Jurnal Analisis Komputasi Digital

Vol.9 No.1, 2025
ISSN: 24455131

CLASSIFICATION GAME GENRE USING TF-IDF AND NAÏVE BAYES

Aditya Nugraha Irwan[1], Hasanul Fahmi[2]
President University
E-mail: aditya.irwan@student.president.ac.id[1], hasanul.fahmi@president.ac.id[2]

Abstrak

Penelitian ini mengklasifikasikan genre game berdasarkan deskripsi teks game yang tersedia di platform Steam. Penelitian ini bertujuan untuk mengklasifikasikan tiga genre game: Adventure, Casual, dan Sports. Metode yang digunakan adalah Term Frequency-Inverse Document Frequency (TF-IDF) untuk merepresentasikan teks dan algoritma Naïve Bayes untuk klasifikasi. Penelitian ini mencakup pengumpulan dataset dari platform Steam, melakukan preprocessing data, mengukur kata-kata signifikan dalam dataset, dan melakukan klasifikasi menggunakan Naïve Bayes. Tingkat akurasi yang dicapai berdasarkan deskripsi teks game adalah 64,79%, dengan skor precision, recall, dan F1 yang seimbang di semua genre. Untuk meningkatkan akurasi, peneliti menggunakan pipeline dan GridSearchCV. Akurasi klasifikasi meningkat menjadi 74,65%, menunjukkan efektivitas kombinasi TF-IDF dan Naïve Bayes dalam klasifikasi genre game.
Kata Kunci — Genre, Game, Klasifikasi, Naïve Bayes, TF-IDF.

Abstract

This study classifies game genres based on the text descriptions of games available on the Steam platform. The research aims to classify three game genres: Adventure, Casual, and Sports. It employs Term Frequency-Inverse Document Frequency (TF-IDF) to represent text and uses the Naïve Bayes algorithm for classification. The study involves collecting datasets from the Steam platform, conducting data preprocessing, measuring significant words within the dataset, and classifying using Naïve Bayes. The achieved accuracy rate based on game text descriptions is 64.79%, with balanced precision, recall, and F1-scores across all genres. In an effort to improve accuracy, the researcher employed a pipeline and GridSearchCV. The classification accuracy increased to 74.65%, demonstrating the effectiveness of combining TF-IDF and Naïve Bayes for game genre classification.
Keywords: *Genre, Game, Classification, Naïve Bayes, TF-IDF.*

1. INTRODUCTION

Genre is a category with distinct characteristics in content or themes, found in literature, music, or other forms of entertainment art [1]. The process of genre classification, traditionally performed manually, can now be automated due to advancements in computational power and is widely applied to domains such as music, films, books, and video games. This study focuses on classifying video game genres based on game titles and descriptions.

Video games are a form of entertainment that provide relaxation and have a significant impact on life [2]. Despite debates ranging from their educational benefits across age groups to criticisms concerning health, time management, and violence, such as that linked to Grand Theft Auto (GTA) the video game industry continues to grow, introducing increasingly diverse genres. Genre classification plays a key role in categorizing games based on their unique characteristics [3].

Significant advancements in the gaming industry have led to the evolution and creation of new genres. Video game genres are often more dynamic than those in books, films, and music, influenced by indie developers and the development of gaming platforms, further enriching the industry's diversity.

2. RESEARCH METHODS

This study aims to classify game genres Adventure, Casual, and Sports based on titles and descriptions from the Steam platform. Data is collected using the Steam API, including game titles, genres, and descriptions. Preprocessing steps include text cleaning, tokenization, stopword removal, and stemming to prepare the data for classification. The preprocessed data is transformed into numerical form using the TF-IDF (Term Frequency-Inverse Document Frequency) method. Naïve Bayes is then implemented to classify game genres based on the processed text.

Then, the model's performance is evaluated to analyze its accuracy in classifying game genres, providing insights into the effectiveness of the combined TF-IDF and Naïve Bayes approach. The complete workflow of the research methodology can be seen in Figure 1.

