IT1244 Project Report Team 14

River Goh, Sean Tan, Tan Je-Ric

Introduction

Genre classification plays a crucial role in information retrieval systems, recommendation engines, and digital libraries. As the number of published literature continues to grow, manually tagging books by genre becomes incredibly time consuming. Using machine learning, deep learning models, and Natural Language Processing (NLP) techniques, the process of inferring a book's genre based on its textual synopsis can be automated to a certain degree, allowing for more efficiency.

In this project, we address the **Book Genre Prediction** problem. The goal being to predict a book's genre given its textual synopsis. We adopt a progressive modelling strategy, starting from classical bag-of-words representations to contextual embeddings to Transformers. This allows us to investigate how semantic understanding and contextualisation improves model performance.

We will implement a baseline model using Term Frequency-Inverse Document Frequency (TF-IDF) and Logistic Regression to investigate how well the model performs solely treating text as unordered token counts. We will then incorporate embedding-based models using Global Vectors (GloVe) for word representations fed into a shallow neural network, introducing distributed word representations that capture semantic similarity. Finally, we will transition to Transformer-based models, leveraging Bidirectional Encoder Representations from Transformers (BERT) embeddings and its distilled version (DistilBERT) fine-tuning to understand contextual meaning and contextual relationships across a sentence.

Related Work

TF-IDF's bag-of-words approach was compared against word embeddings which captured the semantic meaning of words, demonstrating how word embeddings such as Word2Vec generally perform better than TF-IDF (Tamanna 2025)**.**

However, recent studies demonstrate that Transformer-based architectures outperform classical text classifiers across multiple domains. BERT-based models are better suited for genre classification in sentences because they capture context and semantic depth that shallow embeddings cannot (Schnick and Schütze 2020).

DistilBERT is a lighter, yet comparably powerful variant optimised for efficiency (Galke et al. 2025). The authors mention that DistilBERT 'retains 97% of BERT's performance while using 40% fewer parameters and running 60% faster, making it a superior and practical alternative for most classification tasks.'

Our project builds upon these insights by comparing these model families under a common dataset and evaluation framework.

Dataset

Dataset Overview

The dataset used contained book synopses and corresponding genre labels, with ten classes: fantasy, science, crime, history, horror, thriller, psychology, romance, sports, and travel. Each record included a text summary (typically 100-450 words) describing the book's plot or theme.

Data Cleaning and Preprocessing

We conducted an extensive cleaning pipeline to ensure textual quality and consistency:

1. **Duplicate removal:** 18 duplicated rows were identified and dropped.
2. **Null check:** No missing values were found.
3. **Length filtering:** Extremely short synopses (<10 words) were removed. Summaries 500 words or longer were summarised via PegasusLarge into 300-word summaries.
4. **Token-level cleaning:** All text was lowercased, URLs and non-alphabetic characters removed, and stopwords filtered using the NLTK library.
5. **Nonsense detection:** Using the NLTK English vocabulary list, synopses

containing mostly non-dictionary tokens were flagged and discarded.

6. **Genre mapping:** Numerical genre IDs were mapped to human-readable labels based on the README specification.

Exploratory Data Analysis

The dataset exhibited moderate class imbalance, with *romance, travel, psychology and sports* genres being more infrequent than the other genres. As well as *thriller* and *fantasy* being more common than the other genres.

Several visual analyses were performed:

Genre distribution plot (Figure 1) shows class balance across genres.

