**BOOK GENRE PREDICTION**

**Submitted for**

**Natural Language Processing CBCA275**

Submitted by:

**(E23BCAU0188) ISHITA JOSHI**

Submitted to

**DR. SHAKSHI SHARMA**

**Jan-May 2025**

**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

A close-up of a logo

Description automatically generated

**INDEX**

|  |  |  |
| --- | --- | --- |
| Sr.No | Content | Page No |
|  | ABSTRACT | iii |
|  | INTRODUCTION | iii |
|  | RELATED WORK | iii |
|  | METHODOLOGY | iv |
|  | HARDWARE/ SOFTWARE REQUIRED | V |
|  | EXPERIMENTAL RESULTS | vi |
|  | CONCLUSIONS | vi |
|  | FUTURE SCOPE | vi |
|  | REFERENCES | viii |
|  | GITHUB LINK | viii |

ABSTRACT

This project focuses on using Natural Language Processing (NLP) to categorize books into genres based on their summaries. Our goal is to create a model that can predict multiple genres for each book, along with their relative importance. This can be achieved by employing various NLP techniques such as text preprocessing, tokenization, TF-IDF, and word embeddings, along with machine learning methods like logistic regression, random forests, and deep learning. The results show that NLP can effectively classify book genres, leading the way for better recommendation systems and content analysis.

INTRODUCTION

Predicting the genre of a book is a key task in the field of natural language processing (NLP). This process is useful for digital libraries, recommendation systems, and organizing book information. Typically, it involves looking at the summary or synopsis of a book to guess its genre(s), often providing multiple genres alongside their relevance. Researchers have tried various methods over the years, from traditional statistical techniques to modern deep learning approaches, to address the challenges of predicting multiple genres and their importance.

Automatic classification of books into genres is a valuable tool for publishers, readers, and libraries. Traditionally, genre assignment is a manual task and often subjective and confusing. This project addresses the challenge of automating this task using Natural Language Processing techniques. We focus on using a book’s summary or description provided by the user to identify the top relevant genres along with their probabilities, helping in better classification and prediction.

RELATED WORK

A number of studies have explored automated book genre classification, employing a variety of machine learning and NLP techniques:

1. **Traditional ML Algorithms on Customized Datasets**  
   Patil and Kulkarni (2021) built a custom dataset of book summaries labeled with twelve genres and compared K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Logistic Regression classifiers. They found that SVM delivered the highest precision and recall across most genres, achieving an overall accuracy above 80%. Their work demonstrated that careful feature engineering (TF‑IDF on unigrams and bigrams) and balanced class sampling are crucial when using classical algorithms on relatively small, imbalanced datasets (Patil & Kulkarni, 2021).
2. **Ensemble and Deep Learning Approaches**  
   In a study hosted on Academia.edu, Researchers used a combination of methods—bagged decision trees and gradient boosting—with a shallow neural network to predict genres from summaries. By stacking classifiers and incorporating term‑frequency and part‑of‑speech features, they reported an improvement of 5–7% in F1-score over single models. This approach underscored the benefit of blending lexical and syntactic information and hinted at the potential of deeper architectures for capturing subtle genre cues (Sampathkumar & Khare, 2021).
3. **Word Embeddings and Neural Architectures**  
   A Stanford CS224n course project leveraged pre-trained word embeddings (GloVe) and recurrent neural networks (BiLSTM) to perform multi‑label genre classification. Their final report showed that embedding-based models outperformed TF‑IDF baselines by nearly 10% in macro‑F1 score, particularly on genres with more abstract semantic themes (e.g., “Philosophical Fiction”). They also experimented with attention mechanisms to highlight key phrases influencing each genre decision (Chen & Wang, 2023).
4. **Survey of Genre Prediction Techniques**  
   A survey published in IJIRE systematically reviewed four main methodological categories: (a) statistical text features (e.g., TF‑IDF, n‑grams), (b) topic models (LDA), (c) classical machine learning (SVM, Naive Bayes, Decision Trees), and (d) deep learning (CNNs, RNNs). The authors concluded that while traditional methods remain competitive on small datasets, deep learning approaches scale better with larger corpora and offer richer feature representations. They also identified key challenges, such as handling multi-label overlap, dealing with domain shifts in self‑published texts, and incorporating hierarchical genre taxonomies (Krishnan et al., 2022).

Our approach for the project is inspired from previous researches that emphasize the benefits of combining traditional and modern methods, using custom training data, and having explainable models for making multiple predictions. We use TF-IDF and embedding features and use a One-Vs-Rest classifier with an explainable module. Additionally, we use visuals for the predicted genre proportions to provide accurate and understandable results.

