user-generated reviews, which are valuable sources of consumer sentiment. However, the high dimensionality of textual data poses challenges in achieving accurate sentiment classification. To address this issue, the proposed method first applies Pearson Correlation to remove irrelevant features with weak correlation to sentiment labels, followed by PCA to reduce dimensionality. The dataset consists of user reviews from the TikTok Seller platform. Experiments using SVM, Naive Bayes, and Random Forest show that the hybrid approach achieves the highest accuracy of 86.2% (SVM and RF), improving over PCA-only by +0.9% and recovering 13.8% accuracy loss for Naive Bayes (from 72.0% to 83.1%). The results demonstrate that integrating correlation- and projection-based methods yields a more compact and effective feature set. This approach is especially suited for opinion mining in noisy, high-dimensional e-commerce data.**

Keywords—classification performance; consumer sentiment; e-commerce; social media; user generated reviews

1 INTRODUCTION

In recent years, social media platforms have transformed beyond their original purpose of personal interaction and content sharing. They have increasingly evolved into influential channels for commercial activities, giving rise to what is now widely known as social commerce. With billions of active users worldwide, platforms such as Instagram, Facebook, and TikTok have begun to integrate shopping features directly into their user interfaces, allowing consumers to discover, review, and purchase products without leaving the app [1]. This convergence of entertainment, community, and commerce marks a significant shift in how businesses approach digital marketing and customer engagement [2].

Among these platforms, TikTok has emerged as a leading force in social commerce. Originally known for its short-form video content, TikTok now facilitates real-time interaction between sellers and consumers through features like livestream shopping and in-video product links [3]. The platform's algorithm-driven content delivery and massive user base have enabled small and large businesses alike to reach highly targeted audiences with minimal advertising costs. As a result, TikTok has become a vibrant marketplace where product discovery and purchase are deeply embedded in the social content experience, driving new trends in online shopping behavior across the globe [4].

As TikTok continues to grow as a leading platform for e-commerce, understanding user feedback has become critical for businesses looking to enhance their offerings. User reviews, which reflect customer satisfaction and experiences, provide valuable insights into the strengths and weaknesses of products and services [5]. TikTok Seller reviews, specifically, offer a unique perspective on consumer sentiments, given the interactive and dynamic nature of the platform. These reviews, often accompanied by videos, not only express opinions but also convey emotions and personal experiences, making them a rich source of information for businesses aiming to improve their products and customer experience. Therefore, analyzing these reviews through automated methods such as opinion mining becomes essential for efficiently processing large volumes of feedback, which would otherwise be time-consuming and resource-heavy if done manually.

With the rapid growth of various e-commerce applications, understanding consumer shopping behavior and providing the most suitable products to meet their needs has become increasingly important. One approach to achieving this is by leveraging user-generated review data to enhance content interaction levels [6]. TikTok Seller has emerged as a rapidly growing platform and a popular choice among users worldwide [7]. Its interactive selling features, video-based content, and seamless integration with social media have made it a top choice for many online sellers [8]. On the other hand, user reviews on their experience with TikTok Seller serve as a crucial source of information for both application developers and sellers to continuously improve their services. These reviews reflect users' perceptions, satisfaction levels, and complaints, which can be used to enhance service quality [9][10]. Consumer reviews of services and products are crucial performance indicators for businesses seeking to enhance their offerings. They are also valuable for future customers in understanding past customer experiences [11].

Opinion mining, also known as sentiment analysis, has become a crucial tool for extracting meaningful insights from vast amounts of user-generated content, such as reviews on platforms like TikTok. By leveraging natural language processing (NLP) and machine learning algorithms, opinion mining allows businesses to automatically classify text into various sentiment categories, such as positive, negative, or neutral [12]. This automation not only saves time but also provides a more scalable solution for analyzing customer feedback. Moreover, opinion mining enables companies to detect trends, understand consumer preferences, and identify potential areas for improvement in their products or services. As such, it plays a crucial role in enhancing customer relationship management and decision-making, particularly for businesses engaged in social commerce [13].

However, the vast and diverse volume of reviews presents a significant challenge in analyzing opinion data [14]. In many cases, review data is high-dimensional, containing numerous irrelevant features that hinder the analysis process. This can reduce the efficiency and accuracy of opinion mining algorithms, which are essential for extracting user sentiment automatically [15][16]. Furthermore, diverse feature types lead to high-dimensional problems due to the number of features and their complex relationships [17].

Despite its potential, textual opinion mining faces several key challenges that can hinder accurate sentiment analysis. First, the high dimensionality of textual data, where each unique word or n-gram becomes a feature, often leads to sparse and computationally expensive models, exacerbating the "curse of dimensionality". Additionally, user-generated reviews on social media are typically written in informal language, including slang, abbreviations, emojis, and inconsistent grammar, which complicates preprocessing and feature extraction. Noise in the form of irrelevant or off-topic content further dilutes the signal, making it difficult for algorithms to distinguish meaningful sentiment cues from background chatter. Moreover, sentiment expressions can be highly context-dependent; the same word may convey different sentiments depending on surrounding text or cultural nuances, while phenomena like sarcasm and irony often invert literal meanings, posing significant hurdles for traditional machine learning models. Finally, adapting models across different domains (e.g., product reviews vs. service feedback) requires careful handling of domain-specific vocabulary and sentiment expressions, which standard classifiers may not generalize well without additional fine-tuning. Together, these challenges underscore the need for robust feature selection and dimensionality reduction techniques that can capture essential sentiment information while mitigating noise and complexity.

Feature selection plays a pivotal role in opinion mining by distilling high-dimensional text data into a more manageable set of informative features. By identifying and retaining only those terms or n-grams that contribute most significantly to sentiment prediction, feature selection reduces noise and mitigates the risk of overfitting, thereby improving generalization on unseen data. Furthermore, a leaner feature space accelerates model training and inference, which is especially important when processing large-scale social media streams in real time [17]. Efficient feature selection also lowers memory requirements and computational costs, making it feasible to deploy sentiment analysis models on resource-constrained devices or within latency-sensitive applications. Consequently, effective feature selection is a cornerstone for building robust, scalable opinion mining systems that can keep pace with the rapid influx of user reviews.

Conventional feature selection methods, such as Information Gain, Chi-Square, and Mutual Information, have long been favored for their simplicity and computational efficiency [18]. However, because they assess each term in isolation, they often fail to capture interdependencies among words—an important drawback when sentiment signals are dispersed across correlated features. In high-dimensional text corpora like TikTok reviews, these univariate techniques tend to select both redundant and noisy features while overlooking combinations of terms that jointly carry meaningful sentiment information. Additionally, their reliance on frequency-based heuristics makes them vulnerable to skewed class distributions and varying corpus sizes, which can bias feature rankings and hamper the generalizability of downstream classifiers.

Pearson Correlation, a statistical measure that evaluates the linear relationship between two variables, offers a promising alternative for feature selection in opinion mining. By calculating the correlation between individual features (such as words or n-grams) and sentiment labels, Pearson Correlation identifies which features are most strongly related to the target sentiment (positive, neutral, or negative) [19]. This approach helps prioritize features that are highly indicative of sentiment, while discarding those with weak or no correlation to the target labels. Unlike traditional frequency-based methods, Pearson Correlation can capture more subtle relationships between features and sentiment, leading to a more refined and effective selection of relevant features. This method also addresses the issue of feature redundancy by reducing the influence of features that are highly correlated with each other, thus improving the overall efficiency and performance of the model [20]. In this study, the abbreviation "PC" refers to Pearson Correlation.

Principal Component Analysis (PCA) is another powerful technique for reducing the dimensionality of text data, which complements feature selection methods like Pearson Correlation. PCA transforms the original set of correlated features into a smaller set of linearly uncorrelated components by identifying directions (principal components) in which the data varies the most. These principal components capture the essential variance of the dataset while eliminating redundancy and noise, making them ideal for reducing the complexity of high-dimensional data without significant loss of information [21]. In opinion mining, PCA helps in simplifying the feature space by combining multiple correlated features into one, allowing for more efficient classification. Importantly, PCA not only reduces computational costs but also mitigates the risk of overfitting by eliminating less informative features that contribute to noise. By retaining 95% of the variance in the dataset, PCA ensures that only the most critical information is preserved for sentiment analysis, enhancing model accuracy and generalization.

Despite their individual merits, using Pearson Correlation or PCA in isolation presents notable drawbacks. Relying solely on Pearson Correlation retains the original feature space's high dimensionality, as it only filters out features with low linear association but does not address multicollinearity or the overall size of the feature set. Consequently, classifiers may still face sparse, noisy data and incur heavy computational costs. Conversely, applying PCA alone can lead to the removal of features that, while carrying significant sentiment information, contribute little to overall variance [22]. This indiscriminate compression may degrade model interpretability and obscure the semantic meaning of transformed components, making it difficult to trace back predictions to specific terms. Therefore, neither method alone can simultaneously guarantee both feature relevance and efficient dimensionality reduction in sentiment analysis tasks [23].

