VOL. 11. NO. 3 FEBRUARY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v11i3.7282

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

SUPPORT VECTOR MACHINE TO CLASSIFY SENTIMENT REVIEWS ON GOOGLE PLAY STORE

Agus Nursikuwagus[1*]; Suherman[2]; Heri Purwanto[3]; Tono Hartono[4]

Informatics Management[1]
Universitas Komputer Indonesia, Bandung, Indonesia[1]
http://www.unikom.ac.id[1]
agusnursikuwagus@email.unikom.ac.id*

Informatics[2]
Universitas Serang Raya, Serang, Indonesia[2]
http://unsera.ac.id[2]
suherman.unsera@gmail.com

Information System[3]
Universitas Sanggabuana, Bandung, Indonesia[3]
http://sanggabuana.ac.id[3]
heri.purwanto@usbypkp.ac.id

Information System[4]
Universitas Komputer Indonesia, Bandung, Indonesia[4]
http://www.unikom.ac.id[4]
tono.hartono@email.unikom.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *This research addresses the "rating-content discrepancy" on the Google Play Store, where numerical star ratings often conflict with the actual sentiment of textual reviews. Utilizing the CRISP-DM framework, the study evaluates the effectiveness of machine learning in resolving these inconsistencies by classifying Instagram user reviews into positive and negative categories. Two primary algorithms were compared using a dataset of 600 reviews. The Support Vector Machine (SVM) model demonstrated high efficacy with an accuracy of 0.84. In contrast, the K-Nearest Neighbors (KNN) model performed poorly, achieving an accuracy of only 0.48. This significant performance gap highlights SVM's superior ability to handle high-dimensional text data through optimal hyperplane separation. The research further integrated the Streamlit library to create an interactive web interface for real-time sentiment prediction and result visualization. Ultimately, this study confirms that a structured CRISP-DM approach combined with SVM provides a robust solution for automated opinion mining, offering a reliable methodology for future data science applications in social media analysis.*

Keywords: *CRISP-DM, Google Play Store, KNN, Sentiment, SVM.*

Intisari—*Penelitian ini membahas "perbedaan konten peringkat" di Google Play Store, di mana peringkat bintang numerik sering bertentangan dengan sentimen aktual dari ulasan tekstual. Memanfaatkan kerangka kerja CRISP-DM, penelitian ini mengevaluasi efektivitas pembelajaran mesin dalam mengatasi inkonsistensi ini dengan mengklasifikasikan ulasan pengguna Instagram ke dalam kategori positif dan negatif. Dua algoritma utama dibandingkan menggunakan kumpulan data 600 ulasan. Model Support Vector Machine (SVM) menunjukkan kemanjuran tinggi dengan akurasi 0,84. Sebaliknya, model K-Nearest Neighbors (KNN) berkinerja buruk, mencapai akurasi hanya 0,48. Kesenjangan performa yang signifikan ini menyoroti*

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.7282

kemampuan unggul SVM untuk menangani data teks berdimensi tinggi melalui pemisahan hyperplane yang optimal. Penelitian ini lebih lanjut mengintegrasikan perpustakaan Streamlit untuk membuat antarmuka web interaktif untuk prediksi sentimen real-time dan visualisasi hasil. Pada akhirnya, penelitian ini menegaskan bahwa pendekatan CRISP-DM terstruktur yang dikombinasikan dengan SVM memberikan solusi yang kuat untuk penambangan opini otomatis, menawarkan metodologi yang andal untuk aplikasi ilmu data di masa depan dalam analisis media sosial.

***Kata Kunci**: CRISP-DM, Google Play Store, KNN, Sentimen, SVM.*

INTRODUCTION

The Google Play Store serves as a platform for software and application distribution, featuring rating and review options for users. This facility allows application users to express their satisfaction through reviews of the apps they have downloaded and used. On the Google Play Store, users can leave comments in the form of star ratings ranging from one to five, along with text reviews [1]. Apps with high star ratings are recommended by Google and can be featured on the front page or listed among the best when sorted by rating. However, some reviews on the Google Play Store may display low ratings, such as one or two stars, despite the accompanying review text being positive. Conversely, there are cases where the ratings are high, but the review text is critical [2]. In this case, Google cannot determine the difference between positive and negative reviews solely based on the review text provided by users, as the written reviews sometimes do not align with the ratings given.