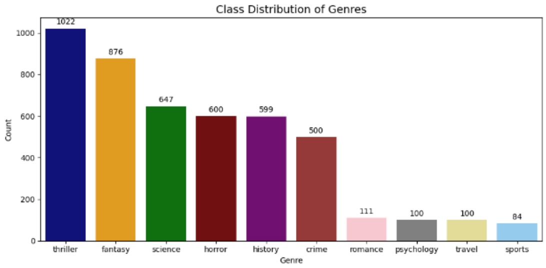
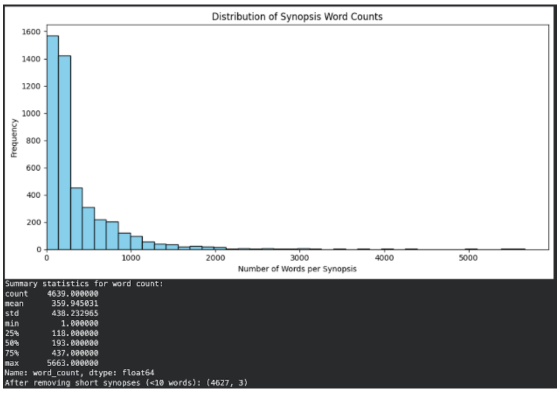


Figure 1

Word count histogram (Figure 2) illustrates the range and typical length of summaries.


Distribution of Synopsis Word Counts
Frequency
Number of Words per Synopsis
1600
1400
1200
1000
800
600
400
200
0
0
1000
2000
3000
4000
5000

```
Summary statistics for word count:
count    4639.000000
mean      359.945031
std       438.232965
min         1.000000
25%       118.000000
50%       193.000000
75%       437.000000
max      5663.000000
Name: word_count, dtype: float64
After removing short synopses (<10 words): (4627, 3)
```

Figure 2

Word clouds per genre (Figure 3) highlights common thematic keywords (e.g., “travel,” “world” in *travel*; “murder,” “death” in *Crime*).

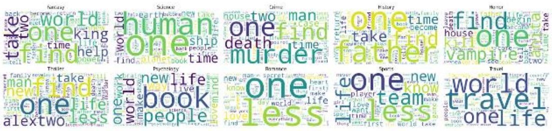


Figure 3

After cleaning, the final dataset consisted of *4613* samples (after Token-level cleaning & nonsense filtering) and 10 genre classes. The processed dataset was saved as **cleaned_books.csv** for downstream modelling.

Methods

Problem Formulation

The goal of this project is to automatically classify books into their respective genres based on their textual synopsis. This is a multi-class text classification problem, where each input, a book synopsis is mapped to a discrete label representing one of ten genres such as *fantasy, thriller, travel, etc.*

The model is then evaluated using **5-fold cross-validation** on the training set and a held-out 20% test set. Accuracy and F1-score are used as evaluation metrics.

Model Pipeline

Our pipeline progresses through three main stages:

Stage 1: Baseline Model (TF-IDF + Logistic Regression)

Converts text into TF-IDF vectors, followed by a linear classifier. Used mainly for its simple and minimalist bag-of-words style processing.

TF-IDF models provide interpretability and speed, establishing a classical baseline and revealing how far simple frequency-based models can go.

Stage 2: Embedding-Based Model (GloVe + Shallow Neural Network)

Each synopsis is represented as the mean of pre-trained word embeddings (GloVe). A one-hidden-layer neural network classifies the averaged embeddings. This captures semantic relationships between words missing from TF-IDF.

Word embeddings address TF-IDF's limitations by introducing semantic similarity and continuous feature spaces.

Stage 3: Transformer-Based Models

Transformer models introduce contextualised understanding of language, where word meaning depends on surrounding text.

Stage 3a: Frozen BERT embeddings + Logistic Regression

Uses pretrained contextual embeddings as fixed feature extractors, highlighting the impact of deeper contextual representation without fine-tuning.

Unlike GloVe, BERT captures word meaning in context, allowing the model to differentiate between polysemous words and nuanced sentence structures.

Stage 3b: Frozen BERT + PCA + Logistic Regression

Extends the frozen BERT approach by applying Principal Component Analysis (PCA) to reduce redundancy in high-dimensional BERT embeddings.

PCA helps the classifier focus on the most informative semantic dimensions, improving feature efficiency and reducing noise.