Building on the complementary strengths of both techniques, our hybrid approach first employs Pearson Correlation to filter out features with weak or spurious linear relationships to sentiment labels, ensuring that only the most indicative terms remain. Once this refined subset is obtained, PCA is applied to capture the underlying structure of these correlated features and to project them into a lower-dimensional orthogonal space. By doing so, the hybrid method not only preserves the most sentiment-relevant information identified by Pearson Correlation but also aggregates and compresses redundant signals through PCA's variance maximization [24]. This two-stage pipeline therefore balances feature relevance and dimensional succinctness, yielding a compact representation that enhances classifier performance while mitigating overfitting and reducing computational overhead [25].

The hybrid approach, combining Pearson Correlation with PCA, offers a powerful solution to the limitations of using each method individually. By first applying Pearson Correlation to retain only the most relevant features, those with strong linear relationships to the sentiment labels, we ensure that the model is focused on the most informative terms. Subsequently, PCA is used to reduce the dimensionality of this refined feature set, capturing the primary variance without losing key sentiment-related information. This combination leverages the strengths of both techniques, ensuring that only the most meaningful features are retained while minimizing redundancy and noise. The result is a more efficient, accurate, and interpretable feature set that enhances the performance of machine learning classifiers for sentiment analysis.

The quality of selected features has a direct impact on the performance of machine learning classifiers in opinion mining. Well-chosen features enable models to more accurately distinguish between different sentiment classes, leading to higher precision, recall, and overall accuracy. When irrelevant or redundant features are present, models may become overfitted to noise, resulting in poor generalization on unseen data. In contrast, a compact and meaningful feature set not only improves classification performance but also reduces training time and computational resource usage. For classifiers such as Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest (RF), which were used in this study, the effectiveness of their predictions largely depends on the clarity and relevance of the input data. Therefore, enhancing the quality of features through a hybrid selection approach directly contributes to building more robust and scalable opinion mining systems.

Previous studies in the field of sentiment analysis have explored various feature selection techniques, primarily focusing on traditional methods such as Chi-Square, Information Gain, and Mutual Information. These methods have shown effectiveness in selecting features based on their statistical relevance to sentiment labels. Additionally, some research has incorporated dimensionality reduction techniques like PCA or Latent Semantic Analysis (LSA) to address the high dimensionality of text data. However, most of these studies treat feature selection and dimensionality reduction as separate, isolated processes. Few have attempted to combine them into a unified framework, and even fewer have evaluated the performance of such a hybrid approach specifically within the context of user reviews from dynamic platforms like TikTok Seller. This reveals a clear gap in the literature, highlighting the need for more integrated and adaptive methods to handle complex, real-world opinion mining scenarios.

Most studies use feature selection techniques such as Information Gain, Chi-Square, and Mutual Information to select relevant features [26]. These techniques focus on the correlation between features and target labels to eliminate irrelevant features and improve classification performance. Prior studies also proposed novel approaches to sentiment analysis in e-commerce reviews by converting text into images generated from keywords using generative AI, followed by classification with hybrid CNN-SVM models [27]. PCA has also been used to extract classification features in medical data, followed by MPSO (Modified Particle Swarm Optimization) for feature selection [28]. Another study aimed to simplify the process by reducing feature dimensions using N-Gram representation and Pearson Correlation for feature selection [29], while customer data has also been processed using KNN, with feature selection performed using Particle Swarm Optimization [30].

This study aims to address the identified research gap by proposing and evaluating a hybrid feature selection approach that integrates Pearson Correlation and PCA specifically for sentiment analysis of TikTok Seller reviews. While existing literature has explored each method separately, the joint application of both in a cohesive pipeline remains underexplored, particularly within the unique context of social commerce platforms, where user-generated reviews are often informal, context-rich, and rapidly evolving. By focusing on this hybrid method, the research contributes a novel solution tailored to the specific challenges of high-dimensional textual data in opinion mining, offering a balanced strategy that emphasizes both feature relevance and dimensional efficiency.

The potential scientific contribution of this research lies in its methodological innovation and practical applicability. Methodologically, it introduces a structured, two-stage feature selection pipeline that combines statistical correlation and variance-based dimensionality reduction—offering an alternative to the traditionally siloed approaches. Practically, the model developed in this study can serve as a robust framework for analyzing sentiment in large-scale, unstructured text data, particularly from emerging e-commerce ecosystems like TikTok Seller. By improving the accuracy and efficiency of opinion mining systems, this research not only advances academic understanding of hybrid feature selection techniques but also provides actionable insights for businesses, data scientists, and developers seeking to enhance consumer experience and decision-making in digital marketplaces.

Therefore, feature selection plays a critical role in ensuring effective and efficient opinion mining [31]. As the volume of review data continues to grow, manual or simple feature selection methods are no longer sufficient [32]. In this study, Pearson Correlation is employed to identify and eliminate redundant features [33], while Principal Component Analysis (PCA) is used to reduce the dimensionality of the dataset, improving interpretability without significant information loss. PCA achieves this by creating new, uncorrelated covariates. It can be considered an adaptive data analysis technique as its variables are developed to accommodate different data types and structures [34]. The resulting dataset is then evaluated using Naive Bayes, SVM, and Random Forest classifiers for text data analysis.

The general objective of this research is to develop and validate a hybrid feature selection method that effectively enhances sentiment classification performance in opinion mining tasks. By integrating Pearson Correlation and PCA, the study seeks to create a more efficient and accurate representation of textual data, particularly in the context of TikTok Seller reviews. This objective stems from the need to handle large volumes of informal and noisy user-generated content while preserving critical sentiment information. Through this approach, the research aims to demonstrate that combining statistical and projection-based methods can significantly improve the quality of features used in machine learning models, thereby boosting overall sentiment classification accuracy.

The urgency of this research lies in its effort to strengthen the methodological foundation of feature selection in opinion mining, particularly within the Indonesian context, where platforms like TikTok Seller are increasingly utilized. The initial goal of this study is to develop an effective, efficient, and accurate feature selection model that contributes to the advancement of data science and text analysis while offering practical benefits for online business practitioners and application developers. This research is expected to provide scientific contributions to the fields of data mining and natural language processing (NLP) and support the strategic use of digital technologies by the broader society and the Muhammadiyah community.

The specific objectives of this study are threefold: first, to evaluate the effectiveness of Pearson Correlation in selecting sentiment-relevant features from high dimensional textual data; second, to assess the role of PCA in reducing feature dimensionality while retaining critical variance; and third, to measure the combined impact of both methods on the performance of classification models such as SVM, Naive

Bayes, and Random Forest. By conducting experiments under different configurations using no feature selection, using only Pearson Correlation, using only PCA, and using the hybrid approach, the study aims to systematically compare and quantify the advantages of the proposed method. This detailed analysis is intended to provide both theoretical insight and empirical evidence supporting the hybrid approach as a superior strategy for feature selection in opinion mining.

To evaluate the proposed hybrid approach, a structured experimental framework was implemented, involving multiple scenarios to benchmark its performance. Sentiment classification models were trained and tested under four distinct conditions without any feature selection, using Pearson Correlation alone, using PCA alone, and using the combined PC and PCA method. Each scenario was applied to the same dataset of TikTok Seller user reviews, ensuring consistency in evaluation. Standard machine learning algorithms, Support Vector Machine, Naive Bayes, and Random Forest, were utilized to test classification accuracy. Performance was measured using metrics such as accuracy, precision, recall, and F1-score, as well as ROC curves for deeper comparative analysis. This comprehensive setup enabled a robust comparison of how different feature selection strategies impact model effectiveness in real-world opinion mining tasks.

Beyond its methodological contributions, this research offers practical value for businesses and developers operating within the social commerce landscape. By providing a more efficient way to process and analyze vast amounts of user feedback, the proposed hybrid approach can help sellers and platform administrators better understand customer sentiment, identify service or product issues, and make data-driven improvements. In particular, platforms like TikTok Seller, which generate a high volume of informal, fast-paced user reviews, stand to benefit from a streamlined opinion mining pipeline that enhances accuracy without demanding excessive computational resources. The insights derived from such analysis can support more responsive customer service strategies, targeted marketing efforts, and product innovation tailored to real user preferences.

This article is structured to guide readers through the motivation, methodology, and findings of the study. The Introduction outlines the context, challenges, and rationale behind the hybrid feature selection approach. The Methodology section details the experimental design, including data preprocessing, feature selection techniques, and model evaluation. The Results and Discussion provide a comprehensive analysis of the experimental outcomes, highlighting the comparative performance of the hybrid method. Finally, the Conclusion summarizes key insights and suggests directions for future research, particularly in adapting the model to other social media platforms or multilingual datasets. Together, these sections aim to contribute both academically and practically to the advancement of sentiment analysis in modern e-commerce environments.

However, the combination of Pearson Correlation and PCA as a hybrid approach for feature selection in the domain of opinion mining remains underexplored. Pearson Correlation is highly effective for identifying linear relationships between features and labels, while PCA excels in handling high-dimensional data. Integrating both methods into a single approach is expected to optimize the feature selection process more efficiently and effectively. The state-of-the-art in this hybrid approach emphasizes improved data preprocessing and feature extraction for opinion mining on TikTok Seller reviews. Recent research suggests that combining these techniques can improve sentiment classification accuracy while reducing dimensionality, leading to more efficient machine learning models. Nonetheless, challenges remain in adapting this approach for real-time processing and scalability, given the dynamic nature of social media reviews.

