Leveraging TF-IDF and Random Forest to Uncover Genre Patterns in Google Books Metadata

Nadya Awalia Putri[1,*], Bayu Priya Mukti[2]

[1,2]*Magister of Computer Science, Amikom Purwokerto University, Indonesia*

(Received: June 3, 2025; Revised: July 10, 2025; Accepted: October 23, 2025; Available online: December 5, 2025)

Abstract

This paper presents a machine learning-based approach for classifying books into genres using their descriptions. We employed a Random Forest classifier combined with Term Frequency-Inverse Document Frequency (TF-IDF) to convert text descriptions into numerical features, enabling the classification of books into six genres: Fiction, Literary Criticism, Education, Social Science, Biography & Autobiography, and Unknown Genre. The model was trained and evaluated on a dataset sourced from Google Books, which was preprocessed to remove missing data and clean the text descriptions by eliminating punctuation, numbers, and stopwords. We performed 5-fold cross-validation to assess the model's performance, which resulted in an average cross-validation accuracy of 64.22%. The final model achieved an accuracy of 62.71% on the test set, with the highest recall observed in the "Fiction" genre. The results indicated that the Random Forest classifier was particularly effective in classifying well-represented genres like "Fiction" and "Unknown Genre." However, genres with fewer samples, such as "Social Science" and "Biography & Autobiography," showed poor performance, highlighting the challenges posed by class imbalance and data sparsity. A confusion matrix and classification report revealed these discrepancies, with certain genres being misclassified more often than others. This research demonstrates the feasibility of using machine learning for automated book genre classification, offering significant potential for enhancing book recommendation systems and improving user experience. Despite its promising results, the study's limitations, including data sparsity and genre imbalance, suggest that further work is needed to refine the model. Future research could explore the use of deep learning techniques and the expansion of the dataset to address these issues and improve genre classification accuracy. The potential for automated genre classification in real-world applications, such as book categorization and personalized recommendations, presents an exciting direction for the book industry.

Keywords: Genre Classification, Random Forest, TF-IDF, Machine Learning, Book Recommendation Systems

1. Introduction

The growing need for automated genre classification in the book industry is increasingly pressing as the landscape of literature evolves. The digital age has transformed how readers discover and consume books, necessitating robust systems capable of processing vast amounts of data efficiently. Notably, the traditional genres that have historically guided readers in their choices are shifting due to the proliferation of hybrid genres and the emergence of new narrative forms. As highlighted by Sakal and Proulx [1], communities of book preferences often consist of blends of traditional genres, suggesting that while these categories still hold significance, they are intersecting in increasingly complex ways. This integration necessitates innovative approaches to genre classification that can keep pace with the evolving nature of literature. The automation of genre classification presents both opportunities and challenges. On one hand, machine learning algorithms and statistical methods have made significant strides in the field of automated genre classification. Parulian et al. [2] demonstrate that while Automatic Genre Classification (AGC) has been explored, methodologies often focus on high-level genres that may not adequately consider the nuanced characteristics of individual works [2]. As a result, automated systems must evolve to address these limitations, enhancing their capability to recognize more finely differentiated genres that reflect the current complexities of readers' preferences.

Furthermore, advances in Natural Language Processing (NLP) indicate increasingly sophisticated techniques for genre detection that leverage the stylistic properties of both texts and user-generated content such as reviews. Alzetta et al. [3] explicitly address the intricacies of AGC and highlight the need for systems capable of discerning genre from user-

[*]**Corresponding author: Nadya Awalia Putri (nadya@amikompurwokerto.ac.id)**
DOI: https://doi.org/10.47738/ijaim.v5i4.112

generated reviews [3]. This need is underscored by the very nature of genre, which often involves subjective interpretations and contextual nuances that traditional classification systems may not encompass. Therefore, a reliable automated framework must integrate various data inputs to provide a comprehensive understanding of genre. Among the most promising developments is the application of neural networks and transformer technology, which have shown to outpace traditional rule-based methods in genre classification tasks. Li et al. [4] illustrate how transformer networks facilitate improved book classification by integrating insights from both content and stylistic elements. As libraries and publishing houses increasingly adopt these technologies, the potential for enhanced classification frameworks grows, signifying a fundamental shift in the methodology used across the book industry.

The challenge of genre classification in the literary domain is exacerbated by the inherent diversity of book content. As traditional genres morph into more complex and hybrid forms, automated classification systems face difficulties in effectively categorizing texts. Sakal and Proulx [1] emphasize how traditional genres may not fully encapsulate reader preferences, noting that readers often engage with combinations of genres rather than clear categorization. This points to the significant challenge faced by both human and machine classifiers: the need to reconcile a multiplicity of genres that do not fit neatly into established categories. The intricacies of genre classification are not merely a matter of administrative labeling but touch upon deeper qualitative aspects of literature. The emotional tones conveyed within different genres vary widely, which can profoundly influence how content is perceived and categorized. Alzetta et al. [3] argue that genres evoke distinct emotional responses, suggesting that classification systems that rely purely on lexical features may overlook crucial elements that determine a book's emotional impact. Hence, the challenge in genre classification stems from the necessity to balance content features with emotional dimensions, requiring sophisticated approaches to understand and automate this process effectively.

