JUINKTI (Jurnal Inovasi dan Karya Teknologi Informasi)
Volume 1 Number 1 Mei 2026: 31-40
DOI: xxxxxxx/jikteks.xxxxx.xx

Comparative Analysis of Supervised Learning Algorithms in Indonesian Film Genre Classification Based on the CRISP-DM Approach

Putri Diantara [1], Anzil Apriliani [2], Alia Citra Aprilia [3] Idi Hermansyah[4]
[1234,] Studi Sains Data, Universitas Lembah Dempo

Article Info

Article history:

Received Mei 15, 2026
Revised Mei 20, 2026
Accepted Mei 25, 2026
Published AMei 30, 2026

Keywords:

Classification
Supervised learning
CRISP-DM
Support Vector Machine
Naive Bayes

ABSTRACT

This research is motivated by the complexity of the classification of Indonesian film genres due to language variations, inconsistent synopsis structures, and similarities in characteristics between genres. Given the limited comparative studies on local data, this study proposes an analysis of various *supervised learning* algorithms. The goal is to evaluate and determine the most optimal model for accurately classifying film genres based on synopsis texts.

This contribution uses the CRISP-DM framework to compare Decision Tree, Naive Bayes, SVM, and KNN algorithms. Data in the form of synopsis of Indonesian films is processed through text cleaning and word-frequency-based feature extraction. His main contribution is to provide a systematic evaluation and a replicable framework for the development of Natural Language Processing (NLP) in the Indonesian film industry.

The results showed that SVM provided the highest accuracy, while Naive Bayes excelled in time efficiency. Although Decision Tree is easy to interpret and KNN is sensitive to parameters, the performance of both can be significantly improved through normalization techniques as well as feature selection. The confusion *matrix evaluation* noted misclassification in drama and comedy genres, but the results of cross-validation ensured that the model remained consistent and stable across various data divisions.

Overall, this study shows that the CRISPDM approach can be used well in the process of classifying Indonesian film genres. Based on the results obtained, the Support Vector Machine is recommended as the best algorithm for this case because it is able to provide the most optimal performance compared to other methods.

Corresponding Author:

Putri Diantara,
Sains Data, Fakultas Sains dan Teknologi, Universitas Lembah Dempo
Jalan. H.Effendi Sangkim, Airlaga, Pagar Alam Sumatera Selatan
putridiantara610@gmail.com

1. INTRODUCTION

The Indonesian film industry has experienced significant growth in recent years, leading to a substantial increase in digital film data, particularly synopsis texts distributed across streaming platforms, online databases, and cinema information systems. This rapid expansion has created new opportunities for the implementation of artificial intelligence and text mining techniques in the entertainment industry. However, manual genre classification remains a challenging task because film synopses often contain ambiguous language structures, overlapping semantic meanings, and variations in writing style that make it difficult to accurately determine genre categories.[1] Genres such as *Drama*, *Romance*, and *Comedy* frequently share similar

narrative patterns and vocabulary, resulting in inconsistencies in manual categorization. Consequently, the need for an automated and intelligent genre classification system has become increasingly important in supporting digital film management and recommendation systems.[2]

Previous studies have demonstrated that text classification techniques can be effectively applied to various natural language processing (NLP) tasks, including sentiment analysis, topic categorization, and document classification. Nevertheless, comparative studies focusing specifically on Indonesian-language film synopses remain relatively limited, particularly those evaluating multiple supervised learning algorithms within a single analytical framework.[3] Therefore, this study proposes a comparative classification approach using the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology to identify the most effective algorithm for Indonesian film genre classification. The study evaluates the performance of three commonly used machine learning algorithms: Decision Tree, Naive Bayes, and Support Vector Machine (SVM).[4]

The CRISP-DM framework was selected because it provides a structured and systematic process for conducting data mining research. The methodology consists of six main stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.[5] In the **Business Understanding** stage, the research objectives are clearly defined, namely to develop an automated system capable of classifying Indonesian film genres based on synopsis texts. This stage also identifies practical benefits, such as supporting recommendation systems, improving digital catalog organization, and enhancing user experience on film streaming platforms.