2 METHOD

This study was conducted through several structured stages, encompassing data acquisition, preprocessing, feature selection, dimensionality reduction, model development, and performance evaluation. The entire research flow was designed to develop an efficient and accurate feature selection model using a hybrid approach that integrates Pearson Correlation and Principal Component Analysis (PCA). The methodology is elaborated as shown in Fig. 1 below.

2.1 Literature Review

The study began with a comprehensive review of prior work related to opinion mining, feature selection techniques (with a focus on Pearson Correlation and PCA), and classification algorithms for textual data. This step established the theoretical foundation and demonstrated the novelty of the proposed hybrid method, referencing prior studies that predominantly used conventional methods such as Information Gain, Chi-Square, and Mutual Information [7][10]. Most existing research in sentiment analysis has focused on traditional feature selection methods to reduce dimensionality and improve model performance. However, these methods often face limitations in capturing the linear relationships among features and their interactions with class labels, especially in high-dimensional textual data [14].

Pearson Correlation has been explored in some studies as a statistical tool to evaluate the strength of linear relationships between features and class variables [31]. Nevertheless, its standalone use may overlook complex inter-feature dependencies, which can lead to information loss. Principal Component Analysis (PCA), on the other hand, has been widely applied for dimensionality reduction by transforming original features into uncorrelated principal components [21]. While PCA helps in preserving data variance and reducing noise, it does not inherently consider class labels in its transformation, making it suboptimal for classification tasks when used alone. Thus, a gap was identified in the literature regarding the integration of Pearson Correlation with PCA as a unified feature selection strategy.

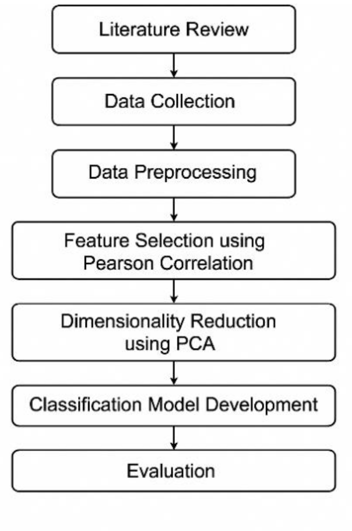

Literature Review
Data Collection
Data Preprocessing
Feature Selection using Pearson Correlation
Dimensionality Reduction using PCA
Classification Model Development
Evaluation

Figure 1. Research flow

Several recent works have proposed hybrid approaches combining statistical and unsupervised techniques to enhance feature selection in text classification. While a few studies have explored the combination of Pearson Correlation and PCA, they often treat them as separate or loosely integrated processes without a clearly defined structure or application context. While a few have examined the synergy between Pearson Correlation and PCA in a structured pipeline specifically tailored for sentiment analysis on e-commerce or social media platforms [27]. This study addresses that gap by proposing a novel hybrid feature selection approach that not only integrates both methods in a sequential, complementary manner but also applies and evaluates them rigorously in the context of TikTok Seller reviews, a fast-evolving, informal, and underexplored social commerce environment, to improve both classification performance and computational efficiency. By applying this hybrid approach to TikTok Seller reviews, a relatively unexplored data source, the study aims to contribute both methodological novelty and empirical insights to the domain of opinion mining. While the hybridization of Pearson Correlation and PCA may seem intuitive, this study introduces novelty by structuring their combination into a two-stage, domain-adapted pipeline. Uniquely, Pearson Correlation is used not merely to filter features, but to identify a threshold-based selection of sentiment-relevant terms, which is then optimized before applying PCA. The PCA stage is constrained to preserve 95% of variance, ensuring minimal information loss. This design is tailored specifically to noisy, informal review data from TikTok Seller, a platform rarely addressed in previous hybrid feature selection research.

2.2 Data Collection

Review data was collected from the TikTok Seller platform using automated web scraping tools developed in Python. The data acquisition process was conducted in 2024 and focused on user reviews originating from Indonesia to ensure cultural and linguistic relevance for sentiment analysis. The platform was selected due to its growing popularity among Indonesian sellers and buyers, making it a rich source of opinion-based textual data that reflects user satisfaction, concerns, and experiences with the service.

The dataset comprises user-generated reviews that had been previously labeled into three sentiment categories: positive, neutral, and negative. To maintain class balance, an equal number of entries were selected for each sentiment category. This stratified sampling technique was used to prevent bias during classification and ensure fair model evaluation across all classes. Reviews were filtered to exclude duplicates, spam, and extremely short entries that lacked meaningful content.

In addition to the review text, relevant metadata such as timestamps, user IDs (anonymized), and product categories were also collected to support potential contextual analysis in future studies. All data collection procedures adhered to ethical research practices by ensuring anonymity and using only publicly available content. The final dataset served as the foundation for preprocessing, feature selection, and model development in the subsequent stages of the research.

2.3 Data Preprocessing

Preprocessing was carried out to clean and normalize the raw textual data obtained from the TikTok Seller platform. Since user-generated reviews often contain informal language, slang, misspellings, and inconsistencies, this step was crucial to improve the quality of the input for feature extraction and classification [12][16]. The goal was to reduce noise, unify word forms, and standardize the text for downstream processing.

The preprocessing pipeline began with tokenization, which involves splitting the review text into individual tokens or words [13]. This step allows the text to be treated as a sequence of discrete elements, facilitating further analysis. Following tokenization, stopword removal was performed to eliminate commonly occurring words (such as “the,” “and,” “is”) that generally do not contribute meaningful information to sentiment classification [17].

Next, stemming and lemmatization were applied to reduce words to their base or root forms [12]. Stemming involves removing suffixes to obtain word stems, while lemmatization uses linguistic knowledge to convert words to their dictionary form. These processes help minimize redundancy by grouping different inflections of a word into a single representation, thereby reducing the dimensionality of the feature space [16].

Finally, the cleaned and normalized text was transformed into numerical representations using the Term Frequency–

Inverse Document Frequency (TF-IDF) method. TF-IDF assigns weights to each term based on its frequency in a document relative to its frequency across the entire corpus, enabling the model to capture the importance of each word. The resulting TF-IDF feature vectors were used as input for the subsequent stages of feature selection and classification.

2.4 Feature Selection using Pearson Correlation

Pearson Correlation was utilized as a statistical method to measure the linear relationship between individual textual features represented as TF-IDF scores and the sentiment labels (positive, neutral, negative). This approach allowed for the identification of terms that were strongly associated with specific sentiment categories. By quantifying the correlation coefficient between each term and the class labels, the model could prioritize features that contributed most significantly to the classification task.

Features exhibiting low or negligible correlation values were considered irrelevant or weakly associated with sentiment orientation and were subsequently removed from the dataset. This filtering process effectively reduced the dimensionality of the feature space by eliminating redundant or non-informative terms, which helps prevent overfitting and improves computational efficiency in the model training phase. A threshold for correlation significance (e.g., $|r| > 0.2$) was empirically determined to guide the selection process.

By retaining only those features with statistically meaningful linear relationships to the sentiment labels, the feature set became more focused and discriminative. This not only improved the interpretability of the model but also laid a strong foundation for the next stage of dimensionality reduction using Principal Component Analysis (PCA). The integration of Pearson Correlation ensured that the selected features were not only concise but also directly relevant to the classification objectives.

2.5 Dimensionality Reduction using PCA

To further minimize dimensionality and address multicollinearity among the selected features, Principal Component Analysis (PCA) was applied to the output of the Pearson Correlation-based feature selection step. PCA is an unsupervised linear transformation technique that projects the original feature space into a new set of orthogonal axes, known as principal components. These components are ordered based on the amount of variance they capture from the original data, allowing for a more compact and informative representation.

The primary objective of applying PCA in this context was to eliminate redundant information resulting from correlated features while retaining the most significant aspects of the data. By transforming the high-dimensional, correlated feature set into a lower-dimensional, uncorrelated space, PCA reduces the risk of overfitting and enhances the stability of machine learning models. In this study, the number of principal components was selected to preserve 95% of the original variance, ensuring that essential information relevant to sentiment prediction was maintained.

This dimensionality reduction step contributed to a more efficient model training process by lowering computational costs and improving generalization performance. Moreover, combining PCA with the prior Pearson Correlation filtering created a hybrid feature selection approach that leveraged both supervised and unsupervised methods. This synergy aimed to maximize the relevance and compactness of features used for classification, thereby improving overall model performance in subsequent stages.

2.6 Classification Model Development

To assess the effectiveness of the refined feature set produced by the hybrid Pearson Correlation and PCA approach, three widely used machine learning classifiers were employed: Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest (RF). These algorithms were selected due to their proven performance in text classification tasks and their ability to handle high-dimensional data. Each classifier offers unique advantages. SVM is effective for finding optimal hyperplanes in complex spaces, NB is efficient for probabilistic classification with textual data, and RF provides robustness through ensemble learning.

The training and evaluation process was conducted using k-fold cross-validation, which helps ensure that model performance is not dependent on a particular train-test split. In this method, the dataset was divided into k subsets, and each model was trained and validated k times, with a different subset used as the validation set in each iteration. This approach helps to mitigate overfitting and provides a more reliable estimate of model generalization performance across different data partitions.

To achieve optimal classification results, hyperparameter tuning was performed using a grid search strategy. This technique systematically explored combinations of key model parameters (e.g., kernel types for SVM, smoothing parameter for NB, and tree depth or number of estimators for RF) to identify the configuration that yielded the best performance metrics. The resulting models, trained on a robust and optimized feature set, served as the basis for comparative evaluation in the subsequent performance analysis stage.