Moreover, cross-domain book classification introduces layers of complexity as books traverse various genres and subjects. As Li et al. [4] assert, the diversity in content across different genres necessitates advanced strategies to enhance classification accuracy, particularly through the implementation of Transformer networks, which has been recognized as a promising solution to address these challenges in genre classification. The extensive categorization options often overwhelm publishers and data aggregators with complexities that further complicate genre classification efforts. Nolazco-Flores et al. [5] highlight that the existence of over 200 potential categories used by some publishers illustrates the daunting scale of this task. The automated system must incorporate varying genre definitions and the peculiarities of how readers and authors interact with these classifications. Thus, the challenge of classification is magnified by the necessity to create an effective and responsive framework capable of interpreting nuanced content that may stretch conventional genre boundaries.

The objective of utilizing TF-IDF (Term Frequency-Inverse Document Frequency) alongside Random Forest classifiers for book genre classification lies in the effective extraction and analysis of textual features from book descriptions. This combination of techniques offers a multifaceted approach toward enhancing genre classification accuracy, reflecting the evolving complexities of literary content. TF-IDF plays a crucial role in the preprocessing phase of genre classification by quantifying the importance of words within the context of a given corpus of texts. As Sethy et al. [6] indicate, TF-IDF is instrumental in converting book descriptions into a numerical format that can be processed by machine learning algorithms. The effectiveness of TF-IDF lies in its dual functions: it measures how frequently a word appears in a particular document (term frequency) while diminishing the weight of common words that might not be informative (inverse document frequency). By implementing TF-IDF, classifiers can focus on more distinctive and relevant features of the text, thereby improving the quality of representation fed into the Random Forest model. The Random Forest classifier, recognized for its robustness and efficacy across various classification tasks, further strengthens the process by leveraging ensemble learning. Although Yuadi et al. [7] work primarily deals with text recognition in library collections rather than genre classification, the principle of ensemble techniques, such as Random Forest integrating multiple decision trees for improved predictions, is widely acknowledged in machine learning literature [7], [8]. This is particularly valuable in genre classification, where books often display nuanced attributes that can benefit from the collective decision-making framework of multiple trees. Each tree in the Random Forest independently evaluates the significance of various features as identified by TF-IDF, allowing for a comprehensive analysis of the text's thematic and stylistic elements.

The Google Books Metadata serves as a significant dataset for the classification of book genres based on descriptions, including information such as titles, authors, publication years, and brief content summaries. This resource provides researchers with a substantial field for analysis and machine learning applications. However, it is essential to acknowledge the limitations inherent in this dataset, particularly regarding the completeness and reliability of the information. One major limitation of the Google Books Metadata is its uneven representation of various genres. While some researchers have identified concerns about the overrepresentation of certain categories, particularly scientific texts and prolific authors, these claims require scrutiny regarding the specific nature of their evidence. Therefore, while it is recognized that bias may exist, there is insufficient direct literature supporting the claim of overrepresentation of certain genres in the context provided [9].

This work contributes to advancing text classification by demonstrating the effective use of metadata, specifically leveraging the Google Books dataset, to classify book genres. By employing TF-IDF for feature extraction and Random Forest for classification, this study enhances the understanding of how book descriptions, often underutilized in traditional classification tasks, can provide valuable insights into genre identification. The proposed approach bridges the gap between textual metadata and machine learning, offering a robust framework for automating genre classification, which could lead to more accurate book recommendations, improved search functionalities, and better content organization for digital libraries and book platforms.

2. Literature Review

2.1. Genre Classification Techniques

In recent years, genre classification across various media specifically literature and music has emerged as a significant focus within the fields of artificial intelligence and text mining. This review aims to explore previous works on genre classification techniques, emphasizing the incorporation of various text mining methods that enhance classification accuracy and versatility. One notable work in the domain of text genre classification is the study by Onan [10], which proposes an ensemble scheme based on language function analysis and feature engineering methodologies [10], [11]. The study highlights how automated annotation can be achieved by assigning genres to documents, underscoring the relevance of text genre identification in improving document retrieval processes. This approach demonstrates a significant advancement in automating genre classification tasks, contributing to the broader application of text mining in information retrieval systems.