Google Play Store is a digital content service, where you can install applications and other online products for free or for a fee. The Google Play Store was developed and launched by Google on March 6, 2012 [3]. The Google Play Store offers a diverse range of apps developed for various needs, including games, movies, and music applications. Instagram is a social media application available on the Google Play Store that enables users to connect and interact with one another. Users can engage in different activities using the app's features, such as uploading photos and videos and broadcasting live. According to a survey conducted by NapoleonCat in Warsaw, Poland, the number of Instagram users reached 61,610,000 in 2019 and has continued to grow each year.

To address this issue, a classification system was developed to categorize opinions from reviews based on customer satisfaction, identifying whether the sentiments are negative or positive. The approach utilized the CRISP-DM research method, employing both the support vector machine (SVM) algorithm and the K-nearest neighbor (KNN) algorithm [4], [5]. Cross-Industry Standard Process for Data Mining (CRISP-DM) is a method that provides a standard data mining process as a general problem-solving process of a business or research unit [6], [7].

The sources outline the standard pre-processing framework often used to convert raw text into numerical data through the TF-IDF [4] method. These steps include: Case Folding & Cleaning: Standardizing the font format and removing non-text characters. Tokenization: Breaking sentences into units of words or tokens. Stemming: Converting words into basic forms (e.g., using the IterativeLovinsStemmer algorithm). Stopword Removal: Removing common words with low information values such as "and" or "the" .

A major drawback in traditional text classification studies is the limitations in dealing with contextual nuances and complex linguistic phenomena: Traditional NLP Weaknesses: Lexicon-based methods (such as VADER) or classical classification often fail to detect sarcasm, contextual dependence, and domain-specific jargon. Large Language Models (LLM) solutions: Models such as GPT-4 offer a deeper understanding of context through structured prompting strategies (such as few-shot or chain-of-thought), which allow for more accurate extraction of aspects and sentiments than rule-based methods or simple machine learning [4]. Gender Linguistic Patterns: There has been discussion of pronoun use and writing complexity (such as the more frequent use of the singular first-person pronoun "I/me" in women), but more research is needed to understand this linguistic pattern as a basis for stronger sentiment analysis [1].

To ensure systematic data analysis, some sources refer to operational frameworks such as: CRISP-DM: Consists of six main phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. KDD (Knowledge Discovery in Databases): Used to identify valid and useful patterns from a set of review data through the stages of selection, transformation, and pattern visualization. Analogy: Understanding the classification of text without a strong semantic orientation is like trying to understand a song just

by counting the number of words in it. Even though we know how many words are used (technical data/accuracy), we still lose the emotional meaning and melody that makes the words have a feeling (semantic orientation).

The objective of this research is to classify reviews of the Instagram application on the Google Play Store using the CRISP-DM methodology, along with SVM and KNN algorithms [4], [8]. This classification focuses on determining whether the reviews are negative or positive in nature. The contributions of this paper are outlined below, providing clarity on the issue at hand:

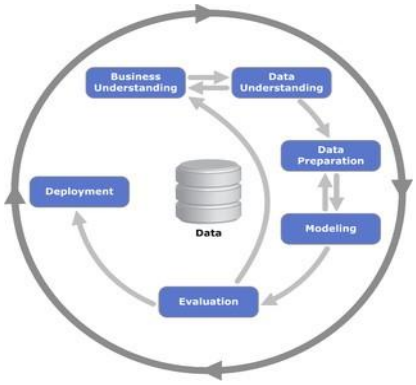
The linear model can distribute data points into defined classes. SVM offers various kernels for data separation [9]. One of these kernels is linear, allowing us to set SVM parameters according to specific objectives. High-dimensional features can be utilized to classify instances labeled by SVM. SVM can effectively enhance the boundaries using the specified labeled classes [4].