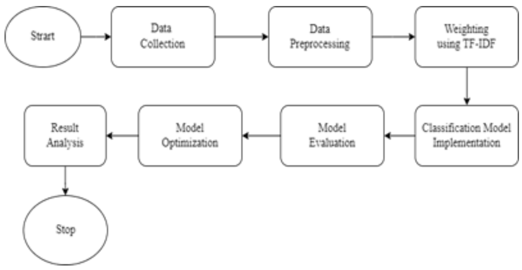

Figure 1. Research Flowchart.

3. RESULT AND DISCUSSION

This research utilizes datasets collected from the Steam platform via SteamSpy API and Steam API, focusing on game titles, genres (Adventure, Casual, and Sports), and descriptions. These datasets form the basis for training a model to classify genres based on specific language patterns. Preprocessing steps include data cleaning, text normalization, tokenization, stopword removal, and stemming to simplify text and prepare it for feature extraction.

The processed data is transformed into numerical format using Term Frequency-Inverse Document Frequency (TF-IDF), which assigns weights to words based on their relevance across the dataset. The Naïve Bayes algorithm is then applied to classify game descriptions into predefined genres, with model performance evaluated using accuracy, precision, recall, and F1-score.

A. Collecting Datasets

The dataset for this research was obtained from the Steam platform using the SteamSpy API and Steam API, designed to ensure comprehensive and accurate data collection for genre classification. The primary focus was on three genres Adventure, Casual, and Sports selected for their distinct characteristics, aiding in effective model training and testing.

Filtering by Genre

To build a genre-specific dataset, games were filtered based on their assigned genres using the SteamSpy API. This filtering ensured that only relevant games were included in the dataset. Each genre was limited to 500 games to maintain balance and prevent bias caused by genre imbalances in the data.

Game Detail Retrieval

After filtering, detailed metadata for each game, including title, genre, and description, was obtained using the Steam API. This step ensured that sufficient contextual information was available for genre classification. The API's appdetails endpoint facilitated the retrieval of game descriptions, which served as the primary textual data for analysis.

Description Text Data Cleaning

The raw game descriptions often included irrelevant elements, such as HTML tags and formatting, which could interfere with natural language processing. A cleaning process was implemented to remove these elements, ensuring that the textual data was in a plain and structured form suitable for further analysis.

Complete Data Collection Process

The entire data collection process was systematically organized to produce a structured dataset. This involved combining the filtered genre-specific lists, detailed game metadata, and cleaned descriptions into a tabular format. The resulting dataset contained three key columns: the game title, the associated genres, and the cleaned description text.

Table 3. Examples Collected Datasets

Title	Genres	Description
Street Warriors Online	Action, Massively Multiplayer, Simulation, Sports	This is the first, realistic PvP brawling game for up to 8 vs 8 players with original, dynamic combat system and fast, round based battles. (…)
Plants vs. Zombies: Battle for Neighborville™	Action, Casual, Strategy	Welcome to Neighborville, where all is well. Except that a crazy, new, coniferous conflict between brain-less and botanicals is brewing! (…)
SpongeBob SquarePants: Battle for Bikini Bottom - Rehydrated	Action, Adventure, Casual	Are you ready, kids? The cult classic is back, faithfully remade in spongetastic splendor! (…)

Data Collection for Spesific Genre

By focusing on Adventure, Casual, and Sports genres and limiting entries to 500 games per genre, the dataset achieved a balance that is critical for unbiased analysis. The structured approach ensured the dataset's relevance, completeness, and quality, forming a robust foundation for genre classification.

B. **Preprocessing Data**

This study focuses on text-based classification of game genres using a dataset comprising game titles, genres, and descriptions. To ensure the quality and consistency of the data used in the analysis, several preprocessing steps were performed, as outlined below.

Selection and Normalization of Main Genres

The initial dataset included games with various genres, but this study focuses only on three main genres: Adventure, Casual, and Sport. To ensure consistency in data grouping, the genre column was normalized by converting all text to lowercase. Sub-genres present in the initial dataset were removed to simplify the classification process by retaining only the main genres. For example, a game with the genre "Action, Adventure, Massively Multiplayer" was simplified to "adventure."