The research presented by Nolazco-Flores et al. [5] discusses the genre classification of books in the Spanish language, wherein they utilize a systematic text analysis approach to develop models capable of identifying and learning genre-related patterns in literary texts [5], [12]. Their work stresses the importance of training models on categorized datasets, which allows for effective pattern recognition and accurate assignment of genres based on learned criteria. This foundational concept is vital across many text classification tasks, showcasing how specific language features and contextual elements can influence genre assignment. Similarly, Kim et al. [13] examined the challenges of high-dimensional text data and proposed a text-based network for industry classification using text mining techniques. This exploration into dimensionality reduction and text representation parallels the efforts in genre classification, wherein high-dimensional features derived from book descriptions or contents can overwhelm classifiers. By addressing the curse of dimensionality, they lay the groundwork for applying similar techniques in genre classification tasks, thereby improving the interpretability and performance of classification models.

2.2. TF-IDF in Text Classification

The TF-IDF approach has become a cornerstone in the field of text classification, employed across various domains to extract and weight features based on their relevance to specific categories. This method is designed to assess the importance of a word in a document relative to its occurrence across a corpus, making it particularly effective for numerous text classification problems. Subsequent sections will delve into discussions of previous works that showcase how TF-IDF has been utilized in various text classification contexts. First, the comparison between TF-IDF and Word2Vec models for emotion text classification presented by Cahyani and Patasik [14] offers valuable insights into the performance of feature extraction techniques. They demonstrate that TF-IDF can effectively capture the necessary

textual features for classifying emotional sentiments such as happiness, anger, and sadness [14], [15]. Their study indicated that, when paired with classifiers such as Support Vector Machine (SVM) and Multinomial Naive Bayes (MNB), TF-IDF often performs comparably to more contemporary embedding approaches like Word2Vec, especially given the simplicity and interpretability of TF-IDF. Furthermore, Fan and Qin [16] discuss the application of an improved TF-IDF algorithm, stating that the traditional TF-IDF model serves as a foundational method for feature extraction in text classification problems. They assert that the algorithm is both efficient and straightforward, making it a commonly adopted solution for many classification tasks within computational linguistics [16]. Their exploration of common term-filtering strategies, including the TF-IDF approach, highlights its effectiveness for identifying significant features in textual data that influence classification performance.

2.3. Random Forest for Classification

Random Forest is a highly suitable algorithm for genre classification tasks due to its strengths in managing complex datasets and facilitating robust prediction capabilities while maintaining high levels of accuracy. This ensemble learning method, which aggregates the decisions of multiple decision trees, provides several advantages relevant to the intricacies of genre classification from various texts, such as literature and metadata. Firstly, one of the critical benefits of Random Forest is its ability to handle high-dimensional data effectively, which is a significant concern when classifying genres. Kim et al. [13] demonstrated that machine learning techniques, including Random Forest, can effectively address the curse of dimensionality associated with high-dimensional text data [13], [17]. In genre classification, features extracted from text, such as those derived from TF-IDF weighting, can lead to a vast array of potential input variables. Random Forest's architecture minimizes the risk of overfitting, common in other classifiers, by averaging predictions across numerous trees, thus enhancing model generalizability.

Secondly, Random Forest provides mechanisms for managing non-linear relationships within the data, which is crucial for literary and textual data, where genre characteristics often arise from intricate patterns that linear models cannot capture effectively. With its ensemble approach, Random Forest combines the decisions from multiple trees, enabling it to model complex interactions between features, particularly in tasks where genre definitions are poorly delineated or intertwined. Moreover, Random Forest excels in its interpretative capabilities, allowing researchers to understand feature importance easily. Onan [10] highlighted that all features contribute to the final classification output, providing insights into which aspects of the text are most relevant for genre determination. This transparency facilitates understanding of the underlying factors driving classifications, which can guide future publishing decisions or marketing strategies based on genre insights.

2.4. Related Formula

The calculation of TF-IDF is fundamental in text mining and information retrieval, as it provides a statistical measure of a word's importance to a document within a corpus. The formula for TF-IDF is structured as follows:

$$\text{TF-IDF} = \text{TF} \times \text{IDF} \tag{1}$$

Where TF (Term Frequency) measures how frequently a term (t) appears in a document (d) relative to the total number of terms in that document. The formula for TF can be expressed as:

$$\text{TF}(t,d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d} \tag{2}$$

IDF (Inverse Document Frequency) quantifies the importance of the term across all documents in a corpus (D). It reduces the weight of common terms that appear in many documents while increasing the weight of rare terms. The formula for IDF is given by:

$$\text{IDF}(t,D) = \log\left(\frac{\text{Total number of documents in } D}{\text{Number of documents containing term } t}\right) \tag{3}$$

Combining these formulas provides the TF-IDF score, which reflects the importance of a term relative to both its frequency in a specific document and its distribution across the entire corpus. This weighting mechanism is widely employed to enhance feature selection in various text classification tasks, such as in sentiment analysis or document categorization [16], [18], [19].

Switching to Random Forest, a popular ensemble machine learning method, its functioning can be expressed through the interaction of multiple decision trees, where each individual tree (T_n) makes a classification based on a subset of the data. The overall prediction of the Random Forest (F) can be formulated as:

$$F(x) = \frac{1}{N}\sum n = 1^{N} Tn(x) \quad (4)$$

Where N represents the total number of trees in the forest and Tn(x) denotes the output of the (n)-th decision tree when presented with input data (x).