The classification using KNN can lead to diversification and dissemination of points by calculating the nearest data point that shares similar characteristics. We need to determine the value of K, which defines the number of neighbors considered. A common choice is K = 5, as this can help improve the quality of the results. By analyzing the distances among points, we can effectively group those that share the same characteristics [10], [11].

Following the paper scheme, we organize the sections according to the argumentation and template flow. For example, section one addresses the urgencies, background problems, problem definition, and contribution. Section two contains the methodology and outlines the approach taken in the paper. Section three discusses the results and their implications, explaining why they appear the way they do. The closing section is a conclusion that summarizes the paper's aims.

MATERIALS AND METHODS

This section writes a path of research. We used the famous methods like CRISP-DM [7], [12]. The reason for using CRISP-DM is that the process is easy to follow and understand. We can directly use the stages of this method and apply them to our work. CRISP-DM is a methodology created by IBM. The Cross-Industry Standard Process for Data Mining, or CRISP-DM, is a data mining process model established in 1996 by the European Commission, with the aim of becoming a standard in data mining for use in the industrial sector. CRISP-DM follows an algorithmic working principle that identifies the most optimal separation space derived from different classes. It consists of six phases: the business understanding phase, the data understanding phase, the data preparation phase, the modeling phase, the evaluation phase, and the dissemination phase [6], [7]. The research method utilized in this study is the Cross-Industry Standard Process for Data Mining (CRISP-DM). In this framework, a data mining project follows a life cycle comprised of six phases. These phases are both sequential and adaptive, meaning that each subsequent phase relies on the outcomes of the preceding one. Arrows depict the critical relationships between the phases [6]. The six phases of this CRISP-DM method are shown in Figure 1.



Source : (R. Wirth and J. Hipp, 2025)
Figure 1. CRISP-DM

Business understanding is the projection of project objectives and also needs in complete within the scope of the business or research unit as a whole is determined [6]. After identifying the problem, sentiment analysis will be conducted on user reviews of the Instagram application.

Data Understanding is a stage that understands the structure of data that helps to solve research problems [13], [14]. In this phase, initial data collection was conducted. Next, an investigative analysis of the data was performed to further understand the information and gather preliminary insights. Following this, the quality of the data was evaluated, and if necessary, a small subset of data exhibiting potential problem patterns was selected.

Data Preparation, the goal of this phase is to lay the foundation for the subsequent phase. Conduct this phase thoroughly by selecting the case and the variables to analyze, ensuring they align with the intended analysis. Adjustments should be made to several variables, and the initial data must be prepared for modeling. Data preprocessing includes the steps of case folding, tokenization,

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.7282

stemming, slang word removal, filtering or stop word removal, encoding target values, and text vectorization (TF-IDF) [15], [16], Dealing with imbalance (SMOTE) [15][[16], [17], [18], [19], [20], [21]Split Data [22], and Word Cloud [23], [24].

Modelling phase, the focus is on selecting and applying modelling techniques that align with the established objectives. This stage aims to optimize the results based on the desired outcomes. If needed, the process may revert to the previous phase, specifically the data processing phase, to transform the data into a format that meets the specifications required by the predetermined data mining techniques [12], [14].

Metric is the evaluation of model data performance is conducted, and all processes that have been implemented are reviewed. This allows us to identify true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) [3], [25][26].

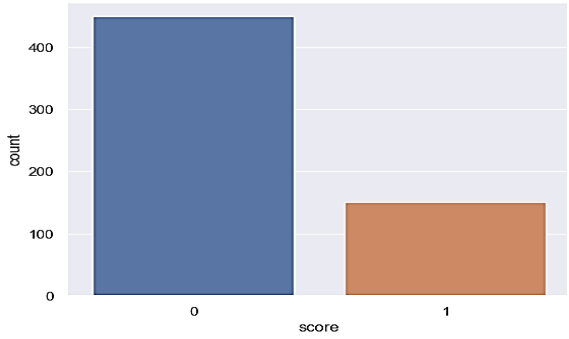
The model's completion does not indicate the project's completion. In this phase, we deploy the model and use Streamlit to access existing data through web application interaction. Streamlit is a web framework designed to facilitate the simple deployment of models or visualizations using the Python programming language, which is both efficient and visually appealing [27].