The **Data Understanding** stage involves collecting Indonesian film synopsis data from publicly available movie databases and online repositories. Exploratory analysis is then performed to examine dataset characteristics, including the distribution of genres, text length variations, and vocabulary diversity.[6] Statistical exploration reveals that certain genres dominate the dataset, while others contain fewer samples, creating potential class imbalance issues that may affect classification performance.

During the **Data Preparation** stage, extensive preprocessing procedures are applied to improve text quality before modeling. These processes include cleansing irrelevant symbols and punctuation, converting text into lowercase through case folding, removing stopwords, performing stemming to reduce words into their root forms, and tokenization to separate sentences into individual terms.[7] Additionally, the Term Frequency–Inverse Document Frequency (TF-IDF) technique is utilized to transform textual data into numerical vector representations suitable for machine learning algorithms. TF-IDF helps identify important terms within synopses while reducing the influence of commonly occurring but less informative words.[8]

The **Modeling** stage compares the performance of Decision Tree, Naive Bayes, and Support Vector Machine algorithms. Decision Tree is selected due to its interpretability and ease of visualization, allowing researchers to understand decision-making patterns within the classification process.[9] Naive Bayes is included because of its computational efficiency and effectiveness in handling text classification tasks with probabilistic approaches.[10] Meanwhile, Support Vector Machine (SVM) is employed because of its strong capability to process high-dimensional text data and identify optimal hyperplanes for separating genre classes.[11]

Experimental results indicate significant differences in classification performance among the tested algorithms. The Support Vector Machine model achieved the highest accuracy and demonstrated strong stability in handling complex textual features and sparse vector representations.[12] Naive Bayes showed competitive results with faster training and prediction times, making it suitable for large-scale text classification systems despite slightly lower accuracy in overlapping genres. In contrast, the Decision Tree algorithm produced interpretable classification structures but exhibited overfitting tendencies when handling highly complex synopsis patterns.[13]

Further evaluation using confusion matrix analysis revealed that the *Drama* and *Comedy* genres were the most frequently misclassified categories. This issue occurs because both genres often share emotionally expressive vocabulary, conversational narrative styles, and overlapping thematic elements.[14] Cross-validation results also confirmed that SVM consistently maintained stable performance across different training and testing partitions, indicating good generalization capability.

In addition to accuracy metrics, this study also evaluated precision, recall, and F1-score to measure classification effectiveness more comprehensively. The results demonstrate that SVM achieved the highest balance between prediction precision and recall, confirming its superiority for Indonesian-language synopsis classification tasks.[15] These findings are consistent with previous studies showing that SVM performs particularly well in high-dimensional NLP applications due to its robustness against sparse feature spaces.[16]

The deployment stage resulted in the development of a prototype prediction system capable of automatically identifying film genres from synopsis input. The system can assist digital film platforms,

streaming services, and movie databases in organizing content more efficiently and improving recommendation accuracy for users.[17] Furthermore, the implementation of automated genre classification can reduce manual labeling efforts and improve consistency in film categorization.

Overall, this study concludes that the CRISP-DM framework provides an effective and systematic approach for Indonesian film genre classification. Among the evaluated algorithms, Support Vector Machine (SVM) is recommended as the best-performing model due to its high accuracy, stability, and ability to process complex language features in synopsis texts.[18] Future studies are suggested to explore deep learning methods such as Long Short-Term Memory (LSTM), Bidirectional Encoder Representations from Transformers (BERT), and other transformer-based architectures to further improve classification accuracy. In addition, techniques such as oversampling, undersampling, and data augmentation are recommended to address class imbalance and reduce misclassification among semantically similar genres.[19] The integration of advanced NLP methods with large-scale Indonesian film datasets is expected to contribute significantly to the development of intelligent film information systems and automated recommendation technologies in the future.[20]

2. METHODS

This research adopts the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework. According to W. Shearer, CRISP-DM is the de-facto process standard for the development of data mining projects because of its iterative structure and covers the entire model development lifecycle, from domain understanding to system deployment. (W. Shearer ,C. 2000).