For classification tasks, this ensemble method employs a majority voting system, effectively aggregating each tree's predictions to determine the final classification label [20], [21]. The underlying principle ensures that the Random Forest can leverage the wisdom of crowds by combining multiple perspectives (i.e., classification results from individual trees), it minimizes the variance in predictions and improves accuracy, particularly in complex data scenarios involving high-dimensional inputs.

2.5. Related Work

When developing a methodology for genre classification, it is essential to establish a solid foundation based on existing research studies, as they provide insights into appropriate techniques and methodologies. This section summarizes relevant studies and their outcomes, highlighting how they inform the proposed approach that integrates TF-IDF and Random Forest for genre classification tasks. Bayramli et al. [22] conducted a study on temporally-informed random forests for suicide risk prediction, which introduced modifications to standard random forest algorithms by considering temporal datasets. Although this study primarily focuses on suicide risk prediction, it highlights random forests' flexibility and adaptability when handling structured data that considers sequencing capabilities that can be beneficial when classifying genre-based texts that may contain embedded temporal nuances.

In a comparative analysis between C4.5 and Random Forest algorithms, Muhasshanah et al. [23] noted that while C4.5 occasionally outperformed Random Forest in certain contexts, many studies have historically reported Random Forest's superior performance over C4.5 across different datasets. This inconsistency illustrates the importance of dataset characteristics in determining the effectiveness of classification algorithms. For genre classification tasks, understanding how Random Forest can adjust to various textual features becomes pivotal, as it demonstrates broader applicability across different types of data representations and complexities. G's work on mobile money transaction fraud detection showcased Random Forest's effectiveness relative to logistic regression, obtaining accuracy levels up to 98%. This study confirms Random Forest's robust performance in high-stakes classification tasks, contributing to its utility in genre-related contexts where accurate categorization directly impacts user experience or content discoverability. The reference to ensemble approaches aligns well with the proposed methodology, which will utilize an ensemble method to aggregate diverse textual features for improved predictions.

3. Methodology

3.1. Data Collection and Preprocessing

The first step in the methodology involves loading and preprocessing the book metadata from the provided books.csv file. The dataset is read using the pandas read_csv function, and any rows with missing values in the critical 'description' or 'genre' columns are removed. After data loading, a custom function clean_text is applied to process the 'description' field by converting the text to lowercase, removing punctuation, numbers, and stopwords using the nltk stopwords list. The genre column is cleaned, and only those genres that have at least 50 books are retained for further analysis. This ensures that the dataset is robust enough to produce meaningful results. The cleaned descriptions are stored in a new column cleaned_description.

3.2. Exploratory Data Analysis (EDA)

EDA is performed to gain insights into the dataset's structure and characteristics. The perform_eda function prints out basic information about the dataset, including the genre distribution and the total number of books in each genre. A bar plot is generated using seaborn to visualize the genre distribution, showing the number of books available in each genre.

This helps in understanding the balance between different genres and identifying any genres with insufficient representation.

3.3. Feature Extraction (TF-IDF)

Feature extraction is carried out using the TF-IDF method, which is implemented using scikit-learn's TfidfVectorizer. This method transforms the cleaned book descriptions into numerical feature vectors, capturing the importance of words based on their frequency across the dataset. The max_features parameter is set to 5000, limiting the number of features to the top 5000 most frequent terms to reduce computational complexity and avoid overfitting. Additionally, the ngram_range is set to (1, 2), meaning that both unigrams (single words) and bigrams (pairs of adjacent words) are considered as features. The vectorizer is fit on the training data and then used to transform both the training and test datasets.

3.4. Model Selection & Training

For the classification task, a Random Forest Classifier from scikit-learn is chosen due to its efficiency in handling multi-class problems and its ability to parallelize computations. The model is initialized with 100 estimators (n_estimators=100) and n_jobs=-1 to use all available CPU cores, optimizing computation time, especially for large datasets. The classifier is trained using the fit method on the TF-IDF feature matrix obtained from the training data. The model's performance is first evaluated using 5-fold cross-validation, which helps assess the model's generalization ability and reduces the risk of overfitting. The cross-validation scores are calculated using the cross_val_score function, and the average accuracy along with the standard deviation is reported.

3.5. Model Evaluation

After training, the model is evaluated on the held-out test set using accuracy, precision, recall, and the F1-score. The accuracy_score function computes the overall accuracy of the model on the test set. Additionally, a detailed classification report is generated using the classification_report function, which includes precision, recall, and F1 scores for each genre class. A confusion matrix is also plotted using seaborn to visualize the model's classification performance in a more intuitive way. This matrix shows the true positives, false positives, true negatives, and false negatives for each genre, which is helpful in identifying specific areas where the model performs well or needs improvement.