RESULTS AND DISCUSSION

In the business understanding phase, this study aims to classify reviews of Instagram applications sourced from the Google Play Store. The researcher plans to employ the SVM algorithm to categorize the sentiment of these reviews as either negative or positive [25], [26].

Data Understanding Phase, this phase begins with the collection of data such as stemming, slang, punctuation, which is then combined into a single dataset for analysis. In this study, the data collection phase focuses on reviewing the Instagram application. Web scraping is a technique employed to extract information (text) from a website, facilitating further analysis of the resulting dataset. The review dataset was sourced from the Instagram application on the Google Play Store using the web scraping technique provided by google-play-scraper. Subsequently, this dataset was converted into a CSV file containing 600 reviews utilized in this study. The CSV file comprises the fields userName, score, at, and content. The userName column contains usernames from the Google Play Store; the score column reflects the ratings given by users for the Instagram application, with values ranging from 0 to 1; the at column indicates the timestamp of when the review was submitted; and the content column includes the reviews or comments provided by users regarding the Instagram application [3], [23], [28][29].

Class distribution refers to the method of dividing the dataset into classes based on the score column. Initially, the data is represented by values ranging from 1 to 5, which are then transformed into binary values of 0 and 1. A value of 0 is assigned to data that combines scores of 1 and 2, while a value of 1 is assigned to data that combines scores of 3, 4, and 5. In total, there are 449 instances with a score value of 0 and 151 instances with a score value of 1. The results of the class distribution are illustrated in Figure 2.



Source: (Research Results, 2025)
Figure 2. Class Distribution

Case folding is a process that standardizes letters by converting all characters from "a" to "z" into lowercase form. This uniformity helps prevent differences in meaning that may arise from the presence of uppercase and lowercase letters in the dataset. Table 1 illustrates the result of case folding.

Table 1. Case Folding

No	Score	Content	Content Length	Case_folded
1	0	Sumpah filter gw yang awalnya dibagian cerita ...	237	sumpah filter gw yang awalnya dibagian cerita ...
2	0	Awal main instagram cuma buat nyimpen data (fo...	281	awal main instagram cuma buat nyimpen data fot...
3	0	G bisa login ke akun awal, muncul kesalahan trs...	138	g bisa login ke akun awal muncul kesalahan trs...
4	0	Reel udh di update,	210	reel udh di update

No	Score	Content	Content Length	Case_folded
		besokannya lagu tidak terse...		besokannya lagu tidak terse...
5	0	Kecewa si... pas di update malah jadi gak bisa lo...	146	kecewa si pas di update malah jadi gak bisa lo...

Source: (Research Results, 2025)

Tokenizing is a stage that involves breaking down each word in a sentence derived from the input text. This process results in a series of individual words, as illustrated in Table 2.

Table 2. Tokenizing

Tokenized
[sumpah, filter, gw, yang, awalnya, dibagian, cerita, ...]
[awal, main, instagram, cuma, buat, nyimpen, data, ...]
[g, bisa, login, ke, akun, awal, muncul, kesalahan, ...]
[reel, udh, di, update, besokannya, lagu, tidak, terse, ...]
[kecewa, si, pas, di, update, malah, jadi, gak, ...]

Source: (Research Results, 2025)

Stemming is a process used to eliminate certain words in the review data that have suffixes, such as the word "reduced," which is returned to its base form, "less." At this stage, the Python library is utilized because it is specifically designed for the Indonesian language. The results of the stemming process can be observed in Table 3.

Table 3. Stemming

Stemmed
[sumpah, filter, gw, yang, awal, bagi, cerita, ...]
[awal, main, instagram, cuma, buat, nyimpen, data, ...]
[g, bisa, login, ke, akun, awal, muncul, salah, ...]
[reel, udh, di, update, besok, lagu, tidak, sedia, ...]
[kecewa, si, pas, di, update, malah, jadi, gak, ...]

Source: (Research Results, 2025)

Slang word removal involves altering slang words or terms in the dataset by utilizing the Indonesian language. The outcomes of the slang words removal process are illustrated in Table 4.