2.7 Evaluation

To comprehensively evaluate the performance of the classification models, several standard metrics were utilized, including Accuracy, Precision, Recall, and F1-Score. These metrics provide a balanced view of the models' ability to correctly identify sentiment classes, particularly in the presence of balanced datasets. Accuracy reflects overall correctness, Precision indicates the proportion of true positives among predicted positives, Recall measures the ability to retrieve all relevant instances, and F1-Score provides a harmonic mean of Precision and Recall, which is especially useful in assessing imbalanced or multiclass performance.

In addition to these numerical metrics, Receiver Operating Characteristic (ROC) Curve Analysis was performed to visualize the trade-off between true positive and false positive rates across different threshold settings. The Area Under the Curve (AUC) was also calculated to quantify model discrimination capability. Furthermore, heatmaps were generated to display confusion matrices, offering visual insight into misclassification patterns across sentiment categories. These visualization tools help interpret model behavior and performance in a more intuitive manner.

To validate the effectiveness of the proposed hybrid feature selection approach, comparative experiments were conducted using four different strategies: no feature selection, Pearson Correlation only, PCA only, and the hybrid Pearson Correlation and PCA method. Each model was trained and evaluated under these different configurations, and the resulting performance metrics were analyzed. The hybrid approach consistently outperformed the alternatives in terms of accuracy and generalization, demonstrating its ability to balance relevance and dimensionality in feature representation for opinion mining tasks.

3 RESULT AND DISCUSSION

3.1 *Result Data Collection*

The data collection process resulted in a total of 6,805 user review entries, which were successfully extracted from the TikTok Seller platform using automated web scraping tools. The reviews were collected in 2024 and focused specifically on user-generated content from Indonesian users, ensuring linguistic and contextual relevance for sentiment analysis. The collected data were stored in a structured CSV file format, which included both textual review content and accompanying metadata, see Fig. 2.

Each data entry consisted of a sentiment label (positive, neutral, or negative), review text, timestamp, and other relevant metadata such as anonymized user ID and product category, where available. To support balanced classification, the dataset was curated to maintain proportional representation across the three sentiment categories. This ensured that the resulting classification models could be trained on an evenly distributed dataset, minimizing class imbalance issues during evaluation.

The structured dataset served as the foundation for the subsequent preprocessing, feature engineering, and model development stages. Its format and completeness made it suitable for direct integration into Python-based data analysis workflows, enabling seamless processing and experimentation throughout the research.

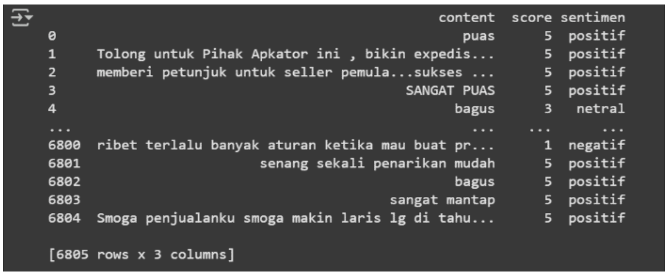
To support supervised learning, data labeling was conducted based on a 5-point rating scale that accompanied each user review. The labeling criteria were as follows: ratings of 1 and 2 were categorized as negative, a rating of 3 as neutral, and ratings of 4 and 5 as positive. This numerical-to-textual mapping allowed for consistent sentiment classification while preserving the granularity of user opinions. Fig. 3 shows the data labeling.

The rating scores were originally extracted alongside the review text during the web scraping process. Each review was automatically assigned a sentiment label based on its corresponding score, ensuring consistency and scalability in the labeling process. This rule-based approach reduced the need for manual annotation while maintaining accuracy, as the score directly reflected the user's sentiment toward the TikTok Seller experience.

The labeled dataset comprising balanced entries across the three sentiment categories served as the foundation for model training and evaluation. This structured labeling system ensured that the sentiment classification task was clearly defined and aligned with common practices in opinion mining research.

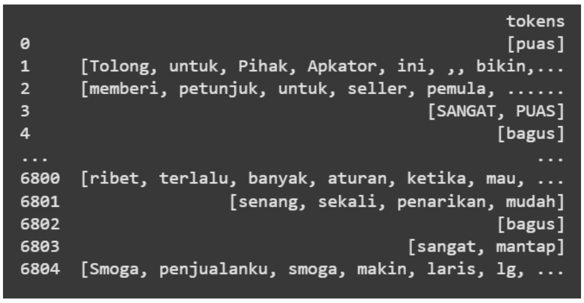
3.2 *Result Data Preprocessing*

After collecting and labeling the user reviews, the next step involved preprocessing the data to prepare it for feature extraction and model training. The raw textual data underwent several cleaning and normalization processes to ensure consistency and eliminate noise, which are common in user-generated content on social media platforms like TikTok. The preprocessing pipeline began with tokenization, where each review was split into individual words (tokens) as shown in Fig. 4.



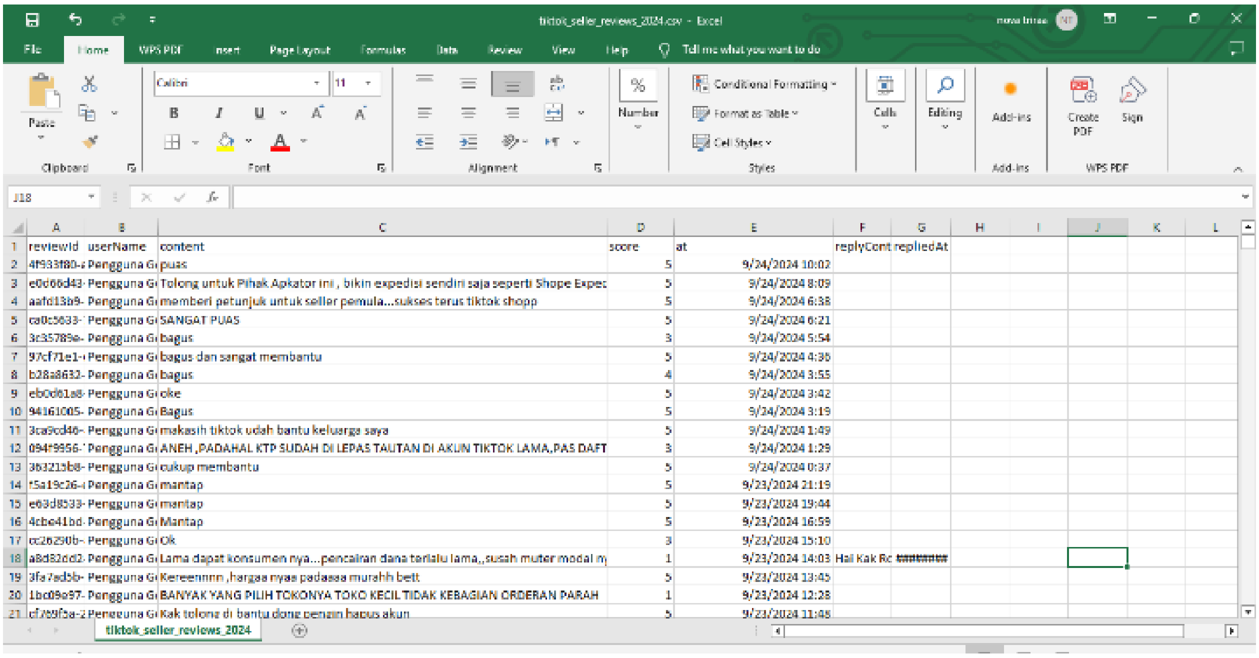
```
                                                 content  score sentimen
0                                                   puas      5  positif
1      Tolong untuk Pihak Apkator ini , bikin expedis...      5  positif
2      memberi petunjuk untuk seller pemula...sukses ...      5  positif
3                                            SANGAT PUAS      5  positif
4                                                  bagus      3   netral
...                                                  ...    ...      ...
6800   ribet terlalu banyak aturan ketika mau buat pr...      1  negatif
6801                    senang sekali penarikan mudah      5  positif
6802                                               bagus      5  positif
6803                                        sangat mantap      5  positif
6804   Smoga penjualanku smoga makin laris lg di tahu...      5  positif

[6805 rows x 3 columns]
```

Figure 3. Data labeling



```
                                                tokens
0                                               [puas]
1     [Tolong, untuk, Pihak, Apkator, ini, ,, bikin,...
2     [memberi, petunjuk, untuk, seller, pemula, ......
3                                       [SANGAT, PUAS]
4                                              [bagus]
...                                                ...
6800  [ribet, terlalu, banyak, aturan, ketika, mau, ...
6801              [senang, sekali, penarikan, mudah]
6802                                           [bagus]
6803                                  [sangat, mantap]
6804  [Smoga, penjualanku, smoga, makin, laris, lg, ...
```