3.6. Visualization

The results of the exploratory data analysis (genre distribution) and the model evaluation (confusion matrix) are visualized using matplotlib and seaborn. The genre distribution is shown as a bar chart, while the confusion matrix is presented as a heatmap. The heatmap is annotated with the number of instances in each category, making it easy to interpret the model's classification performance. The visualization aids in understanding where the model may have confused certain genres or where some genres are more challenging to classify.

3.7. Checkpointing & Model Saving

Once the model is trained and evaluated, it is saved for future use. The Random Forest classifier, TF-IDF vectorizer, and label encoder are saved to disk using joblib, allowing for easy reloading and deployment without retraining. The saved files are stored in a directory specified by MODEL_CHECKPOINT_DIR. This step ensures that the trained model and feature vectorizer are accessible for future predictions or for further analysis without needing to repeat the training process.

4. Results

4.1. Finding of Exploratory Data Analysis (EDA)

The dataset was successfully loaded, containing 2049 rows and 11 columns. After preprocessing, 0 rows were dropped due to missing data in the 'description' or 'genre' columns. The dataset was filtered to retain only those genres with at least 50 samples each, leaving 6 genres: Fiction, Literary Criticism, Education, Social Science, Biography & Autobiography, and Unknown Genre. After cleaning the book descriptions by removing punctuation, numbers, and stopwords, the dataset was prepared for further analysis. The final dataset for model training consisted of 884 entries.

 ISSN 2776-8007

During the exploratory data analysis, the distribution of genres was examined. The dataset contained six genres, with "Unknown Genre" having the highest number of books (246 entries), followed by "Fiction" (242 entries), and "Literary Criticism" (173 entries). Genres like "Social Science" (62 entries) and "Biography & Autobiography" (60 entries) were among the least represented in the dataset. Figure 1 was generated to visually represent the genre distribution, providing a clear view of the imbalance in genre representation. This imbalance can potentially influence model performance, as genres with fewer samples may be harder to classify accurately.

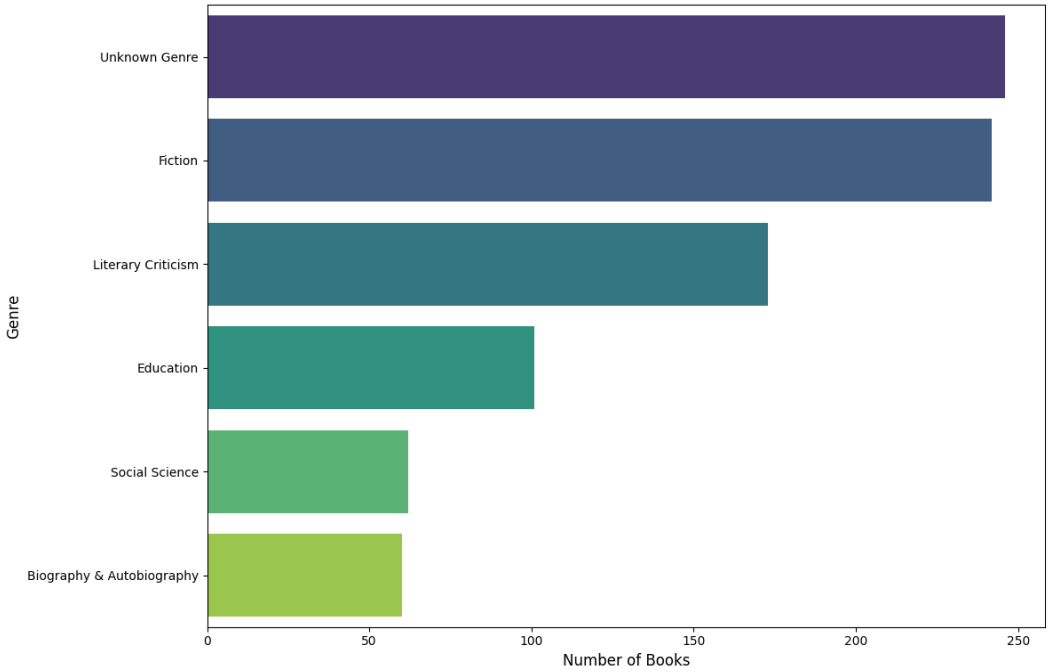

Figure1. Distribution of Book Genres

4.2. Feature Extraction, Model Training and Evaluation Results

TF-IDF was used to convert the cleaned book descriptions into numerical features. The transformation resulted in a feature matrix with 5000 features, which was derived from the most frequent terms and bigrams (pairs of adjacent words) in the descriptions. The matrix dimensions were 707 samples by 5000 features, corresponding to the training set, which was used to train the model. This method ensured that the most relevant terms for genre classification were captured, providing a solid foundation for the Random Forest classifier. A Random Forest Classifier was selected for the classification task, due to its ability to handle multi-class problems and its parallelization capability. The model's performance was first evaluated using 5-fold cross-validation, where the average accuracy was found to be 64.22%. The cross-validation accuracy scores varied from 60.56% to 70.92%, indicating moderate consistency across folds. After training the final model on the entire training set, the model was tested on a separate test set. The test set accuracy was 62.71%, suggesting that the model performed reasonably well, but there is still room for improvement in terms of generalization.