Table 4. Early datasets for the Adventure Genre

Title	Genres	Description
PUBG: BATTLEGROUNDS	Action, Adventure, Massively Multiplayer, Free To Play	LAND, LOOT, SURVIVE! Play PUBG: BATTLEGROUNDS for free. Land on strategic locations, loot (…)
Apex Legends™	Action, Adventure, Free To Play	APEX LEGENDS: SHOCKWAVE About the Game Conquer with character in Apex Legends, (…)

This process was applied separately to each main genre (Adventure, Casual, and Sport). The filtered datasets were then combined into a single dataset, and duplicates were removed based on the title column to eliminate redundancy.

Table 5. Example Datasets after Merge

Title	Genres	Description
PUBG: BATTLEGROUNDS	adventure	LAND, LOOT, SURVIVE! Play PUBG: BATTLEGROUNDS for free. Land on strategic locations, loot weapons and supplies, and survive to become the last team standing across various, diverse Battlegrounds. (…)
Among Us	casual	Play with 4-15 player online or via local WiFi as you attempt to prepare your spaceship for departure, but beware as one or more random players among the Crew are Impostors bent on killing everyone! (…)
Football Manager 2021	sports	The manager is the beating heart of every football club. In Football Manager 2021 , dynamic, true-to-life management experiences and next-level detail renews that focus on you, the manager, equipping you with all the tools you need to achieve elite status. (…)

Case Folding

All text in the dataset was converted to lowercase to eliminate discrepancies caused by capitalization. Non-alphabetic characters, such as numbers, punctuation marks, and URLs, were also removed as they are irrelevant to text classification.

Table 6. Examples from Several Game Genre

Title	Genres	Description
pubg battlegrounds	adventure	land loot survive play pubg battlegrounds for free land on strategic locations loot weapons and supplies and survive to (…)
among us	casual	play with player online or via local wifi as you attempt to prepare your spaceship for departure but beware as (…)
football manager	sport	the manager is the beating heart of every football club in football manager dynamic true to life management experiences (…)

Stopword Removal

Common words that frequently appear but do not carry significant meaning for classification, such as "and," "in," and "to," were removed. This step simplified the text and retained only words with meaningful contributions to the classification process.

Tokenization

Game descriptions were split into individual word units through tokenization. Tokenization allowed for analyzing unique words and their frequencies, which are essential for building a text-based classification model.

Stemming

The stemming process converted words into their root forms using the Snowball Stemmer method. For instance, words like "playing," "played," and "plays" were reduced to their root form, "play." This process reduced the variability of word forms without losing their core meaning, thus enhancing the accuracy of data analysis.

Table 7. Examples Datasets after Stemming

Title	Genres	Description
pubg battlegrounds	adventure	['land', 'loot', 'surviv', 'play', 'pubg', 'battleground', 'free', 'land', 'strateg', 'locat', 'loot', 'weapon', 'suppli', 'surviv'] (…)
among us	casual	['play', 'player', 'onlin', 'via', 'local', 'wifi', 'attempt', 'prepar', 'spaceship'] (…)
football manager	sport	['manag', 'beat', 'heart', 'everi', 'footbal', 'club', 'footbal', 'manag', 'dynam', 'truetolif', 'manag', 'experi', 'nextlevel', 'detail', 'renew'] (..)

Upon completing all preprocessing steps, the resulting dataset consisted of simplified game descriptions with relevant words in their root forms. This refined dataset was then used for further analysis, specifically applying the TF-IDF and Naive Bayes models to classify game genres based on titles, genres, and descriptions.

Weighting using TF-IDF

In this research, Term Frequency-Inverse Document Frequency (TF-IDF) is employed to quantify the importance of words within game descriptions relative to the entire dataset. This feature extraction method is crucial for text-based analysis, specifically for classifying game genres based on titles and descriptions.