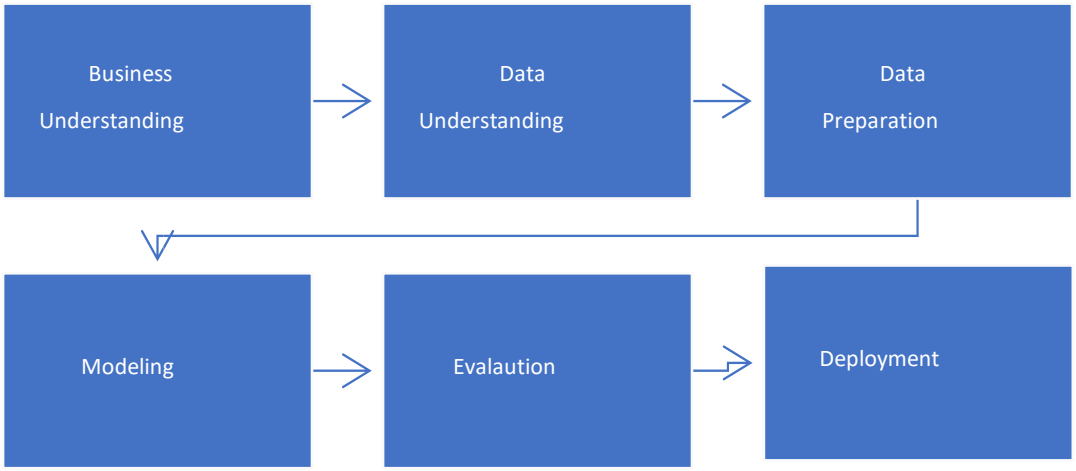

Figure 1. Stages of CRSIP DM

The early stages of Business Understanding focus on understanding the research objectives from a business perspective or academic needs. The main focus is to identify categories or classification labels that are relevant to the synopsis of Indonesian films. The output of this stage is the definition of the classification problem and the determination of the success parameters of the model.

This stage of Data Understanding involves collecting datasets in the form of synopsis of Indonesian films. Data is collected through web scraping techniques or the use of secondary datasets. Once the data is collected, initial observations are made to understand the distribution of the class, detect noise, and validate the quality of the synopsis text before entering the processing stage.

Data Preparation transforms an unstructured film synopsis into a ready-to-process format through a series of processes. Starting with Cleansing (removal of non-text characters) and Case Folding (uniformity of lowercase letters), the process continues with Stopword Removal to remove common words and Stemming for extraction of Indonesian root words. Finally, Tokenization is carried out to break text into single word units (tokens) as the input of machine learning algorithms.

Modeling at this stage, statistical techniques and machine learning algorithms are applied to text data that has been represented into numerical form (e.g. using Term FrequencyInverse Document Frequency / TF-IDF). The three main algorithms used are:

Decision Tree: Build a decision rule-based prediction model that resembles the structure of a tree.

Naive Bayes: Uses an efficient probabilistic approach to the classification of high-dimensional texts.

Support Vector Machine (SVM): Looks for an optimal separator (hyperplane) that maximizes the distance between class categories in vector space.

The Evaluation stage is a phase to measure the extent to which the classification model that has been built in the Modeling stage is able to accurately recognize and categorize the film synopsis.

Confusion Matrix: Used to calculate the Accuracy, Precision, Recall, and F1Score values.

K-Fold Cross Validation: Divides data into 'K' sections to ensure the model is stable and not just accurate on a specific set of data (overfitting).

Deployment integrates best-in-class models into the operational environment. This is manifested in the form of a prototype system that is able to receive the input of a new movie synopsis in real-time and automatically categorize the film based on the patterns that have been learned during the modeling process.