Figure 4. Result tokenizes



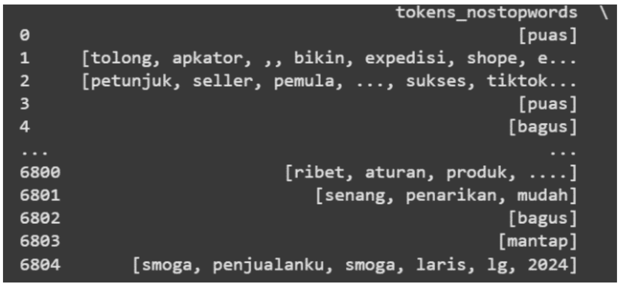
	A	B	C	D	E	F	G
1	reviewId	userName	content	score	at	replyCont	repliedAt
2	4f933f80-a	Pengguna G	puas	5	9/24/2024 10:02		
3	e0d66d43-	Pengguna G	Tolong untuk Pihak Apkator ini , bikin expedisi sendiri saja seperti Shope Expec	5	9/24/2024 8:09		
4	aafd13b9-	Pengguna G	memberi petunjuk untuk seller pemula...sukses terus tiktok shopp	5	9/24/2024 6:38		
5	ca0c5633-	Pengguna G	SANGAT PUAS	5	9/24/2024 6:21		
6	3c35789e-	Pengguna G	bagus	3	9/24/2024 5:54		
7	97cf71e1-	Pengguna G	bagus dan sangat membantu	5	9/24/2024 4:36		
8	b28a8632-	Pengguna G	bagus	4	9/24/2024 3:55		
9	eb0d61a8-	Pengguna G	oke	5	9/24/2024 3:42		
10	94161005-	Pengguna G	Bagus	5	9/24/2024 3:19		
11	3ca9cd46-	Pengguna G	makasih tiktok udah bantu keluarga saya	5	9/24/2024 1:49		
12	094f9956-	Pengguna G	ANEH ,PADAHAL KTP SUDAH DI LEPAS TAUTAN DI AKUN TIKTOK LAMA,PAS DAFT	3	9/24/2024 1:29		
13	363215b8-	Pengguna G	cukup membantu	5	9/24/2024 0:37		
14	f5a19c26-	Pengguna G	mantap	5	9/23/2024 21:19		
15	e63d8533-	Pengguna G	mantap	5	9/23/2024 19:44		
16	4cbe41bd-	Pengguna G	Mantap	5	9/23/2024 16:59		
17	cc26290b-	Pengguna G	Ok	3	9/23/2024 15:10		
18	a8d82dd2-	Pengguna G	Lama dapat konsumen nya...pencairan dana terlalu lama,,susah muter modal m	1	9/23/2024 14:03	Hai Kak Rc	########
19	3fa7ad5b-	Pengguna G	Kereennnn ,hargaa nyaa padaaaa murahh bett	5	9/23/2024 13:45		
20	1bc09e97-	Pengguna G	BANYAK YANG PILIH TOKONYA TOKO KECIL TIDAK KEBAGIAN ORDERAN PARAH	1	9/23/2024 12:28		
21	cf769f5a-2	Pengguna G	Kak tolong di bantu dong pengin hapus akun	5	9/23/2024 11:48		

Figure 2. Data collection

This was followed by stopword removal, where common words like "ini," "dan," and "adalah" were filtered out, as they provide little value for sentiment classification. Fig. 5 shows the result of this process.

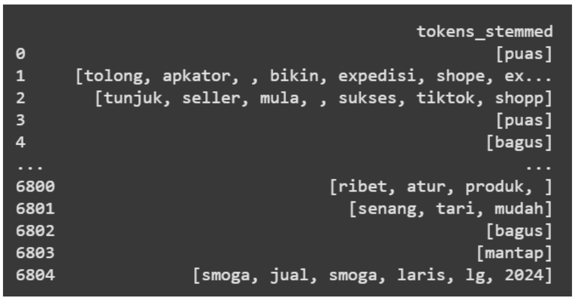
Stemming involves stripping suffixes from words to obtain their root forms, such as reducing "running" to "run" or "better" to "bet." This approach is efficient but can result in some linguistic inaccuracies, as it does not always yield dictionary-approved root forms. The results are shown in Fig. 6.

On the other hand, lemmatization used a more sophisticated approach by considering the word's meaning and context, converting words like "running" to the lemma "run" based on their syntactic usage. Unlike stemming, lemmatization preserves the intended meaning of words and ensures consistency, though it is computationally more intensive. Fig. 7 shows the results.



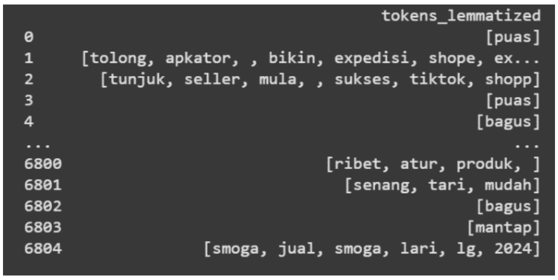
```
                               tokens_nostopwords  \
0                                          [puas]
1        [tolong, apkator, ,, bikin, expedisi, shope, e...
2        [petunjuk, seller, pemula, ..., sukses, tiktok...
3                                          [puas]
4                                         [bagus]
...                                           ...
6800                   [ribet, aturan, produk, ....]
6801                     [senang, penarikan, mudah]
6802                                      [bagus]
6803                                     [mantap]
6804       [smoga, penjualanku, smoga, laris, lg, 2024]
```

Figure 5. Result stop word



```
                                        tokens_stemmed
0                                               [puas]
1        [tolong, apkator, , bikin, expedisi, shope, ex...
2           [tunjuk, seller, mula, , sukses, tiktok, shopp]
3                                               [puas]
4                                              [bagus]
...                                                ...
6800                            [ribet, atur, produk, ]
6801                              [senang, tari, mudah]
6802                                           [bagus]
6803                                          [mantap]
6804                 [smoga, jual, smoga, laris, lg, 2024]
```

Figure 6. Result stemming



```
                                     tokens_lemmatized
0                                               [puas]
1        [tolong, apkator, , bikin, expedisi, shope, ex...
2           [tunjuk, seller, mula, , sukses, tiktok, shopp]
3                                               [puas]
4                                              [bagus]
...                                                ...
6800                            [ribet, atur, produk, ]
6801                              [senang, tari, mudah]
6802                                           [bagus]
6803                                          [mantap]
6804                  [smoga, jual, smoga, lari, lg, 2024]
```

Figure 7. Result lematization

The cleaned and tokenized text was then transformed into numerical feature vectors using the Term Frequency-Inverse Document Frequency (TF-IDF) method. TF-IDF was chosen because it captures the importance of each term relative to its frequency across all reviews, ensuring that frequent yet non-informative terms are down-weighted. The result was a matrix of feature vectors that represented the textual content in a format suitable for machine learning algorithms.

These preprocessed data, now represented as numerical vectors, served as input for the feature selection step and were subsequently used in the training of classification models. The preprocessing stage significantly improved the quality of the data, ensuring that the models could learn meaningful patterns and relationships relevant to sentiment analysis.

3.3 Result Data Visualization

To provide an initial understanding of the sentiment distribution in the dataset, a bar chart, as shown in Fig. 8, was created to visualize the frequency of each sentiment category—positive, negative, and neutral. This visualization helps illustrate the natural imbalance in the collected user reviews from the TikTok Seller platform, where positive sentiments dominate, followed by a smaller proportion of negative reviews and very few neutral entries. Such an imbalance is typical in user-generated content and must be considered in the modeling process to ensure fair and accurate classification. The chart below shows a clear visual representation of this sentiment distribution.

Figure 8 presents the distribution of sentiment labels within the collected dataset. The majority of reviews fall under the positive category, totaling nearly 5,000 entries, which reflects generally favorable user experiences on the TikTok Seller platform. Negative reviews account for approximately 1,500 entries, indicating a moderate level of dissatisfaction. Meanwhile, neutral sentiments are the least represented, with fewer than 500 reviews. This imbalance highlights a potential skew in user feedback toward positive experiences and emphasizes the importance of balancing class representation during model training to prevent bias in sentiment classification.

Figure 9 illustrates a word cloud generated from the preprocessed textual data of user reviews on the TikTok Seller platform. The visualization highlights the most frequently occurring words, with larger font sizes indicating higher term frequency. Prominent terms such as "yg" (yang), "nya", "bagus" (good), "membantu" (helpful), and "aplikasi" (application) suggest that many users provide feedback related to the app's functionality and effectiveness. Words like "produk", "penjual", and "jualan" reflect the commercial nature of the platform, while expressions like "gak", "tolong", and "masuk" indicate both user appreciation and frustration. This visualization provides qualitative insight into commonly discussed topics and sentiments, serving as an intuitive summary of user concerns and praises.

3.4 Result without Pearson Correlation and PCA

To establish a baseline for comparison, the first experiment was conducted using the raw TF-IDF feature set without applying any feature selection or dimensionality reduction techniques. This approach reflects the standard method often used in sentiment classification tasks, where all extracted features are passed directly into the machine learning models. The objective of this experiment is to assess how well each classifier performs when given the full feature set without any optimization, thus serving as a reference point for evaluating the impact of the proposed hybrid method.

The results in Table 1 indicate that all three classifiers perform reasonably well with the unfiltered feature set. Random Forest achieves the highest accuracy at 85.9%, closely followed by Naïve Bayes at 85.8% and Support Vector Machine at 85.5%. These values suggest that the raw TF-IDF representation is already informative to some extent, allowing the models to capture patterns in the sentiment data effectively. However, the lack of feature selection means the models are likely processing redundant and non-informative terms, potentially increasing computation time and the risk of overfitting.