Data Preparation

The dataset, prepared through prior preprocessing steps, includes cleaned and normalized game descriptions. The target column for TF-IDF weighting is the "description_snowball," which contains text in its simplified, stemmed form. This ensures consistency and relevance in the subsequent feature extraction process.

TF-IDF Application

The TF-IDF method transforms text into a numerical matrix, where each term is assigned a score based on its frequency within a document and its rarity across the corpus. This transformation highlights terms that are significant in specific documents but not overly common in others.

The resulting TF-IDF matrix is analyzed to extract important features:

Document-Level Analysis, The highest-scoring terms for each game description are identified. These terms reflect the unique aspects of a game's content that are relevant to its classification.

Corpus-Level Analysis, The total TF-IDF scores across all documents are summed for each term. This identifies the most frequent and significant terms in the dataset, providing an overview of prominent vocabulary.

Results and Insights

High TF-IDF scores in individual documents emphasize terms that distinguish one game from another, such as "hunt," "reserv," or "trophi" for a hunting simulation game. At the corpus level, common terms like "game," "player," and "play" rank highest due to their relevance across multiple genres.

Table 8. Highest Total TF-IDF Scores

	term	tfidf_sum
5679	game	68.301820
10934	player	42.818639
9787	new	41.904455

The TF-IDF weighting process provides a numerical representation of text features, allowing machine learning models to identify patterns and classify game genres effectively. By prioritizing significant terms, the model can focus on features that uniquely define each genre, improving classification accuracy.

Implementation of Naïve Bayes Algorithm

This study employs the Multinomial Naïve Bayes algorithm to classify game genres based on textual descriptions. The algorithm is chosen for its efficiency in handling high-dimensional data, making it suitable for text-based classification tasks.

Data Preparation and Transformation

The dataset utilized includes preprocessed text descriptions, represented numerically using the Term Frequency-Inverse Document Frequency (TF-IDF) method. This transformation quantifies the importance of terms relative to their frequency in a document and across the corpus. The dataset is then split into training and testing subsets, with an 80:20 ratio, ensuring a balanced distribution across genre classes. This split results in 1.133 samples for the training data and 284 samples for the testing data.

Model Training and Evaluation

The Multinomial Naïve Bayes model is trained on the training dataset to learn distinctive patterns associated with each genre class. The testing dataset is subsequently used to evaluate the model's performance, providing insights into its accuracy and ability to generalize to unseen data.

The model accuracy achieved an overall accuracy of 64.79%, reflecting the proportion of correct classifications. Confusion Matrix of the analysis reveals better performance in predicting "Adventure" and "Sport" genres, while the "Casual" genre exhibits higher misclassification rates.

Classification Report of the precision, recall, and F1-scores indicate variability in model performance across genres. Notably, the "Casual" genre has a low recall of 15%, highlighting challenges in identifying its unique characteristics.

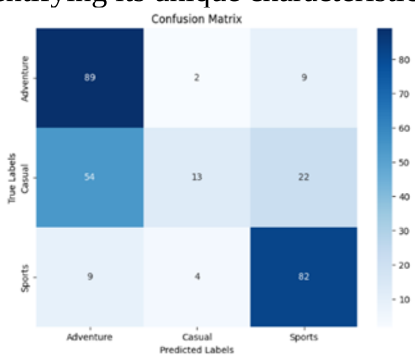

Figure 3. Confusion Matrix

While the model demonstrates reasonable performance for "Adventure" and "Sport" genres, with F1-scores of 0.71 and 0.79 respectively, its performance for the "Casual" genre is less effective (F1-score: 0.24). This suggests the need for further refinement, such as incorporating additional features or addressing class imbalances, to improve classification accuracy for underperforming genres.

Table 9. Classification Report

	Precision	Recall	F1-Score	Support
Adventure	0,59	0,89	0,71	100
Casual	0,68	0,15	0,24	89
Sport	0,73	0,86	0,79	95
Accuracy			0,65	284
Macro Avg	0,67	0,63	0,58	284
Weighted Avg	0,66	0,65	0,59	284

Model Improvement Using Pipeline and GridSearchCV

This research aims to classify game genres using textual features derived from game titles and descriptions. To enhance classification performance, a pipeline was implemented to integrate data transformation and model training, coupled with hyperparameter tuning using GridSearchCV for optimal parameter configuration.