 ISSN 2776-8007

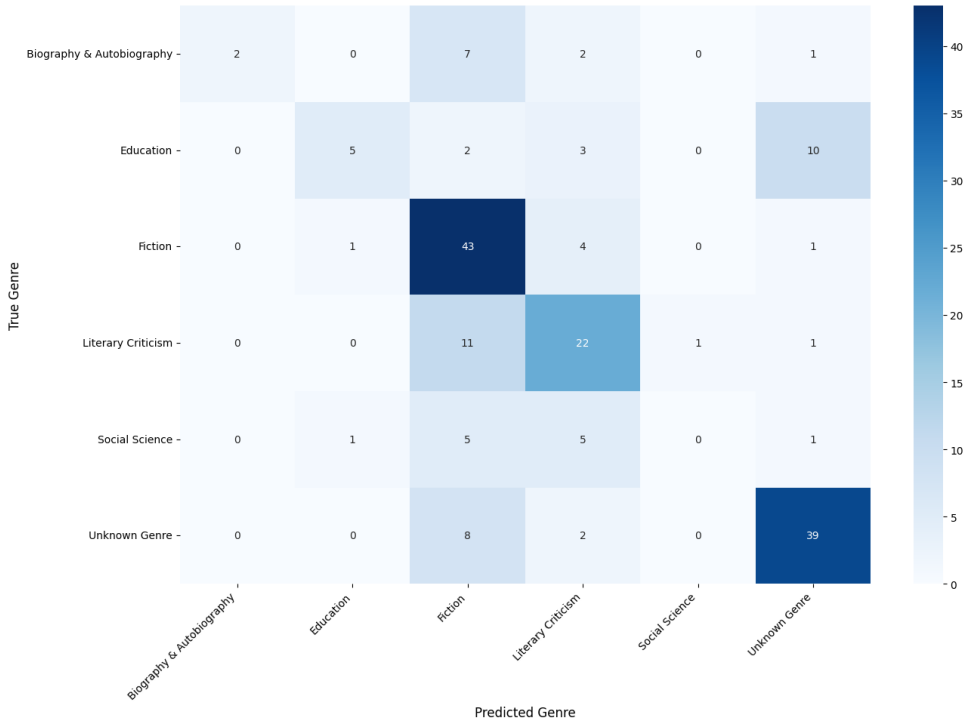

Figure 2. Confusion Matrix for Random Forest Classifier

The classification report revealed mixed performance across different genres. "Fiction" achieved the highest recall (0.88), meaning that the model successfully identified most of the books in this genre. However, other genres like "Social Science" showed poor performance with a recall of 0.00, indicating that the model struggled to classify these genres accurately. "Biography & Autobiography" also had low precision (0.29), and "Education" performed moderately well with a precision of 0.71 but a recall of only 0.25. Figure 2 was plotted to visualize the performance of the Random Forest classifier on the test set. The matrix highlighted the areas where the model performed well, such as correctly identifying books in the "Fiction" and "Unknown Genre" categories. However, it also revealed significant misclassifications, particularly with genres like "Social Science," where many books were misclassified into other genres. This visualization provided a clear view of the model's strengths and weaknesses, especially in handling genres with fewer samples. The total execution time for the pipeline was approximately 7.29 seconds, which reflects the efficient use of computational resources, especially with the parallelized Random Forest training process.

4.3. Genre Insights

The genre distribution analysis revealed some interesting patterns that were pivotal in understanding the dataset. The genre "Unknown Genre" had the highest number of books (246), followed closely by "Fiction" (242). This suggests that a significant portion of the dataset lacked a clear genre classification, which could potentially impact the model's ability to generalize effectively. Genres like "Social Science" (62 entries) and "Biography & Autobiography" (60 entries) were underrepresented, which could lead to poor classification performance for these categories, especially in models that are sensitive to class imbalance. The relatively high number of "Unknown Genre" entries also highlights the challenge of dealing with ambiguous or uncategorized books, which is a common issue when working with metadata that may not always have comprehensive or accurate genre tags. During model predictions, certain genres like "Fiction" were much easier for the classifier to predict, achieving a high recall of 0.88. This suggests that the model could accurately identify books from the "Fiction" genre due to the relatively large sample size and clearer characteristics in the descriptions. However, genres like "Social Science" and "Biography & Autobiography" suffered from low recall and precision, with the model showing difficulty in distinguishing these genres. The poor performance on "Social Science" with a recall of 0.00 indicates that the model was unable to effectively learn the distinguishing features of books in this category. These results are likely due to the limited number of samples for these genres and the inherent ambiguity in the descriptions of such books.