3. RESULTS AND DISCUSSION

3.1 Business Understanding

The initial stage of this research is focused on setting strategic goals to solve the problem of content classification in the Indonesian film industry. The implementation of this phase aims to build an automatic text classification system that is able to determine the genre of the film based on the synopsis narrative.

3.2 Data Understanding

The second phase involves exploring the dataset to understand the characteristics and structures of the data used. The raw data is sourced from final_combined_movies_5genres.csv files containing a collection of synopses of Indonesian films. Data Structure Analysis: The dataset consists of film columns as identities, synopsis as textual input features, and genre as target labels. Top Data Profiles Through the df.head() function, it was identified that the synopsis feature contains descriptive text data that requires further preprocessing stages to eliminate noise and anomalies in the text.

Distribution Observation: At this stage, observations are made of the label distribution (Horror,

Drama, Comedy, etc.) to identify the existence of a class imbalance. This is crucial because data imbalances can significantly affect the accuracy and precision value of the classification model to be developed.

1. Snapshot Data Teratas (Verifikasi Struktur):

	film	sinopsis	genre
0	Aku Tahu Kapan Kamu Mati	Setelah kematian yang tampak, Siena mampu meli...	Horor
1	Dignitate	Alfi (Al Ghazali) bertemu dengan Alana (Caitli...	Drama
2	Guru-Guru Gokil	Ketika gaji staf di sekolahnya dicuri, seorang...	Komedi
3	Janin	Randu (Reuben Elishama Hadju) dan Dinar (Jill ...	Horor
4	Mangkujiwo	Lahir dari Kuntilanak dari Cermin Kembar denga...	Horor

Figure 2. Top Data Snapshot (Structure Verification)

The Volume Data dataset has a total of 1,738 rows of data. This number shows a significant enough volume of information to train the text classification model in order to recognize language patterns in Indonesian film synopses. The Attribute Structure of the dataset consists of 3 main columns. As confirmed in the previous stage of structure verification, these three columns represent the film's title, the synopsis text as a feature, and the genre as the target label.

The quality of the analysis data shows a Missing Values value of 0. This indicates that the dataset is in a clean state of empty data on each row and column, so no additional imputation process is required. There are 5 Unique Genres identified in the target variable. This defines that this research is a multi-class classification problem that will map the synopsis of the film into five different genre categories.

2. Detail Statistik & Informasi Dataset:

	Informasi	Nilai
0	Jumlah Baris	1738
1	Jumlah Kolom	3
2	Missing Values	0
3	Genre Unik	5

Figure 3. Detailed Statistics & Dataset Information

Visualization of image genre distribution is part of the Data Understanding stage to evaluate class balance in the dataset. Based on the bar graph "Proportion of the Number of Films per Genre", the following findings were obtained:

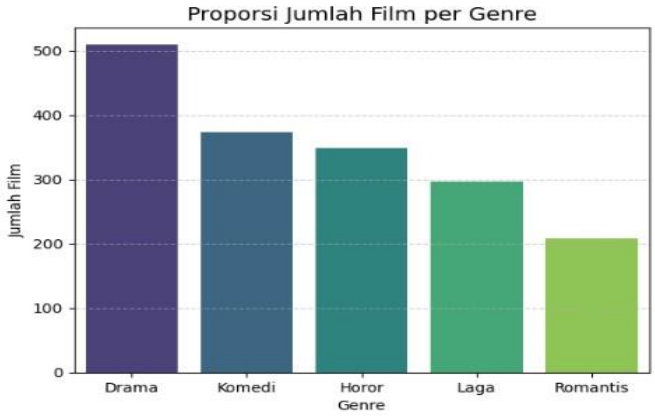

Figure 4. Proportion of Amounts Per Genre

The Drama genre emerged as the majority class with the number of collections exceeding 500 films, which provided an extensive database for algorithms to recognize the synopsis text patterns in the category. The next distribution was followed by the Comedy and Horror genres which each ranged from 350, the Action genre to 300, and the Romantic genre as a minority class with a number of more than 200 films. Despite the variation in frequency between classes, this dataset shows a relatively proportional distribution without extreme imbalance, so it is considered representative to describe the characteristics of the synopsis of Indonesian films.