Data Preparation and Feature Engineering

The preprocessed textual data from “title_snowball” and “description_snowball” columns were merged to create a unified text feature. This merging process enriches the data by combining the contextual information from game titles and descriptions. The merged feature provides a more comprehensive representation of each game, which is critical for accurate classification.

The dataset was split into training (80%) and testing (20%) sets, ensuring that the testing data remains unseen during model training. The split was stratified to maintain class balance, enabling an unbiased evaluation of model performance. The TF-IDF transformation was applied during the pipeline process to convert the text into a numerical matrix that quantifies the importance of each term relative to the dataset.

Pipeline Implementation and GridSearchCV Setup

A pipeline was designed to automate the sequence of text transformation and classification, TF-IDF transformation to converts text into numerical features by considering term frequency and document rarity, capturing both the significance and contextual relevance of terms. multinomial naïve bayes model is a probabilistic classifier tailored for text data, leveraging the TF-IDF features for classification. The pipeline simplifies and ensures consistency in the preprocessing and model training workflow, reducing errors and improving reproducibility.

Hyperparameter Tuning with GridSearchCV

To optimize the pipeline, GridSearchCV was employed to evaluate combinations of key parameters. TF-IDF Parameters, ngram_range: (1, 1) for unigrams and (1, 2) for unigrams and bigrams, allowing the model to capture contextual relationships between words. max_df, threshold for ignoring overly common terms (e.g., stopwords). min_df, minimum frequency cutoff to exclude rare terms that may introduce noise. Naïve bayes smoothing (α) as a regularization to handle zero-frequency issues in the dataset.

Cross-validation (5-fold) was used to assess parameter combinations systematically, ensuring robust evaluation and selection of the optimal configuration.

Results of Hyperparameter Tuning

The best configuration obtained from GridSearchCV includes:

- nb__alpha: 0.5
- tfidf__max_df: 0.85
- tfidf__min_df: 5
- tfidf__ngram_range: (1, 2)

This configuration achieved a cross-validation accuracy of 71.93%, demonstrating its effectiveness in capturing meaningful patterns from the data. The inclusion of bigrams (ngram_range: (1, 2)) improved the model’s ability to understand contextual nuances in the text.

Model Training and Evaluation

Using the optimized parameters, the model was retrained on the training set and evaluated on the test set, yielding an accuracy of 74.65%. The confusion matrix and classification report provide detailed insights into the model's performance across classes:

- Adventure: High precision (0.7), recall (0.88), and F1-score (0.78), indicating reliable classification.
- Sport: The highest F1-score (0.82) with balanced precision (0.81) and recall (0.82).
- Casual: While improved, recall (0.53) remains lower, with an F1-score of 0.62, suggesting challenges in identifying all instances of this genre.

The confusion matrix reveals that most misclassifications occur in the "Casual" genre, likely due to overlapping textual features with other genres or insufficient distinguishing terms in the dataset.

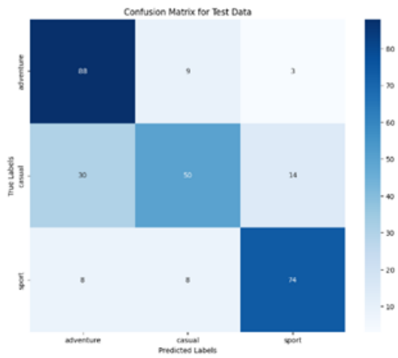

Figure 4. Confusion Matrix after Improve

Table 10. Classification Report after Improve

	Precision	Recall	F1-Score	Support
Adventure	0,7	0,88	0,78	100
Casual	0,75	0,53	0,62	94
Sport	0,81	0,82	0,82	90
Accuracy			0,75	284
Macro Avg	0,75	0,74	0,74	284
Weighted Avg	0,75	0,75	0,74	284

The combination of pipeline automation and hyperparameter tuning successfully enhanced the performance of the Naïve Bayes classifier. With an accuracy of 74.65% on the test data, the model demonstrates significant improvement compared to its unoptimized counterpart.