Table 4. Slank Word Removal

No_slang
[sumpah, filter, gw, yang, awal, bagi, cerita, ...]
[awal, main, instagram, cuma, buat, nyimpen, data, ...]
[g, bisa, login, ke, akun, awal, muncul, salah, ...]
[reel, udh, di, update, besok, lagu, tidak, sedia, ...]
[kecewa, si, pas, di, update, malah, jadi, baik, jadi, ...]

Source: (Research Results, 2025)

Stop words removal is a process aimed at eliminating certain words from documents or datasets, thereby reducing the overall number of words in the corpus. The outcomes of the stop words removal process are illustrated in Table 5.

Table 5. Stopword Removal

No_stop
[sumpah, filter, awal, bagi, cerita, pindah, reel, foto]
[awal, main, instagram, nyimpen, data, video]
[login, akun, muncul, salah, pdh, username]
[reel, udh, update, besok, lagu, tidak, sedia, coba]
[kecewa, pas, update, gak, bisa]

Source: (Research Results, 2025)

Thereafter, change the target in the score column that originally contained 0 and 1 to "negative" and "positive.". TF-IDF is used to determine the frequency value of a word within a dataset or document. We see the class label is not the same amount, so we propose the SMOTE process to employ the address data imbalances. After applying SMOTE, the dataset increased to 896 instances from the original 600. The SMOTE process is conducted based on the results of the TF-IDF analysis [15][16], [30], [31].

Figure 3 shows the result of the word cloud process for the positive review class, while Figure 4 displays the result for the negative review class [23].

Source: (Research Results, 2025)

Figure 3. Word cloud Sentiment Positive

Based on Figure 3 and Figure 4, which present a collection of review words with negative connotations often used by users to comment on the Instagram application, examples include difficulties logging in, the need for updates, and frequent errors or bugs in the application.

The results of the split data process show that when the data is divided into 20% for testing and 80% for training, the testing results yielded a value of 718, while the training value was 180.

Source: (Research Results, 2025)

Figure 4. Word cloud Sentiment Negative

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.7282

The models utilized are Support Vector Machine (SVM) and K-Nearest Neighbor (KNN). SVM, or Support Vector Machine, is a supervised algorithm used for classification tasks. The use of SVM enables the resolution of both linear and non-linear classification problems due to its advanced mathematical concepts. Table 6 illustrates the accuracy achieved with the Support Vector Machine (SVM) model. With a data distribution of 80% for training and 20% for testing, the model yields an accuracy of 84%, a precision (prec) of 87%, a recall (rec) of 86%, and an F1 score (F1) of 85% [11], [23], [28].

Table 6. Table Accuracy of SVM

Item	Prec.	Rec.	F1	support
Negative	0.87	0.83	0.85	95
Positive	0.82	0.86	0.84	85
Accuracy	-	-	0.84	180
Macro.Avg	0.84	0.85	0.84	180
Weighted avg	0.84	0.84	0.84	180

Source: (Research Results, 2025)

Table 7 presents the accuracy results obtained using the KNN (K-nearest neighbor) algorithm. With a data distribution of 80% for training and 20% for testing, the results show an accuracy of 48%, a precision of 100%, a recall of 100%, and an F1 score of 64%.

Table 7. Table Accuracy of KNN

Item	Prec.	Rec.	F1	support
Negative	0.87	0.83	0.85	95
Positive	0.82	0.86	0.84	85
Accuracy	-	-	0.84	180
Macro.Avg	0.84	0.85	0.84	180
Weighted avg	0.84	0.84	0.84	180

Source: (Research Results, 2025)

We present the analysis results using the confusion matrix for the SVM and K-NN models. In Table 8 and Table 9, the value displayed represents the accuracy of the confusion matrix derived from the SVM algorithm and KNN Algorithm.