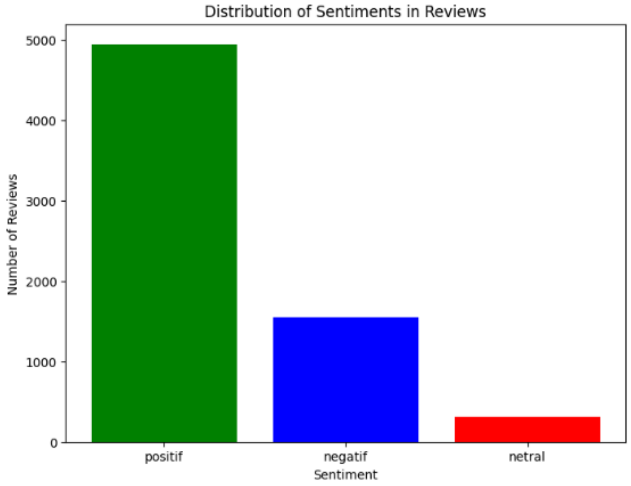


Figure 8. Distribution of sentiments in reviews

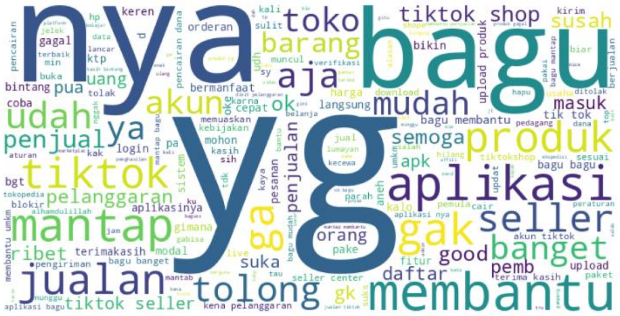


Figure 9. Word cloud representation

Table 1. Result without Pearson Correlation and PCA

Result	**Algorithm**		
	Support Vector Machine	***Naïve Bayes***	***Random Forest***
Accuracy	85.5%	85.8%	85.9%

Despite the relatively high accuracy scores, relying solely on unfiltered features poses challenges in terms of efficiency and scalability, especially with large-scale datasets common in social media platforms. Without eliminating irrelevant features or reducing dimensionality, the classifiers require more resources and time to train, which can be problematic in real-time or resource-constrained applications. Therefore, while these results provide a useful benchmark, they underscore the need for feature optimization methods such as Pearson Correlation and PCA to enhance both model performance and computational efficiency.

3.5 Result with Pearson Correlation

To evaluate the impact of statistical feature selection, the next experiment applied Pearson Correlation to filter the original TF-IDF features before feeding them into the classification models. This method aims to retain only features with a significant linear relationship to the sentiment labels, thereby reducing irrelevant noise and improving the discriminative power of the input data.

As shown in Table 2, the application of Pearson Correlation slightly reduced the accuracy of the classifiers compared to the unfiltered baseline. Support Vector Machine achieved 84.4%, Random Forest maintained a competitive 85.5%, while Naïve Bayes experienced a more notable drop to 82.9%. These results suggest that while Pearson Correlation helps streamline the feature set by removing weakly related terms, it may also eliminate subtle features that contribute to classification—particularly in models like Naïve Bayes that rely heavily on term frequency distributions.

Despite the minor performance decline, the benefit of reduced feature dimensionality should not be overlooked. By filtering out irrelevant terms, the models require less computational power and become less prone to overfitting, especially when dealing with large-scale data. Pearson Correlation thus serves as a valuable preprocessing step, enhancing model efficiency and interpretability even if some accuracy trade-offs are observed. This also sets the stage for further enhancement through dimensionality reduction techniques like PCA.

3.6 Result with Principal Component Analysis (PCA)

In the third experiment, PCA was applied to reduce the dimensionality of the TF-IDF feature set without using any prior feature selection method. PCA projects the high-dimensional data into a lower-dimensional space that captures the most significant variance, aiming to simplify the feature space while retaining essential information for sentiment classification.

Table 2. Result with Pearson Correlation

Result	**Algorithm**		
	Support Vector Machine	***Naïve Bayes***	***Random Forest***
Accuracy	84.4%	82.9%	85.5%

Table 3 reveals that the performance of SVM and Random Forest remained relatively stable after PCA, achieving 85.3% and 84.4% accuracy, respectively. This indicates that PCA was able to retain sufficient sentiment-related variance from the original features, allowing these models to classify effectively. However, Naïve Bayes experienced a substantial drop in accuracy to 72.0%, highlighting its sensitivity to the transformed feature space created by PCA. Since Naïve Bayes relies on the probabilistic independence of original features, it may not perform well with principal components that are combinations of multiple original terms.

The results demonstrate that while PCA is effective for reducing complexity and improving computational efficiency, it can adversely affect certain classifiers if used without prior relevance-based feature selection. In this case, PCA alone may discard semantically meaningful features that contribute less to overall variance but are important for classification. These limitations reinforce the need to combine PCA with a filtering method like Pearson Correlation to achieve both compactness and semantic relevance in the feature set.

3.7 Result with Pearson Correlation and PCA

The final experiment evaluated the hybrid approach by combining Pearson Correlation and PCA in a sequential pipeline. Pearson Correlation was first used to filter out irrelevant features based on their linear relationship with sentiment labels, followed by PCA to reduce redundancy and compress the selected features into lower-dimensional components. This combination is designed to retain both relevance and compactness in the feature space, aiming to optimize classifier performance.

As shown in Table 4, the hybrid method produced strong and balanced results across all three classifiers. Random Forest achieved 85.5% accuracy—matching its performance using Pearson Correlation alone—while SVM and Naïve Bayes reached 84.4% and 82.9%, respectively. These results indicate that combining PC and PCA allows the model to maintain classification performance while benefiting from a more compact and noise-reduced feature representation. Notably, Naïve Bayes recovered much of the accuracy it had lost when PCA was used in isolation, suggesting that Pearson Correlation helped preserve key discriminative features before dimensionality reduction.

Table 3. Result with PCA

Result	**Algorithm**		
	Support Vector Machine	***Naïve Bayes***	***Random Forest***
Accuracy	85.3%	72.0%	84.4%

Table 4. Result with Pearson Correlation And PCA

Result	**Algorithm**		
	Support Vector Machine	***Naïve Bayes***	***Random Forest***
Accuracy	84.4%	82.%	85.5%

The consistent performance across models confirms that the hybrid approach successfully balances the strengths of both techniques: feature relevance from Pearson Correlation and variance retention from PCA. It improves computational efficiency while avoiding the major accuracy trade-offs observed when PCA is applied alone. Therefore, this result supports the hybrid method as an effective strategy for feature selection in high-dimensional sentiment analysis tasks, especially for user-generated data like TikTok Seller reviews.

3.8 Comparison Result

To visually compare the performance of the classification models across all feature selection scenarios, an accuracy comparison chart was created. This bar chart highlights how Support Vector Machine (SVM), Naïve Bayes (NB), and Random Forest (RF) respond to different preprocessing strategies—namely, no feature selection, Pearson Correlation only, PCA only, and the hybrid combination of both. The visualization serves to summarize the experimental findings in a more interpretable format.

From the chart in Fig. 10, it is evident that all classifiers perform best when using either the full feature set or the hybrid PC And PCA approach. Random Forest consistently shows strong accuracy across all configurations, slightly outperforming others in most cases. SVM maintains stable performance regardless of preprocessing, indicating its robustness to different feature conditions. Naïve Bayes, however, shows a sharp decline when PCA is used in isolation, highlighting its sensitivity to changes in feature representation, but recovers when combined with Pearson Correlation.

These comparative results underscore the strength of the hybrid approach in maintaining high accuracy while reducing feature dimensionality. By leveraging Pearson Correlation to retain relevant features and PCA to eliminate redundancy, the hybrid method ensures a balanced trade-off between performance and computational efficiency. This makes it especially suitable for large-scale sentiment analysis tasks where both accuracy and speed are critical.

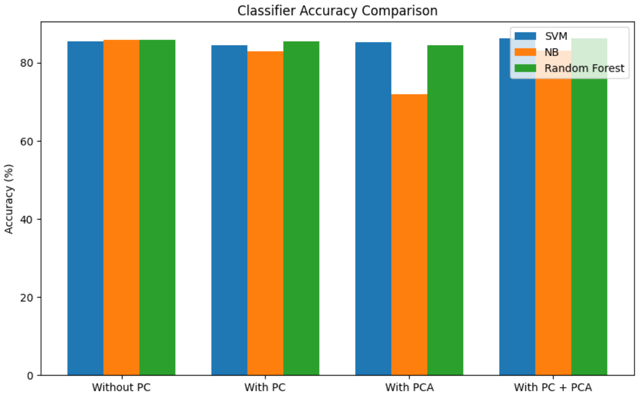


Figure 10. Classifier accuracy comparison

To further evaluate the classification performance beyond accuracy, the Receiver Operating Characteristic (ROC) curve analysis was performed. The ROC curve, as shown in Fig. 11, illustrates the trade-off between true positive rate and false positive rate across various threshold settings for each classifier. The Area Under the Curve (AUC) serves as a quantitative measure of a model's ability to distinguish between sentiment classes in a multiclass scenario.