However, the "Casual" genre remains a challenge due to its relatively low recall. Future research could explore:

- Incorporating additional features, such as metadata or user reviews, to provide more context.

- Experimenting with alternative classifiers, such as Support Vector Machines or neural networks, to address the limitations of Naïve Bayes in this context.
- Using advanced text representation methods, such as word embeddings, to better capture semantic relationships within the text.

The methodology and results highlight the effectiveness of pipeline-based workflows and hyperparameter tuning in optimizing text classification models while identifying areas for further improvement in handling class-specific challenges.

4. CONCLUSION

This study demonstrates the effectiveness of using Term Frequency-Inverse Document Frequency (TF-IDF) for feature extraction and Naïve Bayes as a classification model for predicting game genres based on titles and descriptions. Preprocessing steps, including case folding, tokenization, stopword removal, and stemming, were instrumental in preparing data for the TF-IDF transformation. The Naïve Bayes model achieved an accuracy of 64.79%, which was further improved to 74.65% through pipeline optimization and hyperparameter tuning using GridSearchCV. These results highlight the potential of machine learning in automating genre classification in video games.

Future research can explore advanced classification models such as Support Vector Machines, Decision Trees, and Neural Networks. Enhanced preprocessing techniques, including lemmatization and entity recognition, could improve feature extraction. Incorporating metadata like user reviews or developer notes may provide additional context for classification. Multi-label classification can also be considered to accommodate games belonging to multiple genres.

5. REFERENCES

Y. Jiang and L. Zheng, “Deep learning for video game genre classification,” Multimed Tools Appl, vol. 82, pp. 21085–21099, 2023, doi: https://doi.org/10.1007/s11042-023-14560-5.

M. Barr, “Video games can develop graduate skills in higher education students: A randomised trial,” Comput Educ, vol. 113, pp. 86–97, Oct. 2017, doi: 10.1016/j.compedu.2017.05.016.

R. I. Clarke, J. H. Lee, and N. Clark, “Why Video Game Genres Fail: A Classificatory Analysis,” Games Cult, vol. 12, no. 5, pp. 445–465, Jul. 2017, doi: 10.1177/1555412015591900.

C. Suatap and K. Patanukhom, “Game Genre Classification from Icon and Screenshot Images Using Convolutional Neural Networks,” ACM International Conference Proceeding Series, pp. 51–58, Dec. 2019, doi: 10.1145/3375959.3375988.

W. Seo, S. H. Cho, P. Teisseyre, and J. Lee, “A Short Survey and Comparison of CNN-Based Music Genre Classification Using Multiple Spectral Features,” IEEE Access, vol. 12, pp. 245–257, 2024, doi: 10.1109/ACCESS.2023.3346883.

J. A. Nolazco-Flores, A. V. Guerrero-Galvan, C. Del-Valle-Soto, and L. P. Garcia-Perera, “Genre Classification of Books on Spanish,” IEEE Access, vol. 11, pp. 132878–132892, 2023, doi: 10.1109/ACCESS.2023.3332997.

D. M. Jimenez-Bravo, A. L. Murciego, J. J. Navarro-Caceres, M. Navarro-Caceres, and T. Harkin, “Identifying Irish traditional music genres using latent audio representations,” IEEE Access, 2024, doi: 10.1109/ACCESS.2024.3421639.

F. D. Adhiatma and A. Qoiriah, “Penerapan Metode TF-IDF dan Deep Neural Network untuk Analisa Sentimen pada Data Ulasan Hotel,” Journal of Informatics and Computer Science, 2022.

R. Kosasih and A. Alberto, “Analisis Sentimen Produk Permainan Menggunakan Metode TF-IDF Dan Algoritma K-Nearest Neighbor,” vol. 6, no. 1, 2021, doi: 10.30743/infotekjar.v6i1.3893.