Table 8. Confusion Matrix of SVM

True Predictive	Positive.	Negative.
Positive	73	16
Negative	12	79

Source: (Research Results, 2025)

Table 9. Confusion Matrix of KNN

True Predictive	Positive.	Negative.
Positive	85	94
Negative	0	1

Source: (Research Results, 2025)

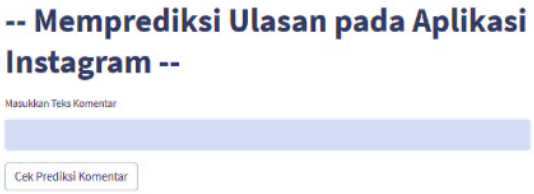
Additionally, this text will address the calculation of the accuracy of the SVM algorithm and KNN:

$$SVM\ Acc. = \frac{73 + 79}{73 + 16 + 12 + 79} = 0.84$$

$$KNN\ Acc. = \frac{85 + 1}{85 + 94 + 1 + 0} = 0.48$$

The SVM classifier is more effective at predicting sentiment than KNN. The sentence labels predicted to be true match the actual true labels for almost 152 sentences, while there are approximately 28 incorrect predictions, as shown in Table 8. This study demonstrates that separating label classes with a linear SVM kernel is more effective when dividing the data labels [4], [8]. Processing imbalanced classes and selecting the appropriate encoding are important steps in determining the value of a dataset. We must guarantee the order of preprocessing to achieve the best classifier and optimal accuracy [4], [5], [9], [10], [32].

In the deployment phase, the best model obtained from the evaluation results is stored in a joblib file (.joblib). This model is then utilized within a prediction framework for the review of the Instagram application on the Streamlit platform [27]. Figure 5 displays the Instagram application prediction in Streamlit [27].



Source: (Research Results, 2025)
Figure 5. Input Text

Figure 6 and Figure 7 show instances used for testing the sentences. We input the sentences via keyboard for this testing. In Figure 6, for example, we entered the sentence "saran ig video musik anggap kerna hak cipta mohon video tayang musiknya nonaktif video." After clicking the button labeled "Cek Prediksi Komentar," the model identifies the sentiment as "positive sentiment." Similarly, in Figure 7, the model predicts the outcome as a negative sentiment [27].

Source: (Research Results, 2025)
Figure 6. Result Reviews Comment in Positive

Source: (Research Results, 2025)

Figure 7. Result Reviews Comment in Negative

On the completion process, we reach the result that SVM is outperform than KNN. The accuracy of SVM is 0.84 and then KNN is 0.48. This event can be said the SVM with the linear kernel is able to good separately the class label within negative and positive comments. Another support is an imbalance method. Imbalance method using SMOTE has encourage the fairly destination on choosing the class label. The balance class will make the class not oriented in one class but focus on another class. This evidence is very importance to make balance in probability. In SVM, of course in parameter tuning is more help to reach a good separated class. Choosing of kernel in SVM is one more a good action to obtain a better result in accuracy. We can comparison between Table 3 and Table 4, the true positive (TP) and true negative (TN), are to be a part to destiny a good accuracy.

CONCLUSION

This study validates that the CRISP-DM framework provides a robust structure for resolving discrepancies between user ratings and review content. The results conclude that SVM significantly outperforms KNN in sentiment classification, particularly when paired with SMOTE for class balancing. However, after manual data labeling, it was found that the negative class significantly outnumbered the positive class, with 449 negative reviews compared to 151 positive reviews. The cleaned data will then undergo a split data phase, where it will be divided into 20% for testing and 80% for training. The comparison of these algorithms indicates that the Support Vector Machine (SVM) is more effective than K-nearest neighbor (KNN), as the accuracy of the SVM algorithm is higher at 84%, while KNN achieves only 48%. Methodologically, this research suggests that future sentiment analysis systems should prioritize hyperplane-based models or deep learning architectures (such as LSTM) to better handle the nuances of non-standard grammar and slang. Further research should explore hybrid models that combine SVM with parameter-tuning frameworks to increase accuracy beyond the 0.84 threshold.

REFERENCES

[1] E. Noei and K. Lyons, "A study of gender in user reviews on the Google Play Store," *Empir Softw Eng*, vol. 27, no. 2, p. 34, 2021, doi: 10.1007/s10664-021-10080-8.