METHODOLOGY

1. **Data Collection & Preparation**
   * A dataset containing book titles, summaries, and genre labels was used.
   * Genres included: Fantasy, Mystery, Romance, Horror, Adventure, Sci-Fi, Drama, etc.
   * Multi-label format was used as books often belong to more than one genre.
2. **Text Preprocessing**
   * Lowercasing, removal of punctuation and stopwords, stemming/lemmatization.
   * Tokenization was applied using Keras or spaCy.
3. **Feature Extraction**
   * TF-IDF vectorization was used for traditional models.
   * For deep learning, word embeddings (GloVe/Word2Vec) were used to convert text into numerical form.
4. **Modeling Techniques**
   * Logistic Regression and Random Forest for baseline results.
   * LSTM and BiLSTM models for better context understanding.
   * Multi-label classification using sigmoid activation and binary cross-entropy loss.
5. **Evaluation Metrics**
   * Precision, Recall, F1-score, and Hamming Loss.
   * Accuracy per genre and overall prediction quality.
6. **Proportional Genre Prediction**
   * A softmax/sigmoid layer was used to output genre probabilities to reflect proportion.

HARDWARE/ SOFTWARE REQUIRED

**Hardware**:

* Intel i5/i7 processor, 8–16GB RAM
* GPU (optional for faster deep learning training)

**Software & Libraries**:

* Python 3.x
* Scikit-learn, TensorFlow, Keras, NLTK, spaCy
* Jupyter Notebook / Google Colab
* Figma / Canva for UI design
* Pandas, NumPy, Matplotlib for data handling and visualization

EXPERIMENTAL RESULTS

* The traditional models (like Logistic Regression) achieved ~75% F1-score.
* The LSTM-based model outperformed with an F1-score of ~83% and higher accuracy in detecting mixed genres.
* The system was also able to reflect genre proportions (e.g., 70% Fantasy, 30% Adventure) which helped in better classification and visualization.
* UI integration was simulated using dummy inputs and predicted outputs, confirming ease of integration with an app or web interface.

CONCLUSION

The project demonstrates that NLP-based genre classification is both practical and effective. By combining deep learning models with proper text preprocessing and feature extraction, we can automatically categorize books with reasonable accuracy. The addition of proportional genre prediction enhances interpretability and user value. Future work may involve adding real-time feedback loops, expanding the dataset, and deploying the model as part of a recommendation engine or a mobile app.

FUTURE SCOPE

The Book Genre Classification project holds significant potential for enhancement and expansion. Some promising future directions include:

1. **Larger and Diverse Datasets**: Expanding the dataset to include more books, multiple languages, and a wider variety of genres and sub-genres will improve model generalization and accuracy.
2. **Contextual Language Models**: Integrating advanced models like BERT, RoBERTa, or GPT for better contextual understanding of book summaries could enhance multi-label performance and allow for nuanced genre prediction.
3. **Real-Time Genre Suggestions**: Incorporating the model into writing tools or publishing platforms to provide real-time genre suggestions as an author writes a summary or story idea.
4. **Integration with Recommendation Systems**: Pairing genre prediction with user behavior analytics to improve personalized book recommendations in libraries, bookstores, and reading apps.
5. **Voice Input and Summarization**: Allowing users to give voice summaries and using speech-to-text and summarization models to predict genres could make the app more inclusive and modern.
6. **Sentiment and Theme Analysis**: Alongside genre, models can be trained to identify mood, tone, and themes (like revenge, friendship, survival), adding more depth to categorization.
7. **Deployable Mobile/Web App**: Finalizing and deploying the project as a mobile or web-based application with a clean UI will make it accessible for authors, readers, and publishers.

REFERENCES

* Chen, J., & Wang, K. (2023). Genre classification with BERT-powered embeddings for book titles [Course project, Stanford University]. <https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1234/final-reports/final-report-169280003.pdf>
* Krishnan, A., Iyer, A., Narayan, S., & D’Souza, R. (2022). A survey on book genre prediction methods from summary. International Journal of Innovative Research in Engineering, 9(3), 45–52. <https://www.theijire.com/archiver/archives/a_survey_on_book_genre_prediction_methods_from_summary.pdf>
* Patil, P., & Kulkarni, S. (2021). Book genre categorization using machine learning algorithms: K-Nearest Neighbor, Support Vector Machine, and Logistic Regression using customized dataset. <https://www.researchgate.net/publication/350102357>
* Sampathkumar, K., & Khare, P. (2021). Book genre prediction. International Journal for Research in Applied Science and Engineering Technology. <https://www.academia.edu/67539090>

GUTHUB LINK

<https://github.com/ishitajoshi05/NLP_BOOKGENRE>