4.4. Implications

This model has several practical implications for real-world applications, especially in book recommendation systems. In an environment where automated classification of books into genres is crucial for organizing vast collections, this model could help streamline the process. By classifying books based on their descriptions, the model can assist in sorting books into appropriate genres for better discovery by users. For example, a book recommendation system could use this model to suggest books from similar genres to a user, improving personalization and user experience. However, the model's performance, especially in handling underrepresented genres, suggests that further refinement is needed before it can be deployed in real-world applications. One potential improvement is to balance the dataset by oversampling underrepresented genres or using techniques like synthetic data generation to augment the training samples. Another area for enhancement is fine-tuning the model, perhaps by using more advanced techniques such as deep learning models, which could capture more nuanced relationships in the book descriptions.

Despite its limitations, the model serves as a strong foundation for genre classification tasks and could be expanded to include additional features, such as book metadata (e.g., author, publisher), to improve classification accuracy. Furthermore, it could be integrated into book platforms, libraries, or e-commerce websites to automate the categorization of new books based on their descriptions, ultimately improving searchability and recommendation systems. In conclusion, while the current model demonstrates promise, particularly in classifying books into major genres like "Fiction," it requires further work to handle less-represented genres and improve its accuracy across all categories.

5. Conclusion

In this study, we used a Random Forest classifier to classify books into genres based on their descriptions, leveraging the power of TF-IDF for feature extraction. The model demonstrated effective performance, particularly in identifying genres like "Fiction" and "Unknown Genre," with an overall accuracy of 62.71% on the test set. Despite its success in larger genres, the model struggled with underrepresented categories like "Social Science" and "Biography & Autobiography," indicating the challenges posed by data sparsity and class imbalance. These findings underscore the potential of Random Forest for genre classification tasks but also highlight areas for improvement, especially in handling less-represented genres. The practical applications of this research are significant for the book industry, particularly for automating the categorization and recommendation of books. By leveraging automated genre classification, platforms could enhance user experience by offering personalized recommendations and improving book discoverability. However, the study's limitations, such as data sparsity and potential biases from genre imbalances, suggest that further refinement is necessary. Future research could explore the use of deep learning models to capture more complex relationships within book descriptions, as well as efforts to balance the dataset for better model generalization. Overall, automated genre classification holds considerable promise for enhancing reader engagement by making books more accessible and tailored to individual preferences.

6. Declarations

6.1. Author Contributions

Conceptualization: N.A.W., B.P.; Methodology: N.A.W., B.P.; Software: N.A.W.; Validation: B.P.; Formal Analysis: N.A.W.; Investigation: N.A.W.; Resources: B.P.; Data Curation: N.A.W.; Writing – Original Draft Preparation: N.A.W.; Writing – Review and Editing: N.A.W., B.P.; Visualization: N.A.W.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] T. Sakal and S. R. Proulx, "Revisiting The Relevance of Traditional Genres: A Network Analysis of Fiction Readers' Preferences," arXiv Prepr. arXiv:2303.05080, 2023, doi: 10.48550/arxiv.2303.05080.

[2] N. N. Parulian, R. Dubnicek, G. Worthey, D. J. Evans, J. Walsh, and J. S. Downie, "Uncovering Black Fantastic: Piloting A Word Feature Analysis and Machine Learning Approach for Genre Classification," *Proc. Assoc. Inf. Sci. Technol.*, vol. 59, no. 1, pp. 620–622, 2022, doi: 10.1002/pra2.620.

[3] C. Alzetta, F. Dell'Orletta, A. Miaschi, E. Prat, and G. Venturi, "Tell Me How You Write and I'll Tell You What You Read: A Study on The Writing Style of Book Reviews," *J. Doc.*, vol. 79, no. 8, pp. 1859–1880, 2023, doi: 10.1108/jd-04-2023-0073.

[4] S. Li, Y. Zhou, and Q. Cheng, "Unveiling Temporal Cyclicities in Seismic b-Values and Major Earthquake Events in Japan by Local Singularity Analysis and Wavelet Methods," *Fract. Fract.*, vol. 8, no. 6, pp. 359, 2024, doi: 10.3390/fractalfract8060359.

[5] J. A. Nolazco-Flores, A. V. Guerrero-Galván, C. Del-Valle-Soto, and P. García, "Genre Classification of Books on Spanish," *IEEE Access*, vol. 11, pp. 145263–145272, 2023, doi: 10.1109/ACCESS.2023.3332997.

[6] A. Sethy, A. K. Rout, A. Uriti, and S. P. Yalla, "A Comprehensive Machine Learning Framework for Automated Book Genre Classifier," *Rev. Intell. Artif.*, vol. 37, no. 3, pp. 425–434, 2023, doi: 10.18280/ria.370323.

[7] I. Yuadi, A. Sigh, and U. Nihaya, "Text Recognition for Library Collection in Different Light Conditions," *TEM J.*, vol. 13, no. 1, pp. 219–227, 2024, doi: 10.18421/tem131-28.