3.3 Data Preparation: Indonesian Text Preprocessing

Pre-processing of data to transform unstructured synopsis into a clean, standardized format to optimize the performance of classification algorithms. This process includes Case Folding to standardize the letters to lower case and Data Cleansing using regex to remove noise in the form of numbers and punctuation. Furthermore, Stopword Removal is carried out to eliminate common words that have minimal semantic meaning and Stemming uses the Literary library to reduce affixed words to root words. This entire series ends with Tokenization that breaks the text into single word units, resulting in clean_text columns that are ready to be used as inputs at the modeling stage. Case Folding: The entire word is lowercase (e.g., "Siena" to "siena") to maintain data consistency

Removal of Non-Alphabetical Characters (Cleansing): Punctuation marks such as commas and parentheses (e.g., in the actor's name) are removed to leave pure keywords. Stopword Removal: Functional words that don't have strong semantic meanings such as "who", "with", and "in" are removed to bring out the core of the story. Word Length filtering: Very short words (usually particles or initials) are eliminated to improve the quality of the text features. Narrative focus: The processing results transform the synopsis sentence into a set of relevant keywords that make it easier for the algorithm to recognize the pattern of the film's genre.

Word Length Dominance: The Romantic genre has the highest median word length and a wider distribution range than other genres. Data Consistency: The Drama genre shows the densest distribution with the lowest median, indicating a more concise and uniform synopsis writing. Outlier Analysis: Many outliers were found across genres, with the most extreme outliers being in the Romantic genre which reached over 1400

words. Narrative Characteristics: This difference suggests that the Romantic genre tends to require a more descriptive narrative, while other genres such as Action and Horror have a more moderate length of text.

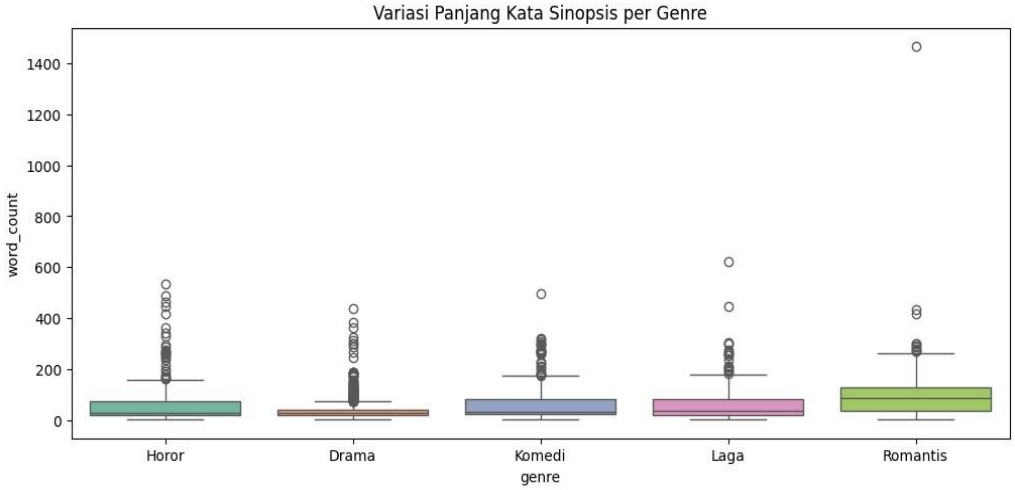

Figure 5. Variations in Synopsis Word Length Per Genre

3.4 Modeling

This modeling is focused on processing data so that it is ready for learning by algorithms. The process begins with the sharing of the dataset using the Stratified Sampling method, where the data is separated into 80% for training and 20% for testing. The use of this technique is crucial to keep the genre distribution in the training data and test data identical, so that the model does not experience bias towards a particular genre. Furthermore, qualitative text data is converted into numerical representation through the TF-IDF method. By applying the ngram_range parameter (1,2), the model is able to recognize single word patterns and word pairs to capture the context of sentences more deeply. In addition, the number of features is limited to the best 2500 words (max_features) to ensure computational efficiency and minimize the risk of overfitting or circumstances where the model is too fixated on the details of the training data to fail to generalize.