Stage 3c: Fine-tuned DistilBERT

Adapts the pretrained Transformer weights to our specific genre dataset, capturing task-specific nuances and providing the strongest end-to-end contextual understanding.

This staged approach allows us to incrementally address the limitations of earlier models while learning the theoretical progression of NLP techniques, from lexical counts to deep contextual representations.

Model Flow

Each modelling stage solved a **specific limitation** of the previous one: TF-IDF (no semantic meaning) → Embeddings (semantic meaning but no context) → BERT (context, fine-tuned for task). The entire model pipeline is captured in the flowchart below (Figure 4).

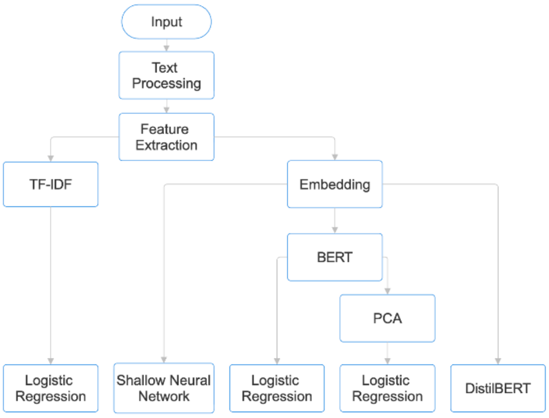


Figure 4

Results & Discussions

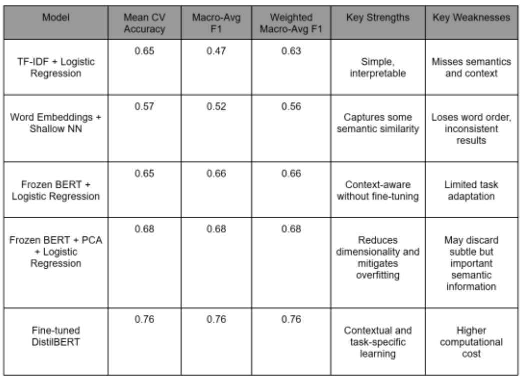
Evaluation Setup

The **DistilBERT model** was fine-tuned on the dataset using a 5-fold cross-validation setup with a weighted cross-entropy loss to address class imbalance. Each fold was trained for four epochs using a learning rate of 3×10^{-5}, a batch size of 8, and a maximum sequence length of 256 tokens.

The increased sequence length (compared to the typical 128) allowed the model to capture more contextual information from longer book summaries, while the smaller batch size balanced memory constraints during fine-tuning. The learning rate of 3×10^{-5} provided stable gradient updates to the pretrained weights, and four epochs offered a balance between sufficient fine-tuning and minimising overfitting.

Weighted loss ensured that underrepresented genres were penalised more heavily, improving the model's ability to generalise across all classes. Cross-validation further strengthened reliability by averaging performance across multiple data splits.

Quantitative Results Summary:



Model	Mean CV Accuracy	Macro-Avg F1	Weighted Macro-Avg F1	Key Strengths	Key Weaknesses
TF-IDF + Logistic Regression	0.65	0.47	0.63	Simple, interpretable	Misses semantics and context
Word Embeddings + Shallow NN	0.57	0.52	0.56	Captures some semantic similarity	Loses word order, inconsistent results
Frozen BERT + Logistic Regression	0.65	0.66	0.66	Context-aware without fine-tuning	Limited task adaptation
Frozen BERT + PCA + Logistic Regression	0.68	0.68	0.68	Reduces dimensionality and mitigates overfitting	May discard subtle but important semantic information
Fine-tuned DistilBERT	0.76	0.76	0.76	Contextual and task-specific learning	Higher computational cost

Table 1

Result Analysis

1. TF-IDF + Logistic Regression

The baseline model achieved a weighted F1 of 0.63, performing best on *fantasy*, *history*, and *science* genres. These categories often contain domain-specific keywords (e.g., *wizard*, *experiment*, *empire*), which are easily picked up by TF-IDF features. However, it struggled with *romance*, *psychology*, and *travel* hence affecting Avg F1. These genres are defined more by tone and narrative style than by specific keywords. This reflects TF-IDF's core limitation: it counts words but doesn't understand meaning or context.