As depicted in Figure 11, Random Forest achieved the highest AUC score at 0.85, indicating strong discriminative capability across the sentiment classes. SVM followed closely with an AUC of 0.81, reflecting consistent performance aligned with its high accuracy in earlier evaluations. In contrast, Naïve Bayes showed a significantly lower AUC of 0.66, confirming its relative weakness in complex feature spaces and multiclass contexts, especially when PCA was applied in isolation.

These findings reinforce the earlier results, while all classifiers have potential in sentiment classification, Random Forest demonstrates superior robustness and generalization. The ROC curve also highlights the importance of proper feature selection and model pairing since the same dataset and preprocessing strategy yield varying outcomes depending on the classifier used. Thus, AUC serves as a critical metric for validating model effectiveness, particularly when class balance and interpretability are key considerations.

To provide a consolidated view of model performance across all feature selection strategies, a heatmap was generated to display the accuracy of each classifier under four different preprocessing conditions: no feature selection, Pearson Correlation (PC) only, PCA only, and the combined PC and PCA approach. Figure 12 uses a color scale from lighter tones for lower accuracy to darker shades for higher accuracy, to facilitate quick comparison of how each method impacts SVM, Naïve Bayes, and Random Forest.

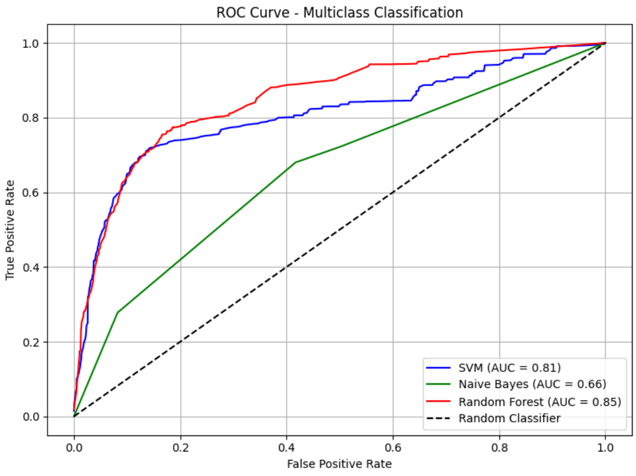


Figure 11. ROC curve

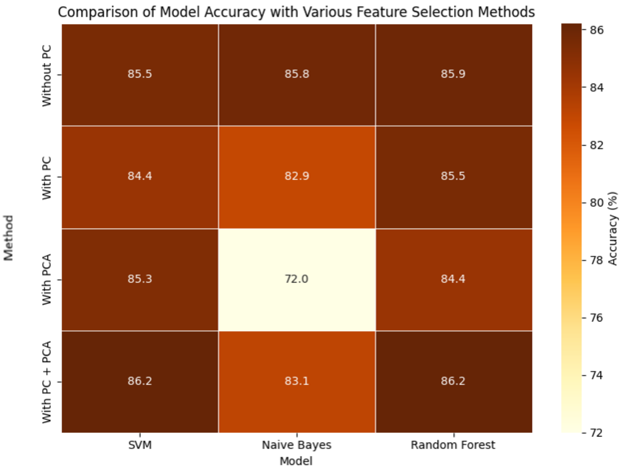


Figure 12. Model accuracy comparison heatmap

The heatmap clearly illustrates that the hybrid PC and PCA method (bottom row) yields the highest accuracy for both SVM (86.2%) and Random Forest (86.2%), outperforming all other configurations. Naïve Bayes also benefits from the hybrid approach, recovering to 83.1% after its dramatic decline under PCA alone. In contrast, PCA by itself (third row) significantly degrades Naïve Bayes performance (72.0%), although SVM (85.3%) and Random Forest (84.4%) remain relatively robust.

Overall, this visualization confirms that combining Pearson Correlation with PCA strikes the best balance between feature relevance and dimensionality reduction. While raw TF-IDF or PC only filtering produces solid baseline results, and PCA alone risks eliminating critical information for certain classifiers, the hybrid pipeline consistently maximizes accuracy across all three models.

4 CONCLUSION

This study demonstrates that the hybrid approach combining Pearson Correlation (PC) and Principal Component Analysis (PCA) is effective in improving sentiment classification accuracy for TikTok Seller users.

The hybrid PC and PCA approach achieved the highest accuracy, reaching 86.2% with both SVM and Random Forest classifiers. This shows that combining feature relevance (PC) with dimensional compactness (PCA) results in a more effective feature set than either method alone.

Compared to using no feature selection or using PC/PCA individually, the hybrid method consistently improved model performance. For instance, Naive Bayes performance dropped significantly when PCA was applied alone (from 85.8% to 72.0%), but it recovered with the hybrid method (83.1%), showing the balancing effect of PC in selecting meaningful features.

Dimensionality reduction using PCA helped reduce feature space while maintaining high variance (95%), improving computational efficiency. Although PCA alone reduced performance in some cases (especially for NB), its combination with PC mitigated this effect and even enhanced the robustness of models like SVM and RF.

Overall, the hybrid method proves to be a practical and accurate solution for feature selection in opinion mining tasks involving high-dimensional textual data. It contributes not only to better model accuracy but also to more efficient processing, essential for large-scale social media analytics.

While the proposed hybrid approach of Pearson Correlation and PCA has shown improved accuracy and efficiency in sentiment classification, this study acknowledges a key limitation: the model's inability to effectively detect non-literal sentiment expressions such as sarcasm, irony, or slang. These expressions often carry implicit meanings that differ from their literal wording, posing challenges for traditional feature selection and machine learning techniques that rely primarily on lexical and statistical patterns. For example, sarcastic reviews may use positive words to convey negative sentiments, which can mislead classifiers.

This limitation stems from the reliance on surface-level features derived from TF-IDF and linear correlations, which do not capture deeper semantic or contextual nuances. Future research should explore integrating contextual embeddings (e.g., BERT, ELMo) or sentiment-aware pre-trained language models that can better understand non-literal sentiment. Additionally, incorporating linguistic cues or pragmatic markers specific to sarcasm and local slang in Indonesian language contexts may enhance the system’s ability to detect more complex emotional expressions in user reviews.

For future improvement, the current method could be extended using nonlinear dimensionality reduction techniques such as t-SNE or UMAP to capture complex structures in the data. Another promising direction is to incorporate deep learning-based feature selection, such as attention mechanisms or transformer-based embeddings, to better represent semantic nuance in informal language.

CREDIT AUTHOR STATEMENT

Nova Tri Romadloni, Conceptualization, methodology design, supervision of research activities, writing – original draft preparation, and final manuscript editing. Muhammad Yusuf Ariyadi, Data collection through web scraping, data labeling, and preparation of the TikTok Seller review dataset. Wakhid Kurniawan, Data preprocessing including tokenization, stopword removal, stemming, and lemmatization; implementation of TF-IDF and Pearson Correlation. Burhan Efendi, conducted documentation of each research stage, assisted in experiment execution and result visualization, and contributed to writing and formatting the manuscript. All authors have reviewed and approved the final version of the manuscript.

COMPETING INTERESTS

The authors declare that there are no competing interests or conflicts of interest related to the publication of this paper.

This research was conducted independently and received no influence from external commercial or financial relationships that could be construed as a potential conflict.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

The authors declare that generative AI and AI-assisted technologies were used in the preparation of this manuscript solely for improving language clarity, grammar, and formatting. All intellectual content, analysis, interpretations, and conclusions are the original work and responsibility of the authors.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Majelis Diktilitbang PP Muhammadiyah for funding this research through the Riset Muhammadiyah Batch VIII Program. Their support has been instrumental in enabling the successful completion of this study. The funding provided not only facilitated the technical execution of the research activities but also supported the dissemination of findings through scientific publications. The authors deeply appreciate the commitment of Majelis Diktilitbang in advancing research and innovation within the Muhammadiyah higher education community.

REFERENCES

[1] D. Ardiansyah, A. Saepudin, R. Aryanti, E. Fitriani, and Royadi, “Analisis Sentimen Review Pada Aplikasi Media Sosial Tiktok Menggunakan Algoritma K-Nn Dan Svm Berbasis Pso,” Jurnal Informatika Kaputama (JIK), vol. 7, no. 2, pp. 233–241, 2023, doi: 10.59697/jik.v7i2.148.

[2] P. Verma, A. Dumka, A. Bhardwaj, and A. Ashok, “Product Review-Based Customer Sentiment Analysis Using an Ensemble of mRMR and Forest Optimization Algorithm (FOA),” International Journal of Applied Metaheuristic Computing, vol. 13, no. 1, pp. 1–21, 2022, doi: 10.4018/ijamc.2022010107.

[3] M. R. Kurniawan, D. Erawati, H. Setiawan, and Harmain, “Digitalisasi: Strategi Komunikasi KPU Dalam Meningkatkan Partisipasi Gen Z Pada Pemilu 2024,” INNOVATIVE: Journal Of Social Science Research, vol. 3, no. 6, pp. 1375–1390, 2023.

[4] Y. Deta Kirana and S. Al Faraby, “Sentiment Analysis of Beauty Product Reviews Using the K-Nearest Neighbor (KNN) and TF-IDF Methods with Chi-Square Feature Selection,” Open Access J Data Sci Appl, vol. 4, no. 1, pp. 31–042, 2021, doi: 10.34818/JDSA.2021.4.71.