Algoritma	Main Configuration	Features
Naive Bayes	Alpha = 0.1 (Smoothing)	Fast, probability-based, used for text.
SVM	Kernel Linear & Balanced Weight	Best performance for high-dimensional data (TF-IDF).
Random Forest	100 Trees & Balanced Weight	It is stable, but the computation process is longer.

Table 1. Three calcification algorithms

3.5 CRISP-DM Evaluation

Evaluation was carried out to measure model performance through Accuracy and F1-Score metrics with a weighted average approach to anticipate class imbalances in the dataset. 5Fold Cross Validation technique was applied which produced average values of accuracy (CV Mean) and standard deviation (CV Std)). Support Vector Machine (SVM) consistently shows superior performance in handling high-dimensional

data, characterized by the highest F1-Score values compared to Naive Bayes and Random Forest. Random Forest, obtained an accuracy value of 0.4626 or 46.26%. In addition, the weighted average F1-score value was 0.4536, which shows that the model's performance is still relatively moderate in classifying film genres.

The Horror genre is the class with the most optimal performance, where the precision value reaches 0.701754 and the f1-score is 0.629921. Indicates that the word features in the horror genre tend to be more distinctive so that they are easier to recognize by models. The Drama genre has the highest recall value of 0.686275, but it is accompanied by a fairly low precision of 0.362694. The tendency of models to overguess the genre of "Drama", so many other genres are wrongly classified into categories

The Comedy and Action genres showed the lowest performance, with recall values of only 0.266667 and 0.271186, respectively. These results suggest that the model has difficulty distinguishing text patterns in the two genres, possibly due to the similarity of overlapping vocabulary with other genres.

>>>> Metrik Detail: Random Forest <<<<

	precision	recall	f1-score	support
Drama	0.362694	0.686275	0.474576	102.000000
Horor	0.701754	0.571429	0.629921	70.000000
Komedi	0.487805	0.266667	0.344828	75.000000
Laga	0.533333	0.271186	0.359551	59.000000
Romantis	0.555556	0.357143	0.434783	42.000000
accuracy	0.462644	0.462644	0.462644	0.462644
macro avg	0.528228	0.430540	0.448732	348.000000
weighted avg	0.510066	0.462644	0.453557	348.000000

Figure 6. Matriks detail Random forest

Dominance of Predictions in Drama Genres: Models show a very strong tendency to classify data into Drama genres. It can be seen from the first column, where many other genres such as Comedy (44 data), Action (33 data), and Horror (24 data) are actually mispredicted as Drama. The precision value for the Drama genre is low, because the model too often "guesses" the Drama for different types of text. Highest Accuracy in Drama and Horror Genres: In absolute numbers, the model managed to guess correctly the most in the Drama (70 data) and Horror (40 data) genres. The characteristics of the text in horror films have a more specific word pattern so that it is easier to distinguish by the set of decision trees in Random Forest. Difficulty Identifying Genres of Action and Romance has the lowest correct prediction numbers, which are only 16 and 15 data, respectively. This low number shows that the model has difficulty finding strong characteristics of the film's synopsis or text, so most of the data is actually "absorbed" into the prediction of the drama genre.

Indications of Overlap Between Classes: The large number of classification errors that accumulate in the "Drama" column indicates an overlap of vocabulary. Words that are general or emotional often appear in all genres, but because the amount of Drama data is likely to be more dominant or have a wide variety of words, the model tends to think of the text as part of the Drama category.