2. Word Embeddings + Shallow NN

The embedding-based model scored slightly lower (weighted F1 = 0.56), showing improvement for *psychology* and *sports* but decline in *fantasy* and *romance*. While embeddings capture semantic similarity, averaging them loses sentence structure and context. For instance, "He won the match" and "He lost the match" become almost identical. This step was crucial for learning how continuous vector spaces represent meaning, but also for understanding why word-level embeddings alone are insufficient for nuanced text classification.

3. Frozen BERT + Logistic Regression

The frozen BERT model produced an F1 of 0.66, roughly matching TF-IDF + Logistic Regression but showing a more balanced performance across genres. Notably, *psychology* and *travel* improved sharply, since contextual embeddings helped the model interpret words differently depending on surrounding text (e.g., "journey" in travel vs. metaphorical use in self-help). This demonstrated that contextualised embeddings outperform static embeddings, even without additional training.

4. Frozen BERT + PCA + Logistic Regression

The Frozen BERT + PCA + Logistic Regression model achieved an overall accuracy of 0.68, representing a slight improvement over the Frozen BERT + Logistic Regression model without PCA. This indicates that applying PCA to the high-dimensional BERT embeddings helped reduce redundancy and noise, enabling the classifier to generalise better across genres. Performance gains are particularly visible in categories such as *fantasy*, *sports*, and *travel*, which showed stronger recall and F1-scores. However, improvements were modest overall, suggesting that while PCA enhances model stability and efficiency, the frozen nature of BERT still limits deeper task-specific feature adaptation.

5. Fine-tuned DistilBERT

Fine-tuning DistilBERT yielded the highest performance (CV accuracy = 0.757, weighted F1 = 0.76). It learned subtle genre-specific linguistic cues such as emotional tone, narrative focus, and character relationships that previous models ignored. However, *romance* remained challenging (F1 = 0.57), likely due to fewer training samples and overlapping vocabulary with *fantasy* and *psychology*. This step illustrates how fine-tuning improves task alignment but also requires careful regularisation to prevent overfitting on smaller genres.

Human vs Model Performance

Humans reading synopses can easily infer genre because they understand narrative intent and cultural references, while models rely solely on surface-level text. Thus, our model's ~76% accuracy is reasonable and already useful for assisting human cataloguing or recommendation systems, even if not surpassing human-level understanding. It need not necessarily replace a human but can also be used as a powerful tool to aid humans.

Ethical & Societal Considerations

Transparency and Trust: If users (like publishers, authors, or readers) can't understand *why* the model assigns a genre, they may lose trust in it or misuse its output. Interpretability helps people see that the system's decisions are based on logical textual patterns rather than arbitrary or biased features.

Impact on jobs: Automated genre classification could streamline the book cataloguing process. Possibly leading to a reduction of human headcount required to perform such tasks. However, it should not completely replace humans as human editors are still important to ensure nuance and cultural sensitivity are preserved.

References

Online Article / Blog Post
Tamanna. 2025. Why Embeddings Usually Outperform TF-IDF. *Medium*, August 6. https://medium.com/@tam.tamanna18/exploring-the-power-of-nlp-why-embeddings-usually-outperform-tf-idf-98742e7b0bce.

Conference Paper
Schnick, T., and Schütze, H. 2020. BERTRAM: Improved Word Embeddings Have Big Impact on

Contextualized Model Performance. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. https://aclanthology.org/2020.acl-main.368.pdf.

Preprint Server

Galke, L., et al. 2025. Are We Really Making Much Progress in Text Classification? *arXiv preprint*. arXiv:2204.03954v6 [cs.CL]. Ithaca, NY: Cornell University Library. https://arxiv.org/html/2204.03954v6.