M. I. Syafaah and L. Lestandy, “Emotional Text Classification Using TF-IDF (Term Frequency-Inverse Document Frequency) And LSTM (Long Short-Term Memory),” 2022. [Online].

Available: https://atapdata.ai/dataset/192/HIMPUNAN_DATA_E
A. E. Irsad, Y. A. Sari, and M. A. Fauzi, "Seleksi Fitur Information Gain untuk Klasifikasi Informasi Tempat Tinggal di Kota Malang Berdasarkan Tweet Menggunakan Metode Naïve Bayes dan Pembobotan TF-IDF-CF," 2019. [Online]. Available: http://j-ptiik.ub.ac.id
[H. Liu, X. Chen, and X. Liu, "A Study of the Application of Weight Distributing Method Combining Sentiment Dictionary and TF-IDF for Text Sentiment Analysis," IEEE Access, vol. 10, pp. 32280–32289, 2022, doi: 10.1109/ACCESS.2022.3160172.
R. Blanquero, E. Carrizosa, P. Ramírez-Cobo, and M. R. Sillero-Denamiel, "Variable selection for Naïve Bayes classification," Comput Oper Res, vol. 135, Nov. 2021, doi: 10.1016/j.cor.2021.105456.
F. Xu, Z. Pan, and R. Xia, "E-commerce product review sentiment classification based on a naïve Bayes continuous learning framework," Inf Process Manag, vol. 57, no. 5, Sep. 2020, doi: 10.1016/j.ipm.2020.102221.
H. Gao, X. Zeng, and C. Yao, "Application of improved distributed naive Bayesian algorithms in text classification," Journal of Supercomputing, vol. 75, no. 9, pp. 5831–5847, Sep. 2019, doi: 10.1007/s11227-019-02862-1.
N. W. Wardani, P. G. S. C. Nugraha, and G. S. Mahendra, "Implementasi Naïve Bayes Pada Data Mining Untuk Mengklasifikasikan Penjualan Barang Terlaris Pada Perusahaan Ritel," JST (Jurnal Sains dan Teknologi), vol. 12, no. 3, Jan. 2024, doi: 10.23887/jstundiksha.v12i3.38605.
H. F. Putro, R. T. Vulandari, and W. L. Y. Saptomo, "Penerapan Metode Naive Bayes Untuk Klasifikasi Pelanggan," Jurnal Teknologi Informasi dan Komunikasi (TIKomSiN), vol. 8, no. 2, Oct. 2020, doi: 10.30646/tikomsin.v8i2.500.
R. A. Ibrahem Alhayali, M. A. Ahmed, Y. M. Mohialden, and A. H. Ali, "Efficient method for breast cancer classification based on ensemble hoffeding tree and naïve Bayes," Indonesian Journal of Electrical Engineering and Computer Science, vol. 18, no. 2, pp. 1074–1080, 2020, doi: 10.11591/ijeecs.v18.i2.pp1074-1080.
N. Salmi and Z. Rustam, "Naïve Bayes Classifier Models for Predicting the Colon Cancer," in IOP Conference Series: Materials Science and Engineering, Institute of Physics Publishing, Jul. 2019. doi: 10.1088/1757-899X/546/5/052068.
N. & W. Y. K. Maneerat, "Brain tumor detection based on Naïve Bayes Classification," IEEE Access, 2019.
V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," Journal of Supercomputing, vol. 77, no. 5, pp. 5198–5219, May 2021, doi: 10.1007/s11227-020-03481-x.
R. R. Burhanuddin, "KLASIFIKASI PENYAKIT PADI MELALUI CITRA DAUN MENGGUNAKAN METODE NAIVE BAYES," Jurnal Informatika dan Teknik Elektro Terapan, vol. 12, no. 2, Apr. 2024, doi: 10.23960/jitet.v12i2.4012.
Sari, "Penerapan Text Mining Dengan Menggunakan Algoritma TF-IF Untuk Klasifikasi Genre Novel," Pelita Informatika : Informasi dan Informatika, vol. 10, no. 1, Jul. 2021.