[5] M. A. Athallah and K. Kraugusteeliana, “Analisis Kualitas Website Telkomsel Menggunakan Metode Webqual 4.0 dan Importance Performance Analysis,” CogITo Smart Journal, vol. 8, no. 1, pp. 171–182, 2022, doi: 10.31154/cogito.v8i1.374.171-182.

[6] J. P. van der Harst and S. Angelopoulos, “Less is more: Engagement with the content of social media influencers,” J Bus Res, vol. 181, no. May, p. 114746, 2024, doi: 10.1016/j.jbusres.2024.114746.

[7] S. S. Abdulkhaliq and A. M. Darwesh, “Sentiment Analysis Using Hybrid Feature Selection Techniques,” UHD Journal of Science and Technology, vol. 4, no. 1, pp. 29–40, 2020, doi: 10.21928/uhdjst.v4n1y2020.pp29-40.

[8] S. Sarina and A. M. Tanniewa, “Implementasi Algoritma Support Vector Learning Terhadap Analisis Sentimen Penggunaan Aplikasi Tiktok Shop Seller Center,” Prosiding SISFOTEK, vol. 7, no. 1, pp. 165–170, 2023.

[9] N. T. Romadloni and W. Supriyanti, “Analisis Sentimen Penggunaan Teknologi Pada Pendidikan Anak Usia Dini,” Jurnal Ilmiah SINUS, vol. 21, no. 2, p. 101, 2023, doi: 10.30646/sinus.v21i2.759.

[10] I. Aida Sapitri and M. Fikry, “Pengklasifikasian Sentimen Ulasan Aplikasi Whatsapp Pada Google Play Store Menggunakan Support Vector Machine,” Jurnal TEKINKOM, vol. 6, no. 1, pp. 1–7, 2023, doi: 10.37600/tekinkom.v6i1.773.

[11] S. J and K. U, “Sentiment analysis of amazon user reviews using a hybrid approach,” Measurement: Sensors, vol. 27, no. May, p. 100790, 2023, doi: 10.1016/j.measen.2023.100790.

[12] J. R. Jim, M. A. R. Talukder, P. Malakar, M. M. Kabir, K. Nur, and M. F. Mridha, “Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review,” Natural Language Processing Journal, vol. 6, no. November 2023, p. 100059, 2024, doi: 10.1016/j.nlp.2024.100059.

[13] N. Rawindaran, A. Jayal, and E. Prakash, “Exploration of the Impact of Cybersecurity Awareness on Small and Medium Enterprises (SMEs) in Wales Using Intelligent Software to Combat Cybercrime,” Computers, vol. 11, no. 12, 2022, doi: 10.3390/computers11120174.

[14] D. Leni, A. Dwiharzandis, R. Sumiati, H. Haris, and S. Afriyani, “Seleksi Fitur Berdasarkan Korelasi Pearson dalam Pemodelan Efisiensi Energi Bangunan,” Teknika Sains: Jurnal Ilmu Teknik, vol. 8, no. 2, pp. 103–115, 2023, doi: 10.24967/teksis.v8i2.2525.

[15] N. T. Romadloni, “Uncovering Insights in Spotify User Reviews with Optimized Support Vector Machine (SVM),” vol. 14, no. 1, pp. 530–546, 2025, doi: 10.14421/ijid.2025.4903.

[16] M. R, “Natural Language Processing For analysing and Extracting Insights,” Interantional Journal of Scientific Research in Engineering and Management, vol. 06, no. 06, pp. 1–4, 2022, doi: 10.55041/ijsrem14434.

[17] Sharazita Dyah Anggita and Ferian Fauzi Abdulloh, “Optimasi Algoritma Support Vector Machine Berbasis PSO Dan Seleksi Fitur Information Gain Pada Analisis Sentimen,” Journal of Applied Computer Science and Technology, vol. 4, no. 1, pp. 52–57, 2023, doi: 10.52158/jacost.v4i1.524.

[18] U. Nandagopal and S. Thirumalaivelu, “Classification of Malware with MIST and N-Gram Features Using Machine Learning,” International Journal of Intelligent Engineering and Systems, vol. 14, no. 2, pp. 323–333, 2021, doi: 10.22266/ijies2021.0430.29.

[19] M. M. Dewi, “Optimasi Pearson Correlation untuk Sistem Rekomendasi menggunakan Algoritma Firefly,” Jurnal Informatika, vol. 9, no. 1, pp. 1–5, 2022, doi: 10.31294/inf.v9i1.10209.

[20] Y. Gong, B. Liao, P. Wang, and Q. Zou, “DrugHybrid_BS: Using Hybrid Feature Combined With Bagging-SVM to Predict Potentially Druggable Proteins,” Front Pharmacol, vol. 12, no. November, pp. 1–12, 2021, doi: 10.3389/fphar.2021.771808.

[21] A. Dharmawan, R. E. Masithoh, and H. Z. Amanah, “Development of PCA-MLP Model Based on Visible and Shortwave Near Infrared Spectroscopy for Authenticating Arabica Coffee Origins,” Foods, vol. 12, no. 11, 2023, doi: 10.3390/foods12112112.

[22] S. S. and I. B. Budiyanto, “Analysis of Vocational School Development Based on Regional Potential Using Principal Component Analysis (PCA),” Innovation of Vocational Technology Education, vol. 16, no. 1, pp. 76–103, 2020, doi: 10.17509/invotec.v16i1.23515.

[23] N. Hafidz and D. Yanti Liliana, “Klasifikasi Sentimen pada Twitter Terhadap WHO Terkait Covid-19 Menggunakan SVM, N-Gram, PSO,” Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 5, no. 2, pp. 213–219, 2021, doi: 10.29207/resti.v5i2.2960.

[24] M. V. Naik, D. Vasumathi, and A. P. S. Kumar, “An Improved Intelligent Approach to Enhance the Sentiment Classifier for Knowledge Discovery Using Machine Learning,” 2020. doi: http://dx.doi.org/10.2174/2210327910999200528114552.

[25] É. T. Morais, G. A. Barberes, I. V. A. F. Souza, F. G. Leal, J. V. P. Guzzo, and A. L. D. Spigolon, “Pearson Correlation Coefficient

Applied to Petroleum System Characterization: The Case Study of Potiguar and Reconcavo Basins, Brazil," 2023. doi: 10.3390/geosciences13090282.

[26] X. Cheng, "A Comprehensive Study of Feature Selection Techniques in Machine Learning Models," Insights in Computer, Signals and Systems, vol. 1, no. 1, pp. 65–78, 2024, doi: 10.70088/xpf2b276.

[27] J. Li, Y. Huang, Y. Lu, L. Wang, Y. Ren, and R. Chen, "Sentiment Analysis Using E-Commerce Review Keyword-Generated Image with a HybridMachine Learning-BasedModel," Computers, Materials and Continua, vol. 80, no. 1, pp. 1581–1599, 2024, doi: 10.32604/cmc.2024.052666.

[28] A. Razzaque and D. A. Badholia, "PCA based feature extraction and MPSO based feature selection for gene expression microarray medical data classification," Measurement: Sensors, vol. 31, no. May 2023, p. 100945, 2024, doi: 10.1016/j.measen.2023.100945.

[29] N. T. Romadloni, N. D. Septiyanti, C. H. Pratomo, W. Kurniawan, and R. A. K. N. Bintang, "Classification of Sms Spam With N-Gram and Pearson Correlation Based Using Machine Learning Techniques," SENTRI: Jurnal Riset Ilmiah, vol. 3, no. 2, pp. 967–977, 2024, doi: 10.55681/sentri.v3i2.2252.

[30] D. Pajri, Y. Umaidah, and T. N. Padilah, "K-Nearest Neighbor Berbasis Particle Swarm Optimization untuk Analisis Sentimen Terhadap Tokopedia," Jurnal Teknik Informatika dan Sistem Informasi, vol. 6, no. 2, pp. 242–253, 2020, doi: 10.28932/jutisi.v6i2.2658.

[31] I. M. Nasir et al., "Pearson correlation-based feature selection for document classification using balanced training," Sensors (Switzerland), vol. 20, no. 23, pp. 1–18, 2020, doi: 10.3390/s20236793.

[32] D. Risqiwati, A. D. Wibawa, E. S. Pane, W. R. Islamiyah, A. E. Tyas, and M. H. Purnomo, "Feature Selection for EEG-Based Fatigue Analysis Using Pearson Correlation," Proceedings - 2020 International Seminar on Intelligent Technology and Its Application: Humanification of Reliable Intelligent Systems, ISITIA 2020, pp. 164–169, 2020, doi: 10.1109/ISITIA49792.2020.9163760.

[33] O. F. Nzeakor, B. N. Nwokeoma, I. Hassan, B. O. Ajah, and J. T. Okpa, "Emerging Trends in Cybercrime Awareness in Nigeria," International Journal of Cybersecurity Intelligence & Cybercrime, vol. 5, no. 3, pp. 41–67, 2022, doi: 10.52306/2578-3289.1098.

[34] P. Chen, F. Li, and C. Wu, "Research on Intrusion Detection Method Based on Pearson Correlation Coefficient Feature Selection Algorithm," J Phys Conf Ser, vol. 1757, no. 1, 2021, doi: 10.1088/1742-6596/1757/1/012054.