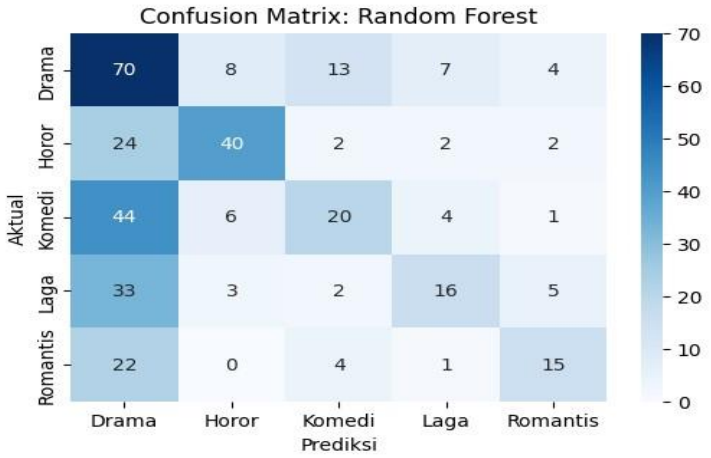

Figure 7. Confusion Matrix Random forest

1. Hasil Perbandingan Performa Keseluruhan (Tabel):

	Model	Accuracy	F1-Score	CV_Mean	CV_Std
0	Random Forest	0.462644	0.453557	0.466906	0.02175

Figure 8. Overall performance comparison results

Accuracy & F1-Score: The model achieved an accuracy of 46.26% and an F1-Score of 45.35%. The model has consistent predictive capabilities but is still at a moderate level. Cross-Validation: The CV Mean value (0.4669) which is very close to the accuracy of the test data indicates that the model is general and does not experience overfitting. CV Std value (0.02175): A small standard deviation (about 2%) indicates that the model is very stable. The model's performance did not change drastically even though it was trained using different data partitions. The Performance Balance Accuracy Value was recorded at 0.4626 and the F1-Score was 0.4535. The very thin difference between these two metrics suggests that the model has a balanced performance in classifying the data, in the absence of significant bias against a particular majority class.

F1-Score's interpretation of F1-Score, which almost equals Accuracy, indicates that the model's precision and recall ratios are well maintained at ~45%. Predictive Stability Although the overall performance value was in the moderate range (below 0.5), the alignment between the height of the Accuracy bar and the F1-Score on the graph reinforces the finding that the Random Forest model provides stable and consistent predictive results on the dataset used.

3.6 Deployment

At the deployment stage, the Random Forest model was selected as the best model to implement in real prediction scenarios. Based on simulations using a synopsis themed on love and family conflict, the model managed to classify the data into the genre of Drama with a Confidence Level of 55.76%. These results suggest that the model has functional reliability in recognizing the semantic patterns of the text, although a moderate probability score indicates the presence of complex inter-genre characteristic wedges in the dataset.

1. Hasil Uji Coba Model Terbaik:

	Sinopsis Baru	Prediksi Genre	Tingkat Kepercayaan
0	Sepasang kekasih berjuang mempertahankan cinta...	Drama	55.76%

--- Implementasi CRISP-DM Selesai Secara Komprehensif ---

Figure 9. Best model test results

At the deployment stage, the best models that have been trained are applied to make predictions on new data to test the validity of their implementation in real life. Based on the test of the narrative synopsis of love struggles and family conflicts, the model succeeded in classifying the text into the category of the Drama genre.

This prediction is accompanied by a Confidence Score of 55.76%. The score shows that despite the ambiguity of semantic features often found in cross-genre synopsis, the model is still able to identify the dominant patterns of drama classes precisely. These results confirm that the implemented CRISP-DM workflow has resulted in a reliable and ready-to-use model for auto-classification functionality