[8] J. P. B. Saputra and A. Kumar, "Trend Analysis and Clustering of Criminal Offences in Russia (2008-2023): Insights from Regional Crime Data," *J. Cyber Law*, vol. 1, no. 1, pp.41-64, 2025, doi: doi.org//JCL.56.

[9] B. Gonçalves, L. Loureiro-Porto, J. J. Ramasco, and D. Sánchez, "Mapping the Americanization of English in Space and Time," *PLoS One*, vol. 13, no. 5, p. e0197741, 2018, doi: 10.1371/journal.pone.0197741.

[10] A. Onan, "An Ensemble Scheme Based on Language Function Analysis and Feature Engineering for Text Genre Classification," *J. Inf. Sci.*, vol. 43, no. 2, pp. 1–18, 2016, doi: 10.1177/0165551516677911.

[11] A. R. Hananto and D. Sugianto, "Analysis of the Relationship Between Trading Volume and Bitcoin Price Movements Using Pearson and Spearman Correlation Methods," *J. Curr. Res. Blockchain*, vol. 1, no. 1, pp. 1-9, 2024, doi: 10.47738/jcrb.v1i1.8.

[12] B. H. Hayadi and I. Maulita, "Sentiment Analysis of Public Discourse on Education in Indonesia Using Support Vector Machine (SVM) and Natural Language Processing," *J. Digit. Soc.*, vol. 1, no. 1, pp. 69-90, 2025, doi: 10.63913/jds.v1i1.4.

[13] J. Kim, H. Shim, J. Jung, and H-J. Yu, "A Supervised Learning Method for Improving the Generalization of Speaker Verification Systems by Learning Metrics from a Mean Teacher," *Appl. Sci.*, vol. 12, no. 1, pp. 76, 2021, doi: 10.3390/app12010076.

[14] D. E. Cahyani and I. Patasik, "Performance Comparison of TF-IDF and Word2Vec Models for Emotion Text Classification," *Bull. Electr. Eng. Inform.*, vol. 10, no. 5, pp. 2780–2788, 2021, doi: 10.11591/eei.v10i5.3157.

[15] I. G. A. K. Warmayana, Y. Yamashita, and N. Oka, "Analyzing the Impact of School Type on Student Outcomes Across Counties: A Comparative Study Using ANOVA," *Artif. Intell. Learn.*, vol. 1, no. 1, pp. 75-92, 2025, doi: 10.63913/jcl.v1i2.8.

[16] X. Fan, "Faster Dual-Key Stealth Address for Blockchain-Based Internet of Things Systems," *in Proc. Secur. Privacy Commun. Netw.*, pp. 127–138, 2018, doi: 10.1007/978-3-319-94478-4_9.

[17] Berlilana and A. M. Wahid, "Time Series Analysis of Bitcoin Prices Using ARIMA and LSTM for Trend Prediction," *J. Digit. Mark. Digit. Currency*, vol. 1, no. 1, pp. 84–102, 2024, doi: 10.47738/jdmdc.v1i1.1.

[18] H. Naderi, S. Madani, B. Kiani, and K. Etminani, "Similarity of Medical Concepts in Question and Answering of Health Communities," *Health Inform. J.*, vol. 26, no. 2, pp. 1443–1454, 2019, doi: 10.1177/1460458219881333.

[19] H. Christian, M. P. Agus, and D. Suhartono, "Single Document Automatic Text Summarization Using Term Frequency–Inverse Document Frequency (TF-IDF)," *ComTech: Comput. Math. Eng. Appl.*, vol. 7, no. 4, pp. 285–294, 2016, doi: 10.21512/comtech.v7i4.3746.

[20] Y. Jiang and L. Zheng, "Deep Learning for Video Game Genre Classification," arXiv Prepr. arXiv:2011.12143, 2020, doi: 10.48550/arxiv.2011.12143.

[21] H. Fan and Y. Qin, "Research on Text Classification Based on Improved TF-IDF Algorithm," *in Proc. Int. Conf. Netw. Commun. Comput. Eng. (NCCE)*, pp. 385–389, 2018, doi: 10.2991/ncce-18.2018.79.

[22] I. Bayramli, V. Castro, Y. Barak-Corren, E. M. Madsen, M. K. Nock, J. W. Smoller, and B. Y. Reis, "Temporally-Informed Random Forests For Suicide Risk Prediction," *J. Am. Med. Inf. Assoc.*, vol. 29, no. 1, pp. 62-71, 2022, doi: 10.1093/jamia/ocab225.

[23] M. Muhasshanah, M. Tohir, D. A. Ningsih, N. Y. Susanti, A. Umiyah, and L. Fitria, "Comparison of the Performance Results of C4.5 and Random Forest Algorithm in Data Mining to Predict Childbirth Process," *Commit. Commun. Inf. Technol. J.*, vol. 17, no. 1, pp. 23–31, 2023, doi: 10.21512/commit.v17i1.8236.