[2] Z. Hadi, E. Utami, and D. Ariatmanto, "Detect Fake Reviews Using Random Forest and Support Vector Machine," *SinkrOn*, vol. 8, no. 2, pp. 623–630, Apr. 2023, doi: 10.33395/sinkron.v8i2.12090.

[3] R. Kurniawan, H. O. L. Wijaya, and R. P. Aprisusanti, "Sentiment Analysis of Google Play Store User Reviews on Digital Population Identity App Using K-Nearest Neighbors," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 13, no. 2, pp. 170–178, Jun. 2024, doi: 10.32736/sisfokom.v13i2.2071.

[4] L. W. Rizkallah, "Optimizing SVM hyperparameters for satellite imagery classification using metaheuristic and statistical techniques," *Int. J. Data Sci. Anal.*, vol. 20, no. 5, pp. 4945–4962, Oct. 2025, doi: 10.1007/s41060-025-00762-7.

[5] A. R. Isnain, J. Supriyanto, and M. P. Kharisma, "Implementation of K-Nearest Neighbor (K-NN) Algorithm For Public Sentiment Analysis of Online Learning," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 15, no. 2, p. 121, Apr. 2021, doi: 10.22146/ijccs.65176.

[6] R. Wirth and J. Hipp, "CRISP-DM: Towards a Standard Process Model for Data Mining," 2025. [Online]. Available: www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf

[7] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Comput Sci*, vol. 181, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

[8] G. Liu, L. Wang, D. Liu, L. Fei, and J. Yang, "Hyperspectral Image Classification Based on Non-Parallel Support Vector Machine," *Remote Sens. (Basel).*, vol. 14, no. 10, May 2022, doi: 10.3390/rs14102447.

[9] R. B. Gumilar and others, "Analisa Perbandingan Algoritma Support Vector Machine dan K-Nearest Neighbors Terhadap Ulasan Aplikasi Vidio," *Journal of Information System Research (JOSH)*, vol. 5, no. 4, pp. 1188–1195, 2024, doi: 10.47065/josh.v5i4.5640.

[10] A. Dafid *et al.*, "Optimizing K-Nearest Neighbors with Particle Swarm Optimization for Improved Classification Accuracy

Corresponding Author," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, vol. 11, no. 2, pp. 238–250, 2025, doi: 10.26555/jiteki.v11i2.30775.

[11] I. G. I. Sudipa, R. A. Azdy, I. Arfiani, N. M. Setiohardjo, and Sumiyatun, "Leveraging K-Nearest Neighbors for Enhanced Fruit Classification and Quality Assessment," *Indonesian Journal of Data and Science*, vol. 5, no. 1, pp. 30–36, Mar. 2024, doi: 10.56705/ijodas.v5i1.125.

[12] F. Martínez-Plumed and others, "CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories," *IEEE Trans Knowl Data Eng*, vol. 33, no. 8, pp. 3048–3061, 2021, doi: 10.1109/TKDE.2019.2962680.

[13] J. Brzozowska, J. Pizoń, G. Baytikenova, A. Gola, A. Zakimova, and K. Piotrowska, "DATA ENGINEERING IN CRISP-DM PROCESS PRODUCTION DATA – CASE STUDY," *Applied Computer Science*, vol. 19, no. 3, pp. 83–95, 2023, doi: 10.35784/acs-2023-26.

[14] M. A. Hasanah, S. Soim, and A. S. Handayani, "Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir," 2021. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[15] M. Liang and T. Niu, "Research on Text Classification Techniques Based on Improved TF-IDF Algorithm and LSTM Inputs," *Procedia Comput Sci*, vol. 208, pp. 460–470, 2022, doi: 10.1016/j.procs.2022.10.064.

[16] J. Zhou, Z. Ye, S. Zhang, Z. Geng, N. Han, and T. Yang, "Investigating response behavior through TF-IDF and Word2vec text analysis: A case study of PISA 2012 problem-solving process data," *Heliyon*, vol. 10, no. 16, p. e35945, 2024, doi: 10.1016/j.heliyon.2024.e35945.