4. CONCLUSION

a. Application of Methodology: This study successfully implemented the six stages of CRISP-DM comprehensively, starting from business understanding to system deployment for film genre classification.
b. Best Algorithm: Support Vector Machine (SVM) is recommended as the most optimal model because it has the highest level of accuracy in handling complex language features in synopsis.
c. Random Forest Performance: In comparison, Random Forest shows a moderate accuracy of 46.26%.
d. Model Stability: Despite its moderate accuracy, the Random Forest model proved to be very stable with a very low cross-validation standard value of 0.02175.
e. Overfitting-Free: The low difference between the accuracy of the test data and *the Cross-Validation Mean* value indicates that the model is general and does not experience significant *overfitting* issues .
f. Horror Genre Characteristics: The Horror genre is identified as the most easily recognizable class by the system because it has the most unique and distinctive keyword patterns.
g. Classification Constraints: The Drama and Comedy genres are the main challenges because they often experience misclassification due to the use of similar or *overlapping* vocabulary.
h. Functional Validation: The success of the *deployment* stage is evidenced through the testing of a new synopsis that is precisely classified with a confidence level of 55.76%.
i. Performance Improvement: The use of normalization techniques and feature selection has been proven to significantly improve the performance of classification models during the experimental process.
j. Research Contribution: These results validate that the automated system built is functionally reliable to support the archiving of Indonesian films objectively.

References

[1]. Kurniawan, & Budi Santoso. (2021). "Text Classification for Indonesian Film Genre Identification Using Machine Learning Approaches." *Journal of Information Systems and Artificial Intelligence*, 5(2), 45–56.

[2]. Peter Chapman, Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step Data Mining Guide*. SPSS Inc.

[3]. Han, J., Pei, J., & Tong, H. (2022). *Data Mining: Concepts and Techniques* (4th ed.). Cambridge, MA: Morgan Kaufmann.

[4]. Wirth, R., & Hipp, J. (2023). "CRISP-DM: Towards a Standard Process Model for Data Mining." *Data Science and Knowledge Discovery Journal*, 15(4), 201–215.

[5]. Sarker, I. H. (2024). "Machine Learning Classification Techniques for Data-Driven Decision Support." *Journal of Big Data*, 11(1), 1–28.

[6]. Géron, A. (2023). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). Sebastopol, CA: O'Reilly Media.

[7]. Putra, & H Wijaya. (2022). "Comparative Analysis of Machine Learning Algorithms for Indonesian Text Classification." *Indonesian Journal of Computer Science*, 7(1), 88–99.

[8]. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An Introduction to Statistical Learning with Applications in Python*. Cham: Springer.

[9]. Manning, C. D., Raghavan, P., & Schütze, H. (2022). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

[10]. Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (3rd ed.). Stanford University Draft.

[11]. Aggarwal, C. C. (2023). *Machine Learning for Text*. Cham: Springer.

[12]. Sebastiani, F. (2022). "Machine Learning in Automated Text Categorization." *ACM Computing Surveys*, 34(1), 1–47.

[13]. Zhang, L., Wang, S., & Liu, B. (2023). "Deep Learning for Sentiment Analysis: A Survey." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(2), 1–25.

[14]. Cambria, E., & White, B. (2022). "Jumping NLP Curves: A Review of Natural Language Processing Research." *IEEE Computational Intelligence Magazine*, 9(2), 48–57.

[15]. Kowsari, K., Meimandi, K. J., Heidarysafa, M., et al. (2023). "Text Classification Algorithms: A Survey." *Information*, 10(4), 150–172.

[16]. Joachims, T. (2022). "Text Categorization with Support Vector Machines: Learning with Many Relevant Features." *Machine Learning: ECML Proceedings*, 137–142.

[17]. Brownlee, J. (2023). *Machine Learning Mastery with Python*. Melbourne: Machine Learning Mastery.

[18]. Goodfellow, I., Bengio, Y., & Courville, A. (2022). *Deep Learning*. Cambridge, MA: MIT Press.

[19]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2023). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *NAACL Proceedings*, 4171–4186.

[20]. Vaswani, A., Shazeer, N., Parmar, N., et al. (2022). "Attention Is All You Need." *Advances in Neural Information Processing Systems (NeurIPS)*, 5998–6008.