[17] D. S. Turan and B. Ordin, "The incremental SMOTE: A new approach based on the incremental k-means algorithm for solving imbalanced data set problem," *Inf Sci (N Y)*, vol. 711, p. 122103, 2025, doi: 10.1016/j.ins.2025.122103.

[18] E. L. T. Tchokote and E. F. Tagne, "Effective multimodal hate speech detection on Facebook hate memes dataset using incremental PCA, SMOTE, and adversarial learning," *Machine Learning with Applications*, vol. 20, p. 100647, 2025, doi: 10.1016/j.mlwa.2025.100647.

[19] P. Sun, Z. Wang, L. Jia, and Z. Xu, "SMOTE-kTLNN: A hybrid re-sampling method based on SMOTE and a two-layer nearest neighbor classifier," *Expert Syst Appl*, vol. 238, p. 121848, 2024, doi: 10.1016/j.eswa.2023.121848.

[20] S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana, "Enhancing SMOTE for imbalanced data with abnormal minority instances," *Machine Learning with Applications*, vol. 18, p. 100597, 2024, doi: 10.1016/j.mlwa.2024.100597.

[21] Md. Shamshuzzoha and others, "A novel framework for seasonal affective disorder detection: Comprehensive machine learning analysis using multimodal social media data and SMOTE," *Acta Psychol (Amst)*, vol. 256, p. 105005, 2025, doi: 10.1016/j.actpsy.2025.105005.

[22] H. K. Chinmayi and others, "Monitoring legume nutrition with machine learning: The impact of splits in training and testing data," *Appl Soft Comput*, vol. 176, p. 113186, 2025, doi: 10.1016/j.asoc.2025.113186.

[23] L. Stanca, D.-C. Dabija, and V. Câmpian, "Qualitative analysis of customer behavior in the retail industry during the COVID-19 pandemic: A word-cloud and sentiment analysis approach," *Journal of Retailing and Consumer Services*, vol. 75, p. 103543, 2023, doi: 10.1016/j.jretconser.2023.103543.

[24] H. Ren, Y. Liu, G. Naren, and J. Lu, "The impact of multidirectional text typography on text readability in word clouds," *Displays*, vol. 83, p. 102724, 2024, doi: 10.1016/j.displa.2024.102724.

[25] G. Indrawan, H. Setiawan, and A. Gunadi, "Multi-class SVM Classification Comparison for Health Service Satisfaction Survey Data in Bahasa," *HighTech and Innovation Journal*, vol. 3, no. 4, pp. 425–442, Dec. 2022, doi: 10.28991/HIJ-2022-03-04-05.

[26] A. Widodo, B. Agus Herlambang, and R. Renaldy, "Optimizing Support Vector Machine (SVM) for Sentiment Analysis of Blu by BCA Reviews with Chi-Square," 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[27] R. Yadav and H. Raheman, "Machine learning-based estimation of agricultural tyre sinkage: A streamlit web application," *J Terramech*, vol. 119, p. 101055, 2025, doi: 10.1016/j.jterra.2025.101055.

[28] J. Li and others, "Sentiment Analysis Using E-Commerce Review Keyword-Generated Image with a Hybrid Machine Learning-Based Model," *Computers, Materials and Continua*,

vol. 80, no. 1, pp. 1581–1599, 2024, doi: 10.32604/cmc.2024.052666.

[29] T. Anderson, S. Sarkar, and R. Kelley, "Analyzing public sentiment on sustainability: A comprehensive review and application of sentiment analysis techniques," *Natural Language Processing Journal*, vol. 8, p. 100097, 2024, doi: 10.1016/j.nlp.2024.100097.

[30] Q. Wan, X. Xu, and J. Han, "A dimensionality reduction method for large-scale group decision-making using TF-IDF feature similarity and information loss entropy," *Appl Soft Comput*, vol. 150, p. 111039, 2024, doi: 10.1016/j.asoc.2023.111039.

[31] E. Delibaş, "Efficient TF-IDF method for alignment-free DNA sequence similarity analysis," *J Mol Graph Model*, vol. 137, p. 109011, 2025, doi: 10.1016/j.jmgm.2025.109011.

[32] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat, "Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications," *J Big Data*, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-024-00973-y.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri