DESIDOC Journal of Library & Information Technology, Vol. 45, No. 3, May 2025, pp. 169-176, DOI : 10.14429/djlit.20026

Leveraging Book Genre Classification using Machine Learning

Jhimli Adhikari

Department of Computer Science, Narayan Zantye College, Bicholim, Goa - 403 529, India
E-mail: jhimli_adhikari@yahoo.co.in

ABSTRACT

One helpful tool for book recommendations is a book summary. This article discusses categorizing books only on the basis of their title and summary, without taking into account the author's background or place of origin. The title and abstract of the book make reference to the machine learning methods used to create the genre. This study assesses the capacity to distinguish between books based on their title and summary using four machine learning models. The dataset that can be found on the Kaggle website includes 10 distinct genre kinds and 4657 instances. The dataset is first subjected to exploratory data analysis, and then a machine learning based strategy is used to extract features from the book's title and abstract using natural language processing techniques. We use 80 % of the samples (3,726 instances) to train the models, and the remaining 20 % (931 instances) are used for testing. Every model's performance is evaluated using a range of metrics, including accuracy, precision, recall, and F1score. The method also determines which words are most frequently used in each genre. Systems for automatically classifying books and making recommendations can be built using this framework.

Keywords: Classification; Genre; Machine learning; Natural language processing; Prediction; Summary

1. INTRODUCTION

Advances in cutting-edge technology has changed the ways we gather, examine, and share knowledge. The necessity for text data classification and categorisation is increasing day by day as the volume of data increases exponentially[1-2]. In an effort to enhance their offerings and competitive edge, libraries are also implementing new technology. As a result, automatic book classification has emerged as a crucial service for libraries[3]. Simple classifications known as book genres enable readers to identify the kind of book they are now reading. They also help book publishers comprehend the kind of book they are expected to write. The main goal of developing this model is to use machine learning techniques to significantly speed up the laborious process of classifying data based on book titles and summaries. Thus, it is necessary and essential to comprehend the genre when submitting a query. It helps by giving readers the important details about the book, which enhances marketing. It also benefits the online catalogues and digital repositories, and target interested consumers in E-Commerce platforms[4].

As Natural Language Processing (NLP) evolves, computer programs can now understand human language[5]. The classification of books based on their names and summaries is the basis of this article. The synopsis of the book is a crucial resource for identifying its genre. One of the main causes of this is the textual content of books, which is significantly more than that of most other text medium. Consequently, instead of working with the entire text, we'll be dealing with book titles and summaries. The proposed method extracts knowledge from book summaries. In this regard this paper introduces four machine learning algorithms such as Multinomial Naive Bayes, Logistic regression, Support Vector Classifier and Random Forest for classification. There were no prior attempts that we could find aimed at tackling the particular problem of classifying books genres by title and summary. The goal of the article is to develop a framework that can automatically find new books and assess how well they do in automatically determining a book's correct category based just on its title and synopsis. In order to develop a classifier that may be used to assign future book descriptions to specific genres, a model is trained on a dataset of books in this article. A selection of the many real world applications for the predictive model employed in this work are listed below.

1. To correctly identify a book's genre, making it simpler to classify books for use in retailers, libraries, and online marketplaces.
2. To recommend books to readers according to their interests
3. To automatically label books with pertinent genres and search terms, facilitating book searches and discovery
4. To assist publishers and marketers in creating more successful marketing plans

Received : 18 March 2024, Revised : 10 April 2025
Accepted : 16 April 2025, Online published : 09 May 2025

The paper proceeds with the following contributions .

1. Examine how the classification of book genres has evolved recently.
2. Before using the models, a thorough Exploratory Data Analysis (EDA) is performed on the dataset .
3. Transform the data into a format that is useful and to produce a list of frequently occurring words for each genre.
4. Evaluates the efficiency for the implemented models using real dataset.
5. Evaluates all classifiers performance using four well known evaluation metrics: Accuracy, Precision, Recall and F1-score.

After presenting related work in section 2, research methodology is described in section 3. Several machine learning algorithms and criteria for evaluating the performance of the proposed methods are described in section 4. Section 5 presents exploratory data analysis and result evaluation about the model. Conclusion and future work is presented in section 6.

2. RELATED WORK

The classification of books based on genres has been a widely explored problem in recent academic literature. Various methods have been proposed, each leveraging different input features such as book titles, cover images,

Table 1. Comparison table for various techniques

Title (Year)	Methodologies	Findings	Dataset used
"Book Genre Classification Based on Reviews of Portuguese-Language Literature[6] (2022)"	Naive Bayes, Decision Tree, Random Forest, SGD, KNN	Random Forest achieved the best accuracy (96%)	PPORTAL - Portuguese literature dataset
"Deep Learning Approaches towards Book Covers Classification[7] (2018)"	Convolutional Neural Networks (CNN)	Genre prediction from covers reached 60% accuracy	IMDb book cover dataset
"Book Genre Classification Based on Titles with Comparative Machine Learning Algorithms[8] (2019)"	Recurrent Neural Networks (RNN), Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), Bi-Directional LSTM (Bi-LSTM), Convolutional Neural Networks (CNN)	LSTM outperformed other models with 65.58% accuracy	GitHub - Book Genre Classification
"Classification of Book Genres using Book Cover and Title[9] (2019)"	Feature fusion + Logistic Regression	Highest accuracy achieved (87.2%) when combining cover and title	Google Books API, OpenLibrary API
"Enriching BERT with Knowledge Graph Embeddings for Document Classification[10] (2019)"	BERT + Knowledge Graph Embedding	F1-score of 64.70 on multi-label classification	GermEval 2019 - 20k+ German books
"Book Genre Categorisation Using Machine Learning Algorithms (K-Nearest Neighbor, Support Vector Machine and Logistic Regression) using Customised Dataset[11] (2021)"	K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Logistic Regression (LR)	SVM demonstrated the best speed and accuracy	CMU book summary + The Blurb Genre Collection
"A Survey on Book Genre Prediction Methods from Summary[12] (2022)"	Supervised Learning and Deep Learning approaches	KNN showed highest classification accuracy	CMU book summary Dataset
"Cover-based multiple book genre recognition using an improved multimodal network[13] (2023)"	CNNs (text-only, image-only, hybrid)	Hybrid model reached 69.09 % (Latin) and 38.12 % (Arabic)	Latin and Arabic book cover datasets

summaries, or a combination of these. Table 1 provides a comparative overview of some notable studies in this domain, focusing on their methodologies, outcomes, and datasets used.

From Table 1 it is seen that book genre classification is made using title[8], book cover and title[9], title and summary[11] and only cover[13]. Various datasets are also used across different studies. In relation to the previously stated literature, this work is important in the following ways.

- The entire book's material is not necessary for the suggested method. This is advantageous because it doesn't rely on intricate methods or a sparse amount of knowledge from the book.
- Model accurately predicts book genres based on limited metadata (title and summary). This approach is particularly useful for applications where the full text is unavailable, such as online bookstores, digital libraries, and recommendation systems.
- We provide a system that uses the book title and summary to classify books according to genre.
- Most common words occurring for each genre are listed.

3. RESEARCH METHODOLOGY

Classification is one of the basic and very important tasks in machine learning field[14]. Machine learning is a two-stage process: the learning step and the prediction step. The model is created in the learning step using the provided training data. The technique is used to forecast the response for the provided data in the prediction step. The study's proposed methodology is shown in Fig. 1, and each of its elements is explained in the paragraphs that follow.

A crucial stage in machine learning is data preprocessing. The more thoroughly cleansed data we feed the computer, the better it will be trained and able to produce an output through data analysis. Data cleansing involves deleting duplicate values, unnecessary tags, commas, and grammar because if we feed a machine meaningless data, it will produce garbage in return. In this case, the titles and summaries were cleaned using the Natural Language Toolkit (NLTK)[15].

3.1 Title and Summary Pre-processing

The text dataset selected for genre classification needs to be pre-processed. The following are the commonly used stages applied for text cleaning:

1. Combine book title and summary
2. Conversion to lowercase
3. Punctuation removal: In this step, all the punctuations from book title and summary are removed.
4. Eliminate Stop Words: Commonly used terms known as "stop words" are removed from the text because they don't provide useful information to the study. These are words with little or no meaning.
5. Stemming: The process of stemming or diminishing words to their root or base form is also referred to as the text standardization phase. This helps to limit the words in vocabulary.
6. Lemmatization: It stems the word but makes sure that it does not lose its meaning. Lemmatization has a pre-defined dictionary that stores the relationship of words and checks the word in the dictionary while diminishing.
7. For feature representation, we employ the term frequency-inverse document frequency method. This method considers a term's importance in context to the entire text.
8. Feature extraction: This method reliably determines the frequency and significance of words by converting textual data into numerical information.
9. Word-to-number conversion: Converting each category 'genre' feature to a numerical value.
10. Data is split into training and testing set. We train the models with 80% of the samples (3,726 instances) and test with the remaining 20% (931 instances)
11. Models are assessed using statistical score

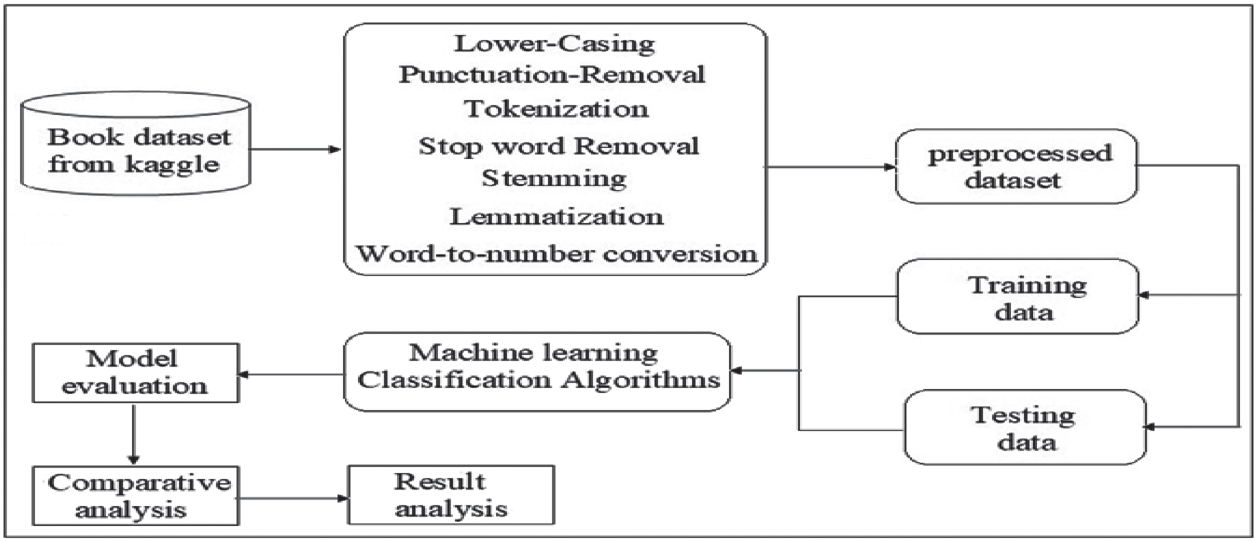

Figure 1. Architecture of the proposed approach.

3.2 Machine Learning Models

In this section we describe four machine learning techniques Multinomial Naive Bayes, Logistic regression, Support Vector Classifier and Random Forest to classify text (title and summary). Such classifiers are among the most known ones and implemented using scikit-learn.

- Multinomial Naive Bayes (MNB)[16] is a straightforward but effective probabilistic classification method that scales to enormous datasets and parallelizes well. This technique can be applied when there are several classes to be classified. It predicts the text's label by calculating the likelihood of each label for the input text and then outputs the label with the highest probability. The statement indicates that the state of one feature inside a class does not impact the state of another feature. It operates on the principle of term frequency, or the number of times a word appears in a document. This model provides two facts: the frequency of the word and if it appears in the document.
- Logistic regression[17] is used to categorise dataset records according to the values of their input fields. In order to anticipate results, it makes predictions about a dependent variable based on one or more sets of independent variables. This applies to both multiclass and binary classification.
- Support Vector Classifier (SVC)[18] is a specific implementation of the Support Vector Machine that is designed specifically for classification tasks. It looks for the hyperplane that best divides the data points into several classes. A space is divided into two subspaces by the decision boundary: one is for vectors that are members of the group, and the other is for vectors that are not.
- Random Forest[19] creates several decision tree classifiers and assigns the class that the majority of the trees select to a document. The algorithm, given a training set of N samples, first creates many randomly selected subsets of N samples, some of which appear more than once or not at all.

4. DATA ANALYSIS

Our proposed system is implemented on the dataset available on www.kaggle.com. This dataset is licensed and available for general usage. Data was collected in Excel file uploaded in Jupyter notebook and analysed with Python software. Table 2 shows brief description of the dataset. Dataset includes 4657 rows with each title corresponding to one of the 10 different genres. The dataset comprises 2 text and 1 categorical attributes. Unlike other collections, this dataset has a wider range of genres that is why we chose it. The data contains no missing values. Since the Index column offers no assistance in characterizing the data, we remove it. To suit this study's requirements, the dataset has undergone further pre-processing, nevertheless.

We parsed the summary using the Natural Language Toolkit (NLTK) to determine how frequently each of these words appeared in the text. We then normalized the results to obtain a frequency value for each word. We also opted to include average sentence length, word count, and average word length in addition to the word count approach after conducting some early study. We believed that by doing this, our model's classification accuracy would improve.

Table 2. Data file description

Data source	Number of instances	Number of feature	Number of target classes	Size of data	File type
www.kaggle.com	4657	4	10	4557 KB	Excel spread sheet

4.1 Exploratory Data Analysis

There are 10 distinct genres, including science fiction, history, horror, romance, psychology, crime, fantasy sports, and travel. Table 3 displays each genre's frequency. Table 3 shows that the category "thriller" makes up 22 % of the dataset, whereas the categories "sports," "psychology," and "travel" each make up

Table 3. Genre wise statistics

Genre	Count	Title length (in words)			Summary length (in words)		
		Max	Mean	Min	Max	Mean	Min
crime	500	9	3.12	1	5475	369.18	7
fantasy	876	13	3.13	1	3086	384.20	2
history	600	24	3.13	1	5089	473.30	2
horror	600	9	2.66	1	5664	429.47	2
psychology	100	26	9.46	2	480	173.08	33
romance	111	7	3.12	1	300	165.70	52
science	647	23	3.97	1	3049	374.27	4
sports	100	24	3.77	1	474	180.03	53
thriller	1023	9	2.80	1	3765	290.01	12
travel	100	28	7.00	1	448	158.57	21

2 %. Our top three genres are “science”, “fantasy,” and “thriller.” We decided to prioritize other studies because the ratio of the largest class (1023 books) to the smallest class (100 books) is approximately 1:10, indicating that the degree of imbalance is not significant enough. We discover a relationship between the target label in the data and the length of the text that is available for more data research. Thus we extract the number of characters and words in summary and title of the book for the same.

Here we define the basic terms such as Title Length, Title Character length, Summary Length and Summary Character Length. Title Length is the number of words in the book’s title. Title Character Length is the total number of characters (including spaces) in the book’s title. Summary Length is the number of words in the book’s summary and Summary Character Length is the total number of characters (including spaces) in the book’s summary.

Table 3 shows that while horror books have the longest summaries, psychology and travel books typically have the longest titles. In this study, we discover a relationship between the target label in the data and the length of the text that is provided. As a result, we derive the book’s title, character and word counts from the summary. In this article, the distribution of observations in a dataset is visualized in Fig. 2 using a Kernel Density Estimate (KDE) plot.

It is clear from Fig. 2 that there is no recognizable relation between the genre of the book and length of the summary and title.

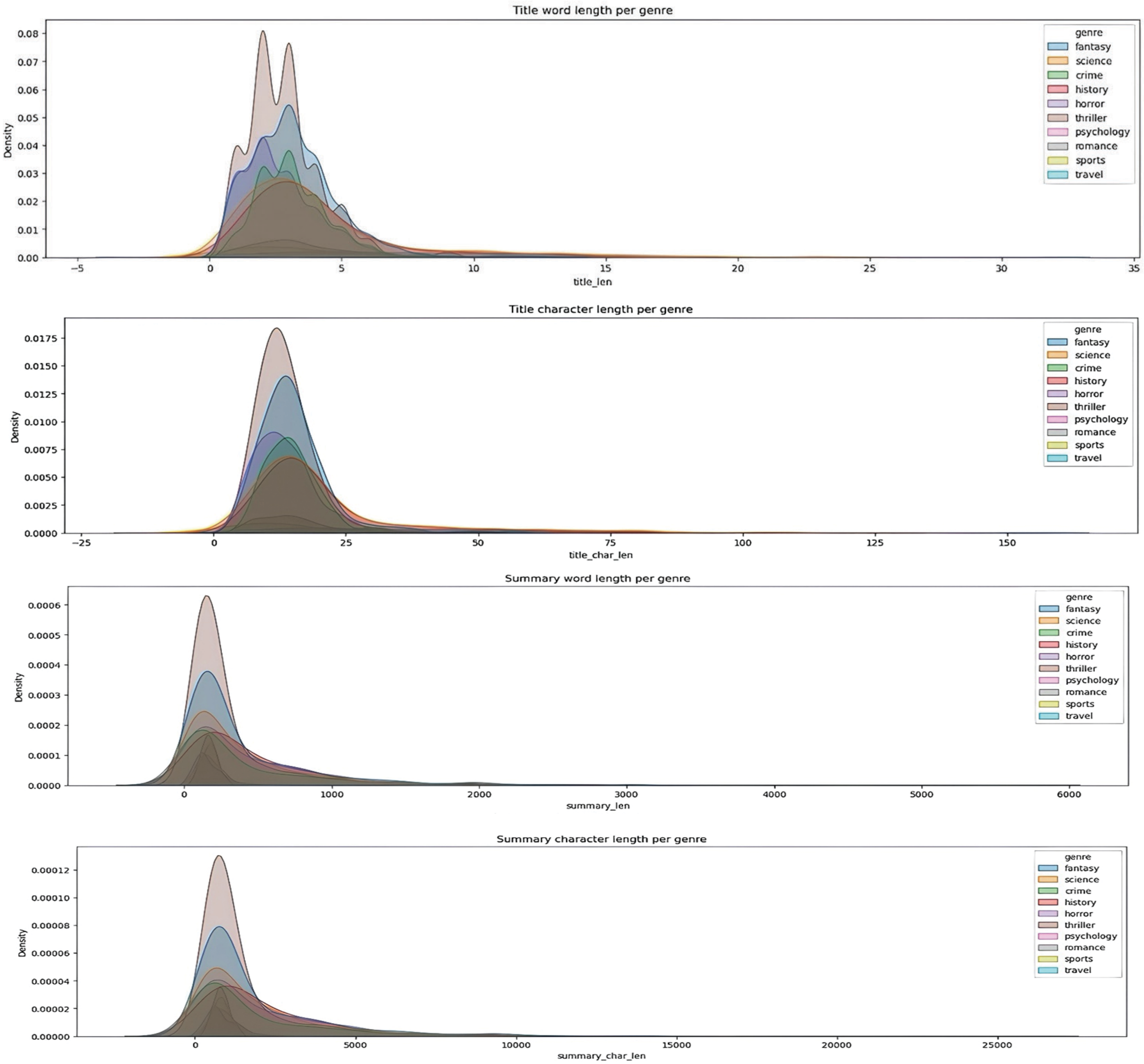

Figure 2. KDE plot between summary and title.

Table 3 shows that while horror books have the longest summaries, psychology and travel books typically have the longest titles. In this study, we discover a relationship between the target label in the data and the length of the text that is provided. As a result, we derive the book's title and character and word counts from the summary. In this article, the distribution of observations in a dataset is visualized in Fig. 2 using a Kernel Density Estimate (KDE) plot.

It is clear from Fig. 2 that there is no recognizable relation between the genre of the book and length of the summary and title.

5. RESULT DISCUSSION

5.1 Experimental Settings

In this study Python programming language is used to implement the proposed technique. Python packages such as scikit-learn, matplotlib, pandas and numpy are used for data analysis. Models were run using Jupyter notebook environment with a processor of Intel Core i5-8300H CPU @ 2:30 GHz and RAM of 8 GB running on Windows 10.

5.2 Experimental Results

Book genre classification is a multiclass classification problem, in which instances are classified into one of many potential classes. Our approach, known as One Vs Rest Classifier, involves fitting a single classifier for each class. Every class is fitted against every other class for every classifier. The main advantage of this method is computational efficiency. Since there is only one classifier per class, it is feasible to learn more about a class by looking at the classifier that corresponds to it. This is a reasonable default option and the most widely employed tactic. We have also compared execution time with One Vs One Classifier approach. Unlike One Vs Rest that splits it into one binary dataset for each class, the One Vs One approach splits the dataset into one dataset for each class versus every other class. Table 4 shows the values related to the confusion matrix, which is a performance measure of the techniques with 80 % percentile splits. Table 5 indicates execution time of the models.

5.3 Performance Evaluation Metrics

The actual and predicted classification done by a classification matrix[20] is usually generated and represented by a confusion matrix. A confusion matrix is a table that summarizes the performance of a classification model by comparing its predicted labels to the true labels. Let's understand the segments of the confusion matrix below.

- True Positives (TP): It means the actual value and also the predicted values are the same.
- False Positives (FP): This means the actual value is negative.
- True Negatives (TN): It means the actual value and also the predicted values are the same.
- False Negatives (FN): This means the actual value is positive.

In this article classification tasks involves ten genre classes and thus known by the name of "multiclass classification". Let's understand

1. Precision (P): How many of the books designated as positive are actually positive i.e. $P=\frac{TP}{TP+FP}$
2. Recall (R): What percentage of real true positive books is anticipated to be true positive books i.e. $R=\frac{TP}{TP+FN}$
3. Accuracy (A): How many predictions have the classifier made right i.e. $$A=\frac{TP+TN}{TP+TN+FP+FN}$$
4. F1 score: It assigns equal weight to precision and recall by measuring their harmonic means. F1 score lies between 0 and 1. When F1 score is equal to 1, it means all classes were correctly predicted, this is a very hard score to obtain with real data. When the model's F1 score is high, it means that it is performing well in both precision and recall; when it is low, it means that it is not performing well in any of these areas. $$\text{F1 score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Tabel 4. Metrics of performance for models

Model	Accuracy	Precision	Recall	F1 score
MultinomialNB	65.88	0.454	0.533	0.461
Logistic regression	67.06	**0.638**	**0.705**	**0.660**
Random forest	**67.81**	0.508	0.657	0.541
SVC	65.12	0.589	0.701	0.609

Table 5. Execution performance for models

Model	One Vs rest classifier time (sec)	One Vs one classifier time (sec)
MultinomialNB	**0.202**	**0.282**
Logistic regression	6.882	16.781
Random forest	69.602	70.465
SVC	74.542	75.238

The efficiency and effectiveness of the models must be improved by increasing precision, recall, and F1 score. Better performance would be indicated by a higher accuracy. It is crucial to assess performance evaluation in addition to accuracy while describing machine learning methods. Specifically, the execution time is a crucial performance metric. While the accuracy of all models is nearly same, there are notable differences in other aspects. With an accuracy of 67.81 %, the Random Forest model performs best on this dataset; but, because it generates and computes several decision trees, its execution time is noticeably longer (69.602). While Multinomial NB takes the least amount of time, it has relatively low Precision, Recall, and F1 score metrics. Additionally, the model SVC offers the lowest accuracy (65.12 %), whereas the logistic regression model yields higher performance and more accurate outcomes. One more interesting thing is noticed that execution time for all models is increasing for One Vs One approach. Since this approach trains more classes it is usually slower than One-vs-Rest. We have also checked the common words occurring in the summary depending on genre of book. Final observation is listed in Table 6 and shows there are some words which belong to both the genres.

Table 6. The most common words occurring for each genre

Genre	Common words	Genre	Common words
Thriller	Life, Find, Less	Crime	Murder, Death
Fantasy	Take Find, World, King	Romance	Life, Love, Less
Science	World, Time, Human	Psychology	Life, Book, People
History	Take, Father, Become	Sports	Life, Team, Less
Horror	Kill, Find, Vampire	Travel	Life, World, Travel, Journey

6. CONCLUSIONS

In this article, four machine learning methods are implemented for categorising different book genres from their titles and summaries. 80 % of the samples are used to train the models, and the remaining 20 % are used for testing. The performance of the techniques is compared based on F1 score, accuracy, precision, and recall. We have also compared execution time between One Vs Rest and One Vs One classifier. The best performing model is the logistic regression model. This approach is quite helpful for e-books as well, since it makes book recommendations depending on genre. The classifications of literary works used by digital libraries and archives, which frequently deceive readers and users, can also benefit from this study. Future work of this study could incorporate other models with hybrid techniques and various parameter adjustments. Furthermore, the models can be applied with innovative approaches to parameter optimization in order to illustrate the efficient information finding. Local rural language books which are hard to classify can be classified with further researches.

REFERENCES

1. Hassan, S.U.; Ahamed, J. & Ahmad, K. Analytics of machine learning-based algorithms for text classification. *Sustainable Operations and Comput.*, 2022, (3), 238-248, doi: 10.1016/j.susoc.2022.03.001
2. Hassani, H.; Beneki, C.; Unger, S.; Mazinani, M.T. & Yeganegi, M.R. Text mining in big data analytics. *Big Data and Cognitive Computing*, 2020, **4**(1). doi: 10.3390/bdcc4010001
3. Gupta, S.; Agarwal, M. & Jain, S. Automated genre classification of books using machine learning and natural language processing. 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2019, pp. 269-272 doi: 10.1109/CONFLUENCE.2019.8776935
4. Zhang, W.; Liu, F.; Zhang, Z.; Liu, S. & Huang, Q. Commodity text classification based E-commerce category and attribute mining. IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Shenzhen, China, 2020, pp. 105-108. doi: 10.1109/MIPR49039.2020.00028
5. Eisenstein, J. Introduction to natural language processing, ISBN: 9780262042840, MIT Press, 2019
6. Scofield, C.; Silva, M.O.; de Melo-Gomes, L. & Moro, M.M. Book genre classification based on reviews of portuguese-language literature. In: Pinheiro, V., *et al.* Computational processing of the portuguese language. PROPOR 2022. *Lecture Notes in Comput. Sci.,* vol 13208. Springer, Cham. doi: 10.1007/978-3-030-98305-5_18
7. Buczkowski, P.; Sobkowicz, A. & Kozlowski, M. Deep learning approaches towards book covers classification. *In* Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods-ICPRAM; ISBN 978-989-758-276-9; ISSN 2184-4313, *Sci. TePress,* 2018, pp. 309-316. doi: 10.5220/0006556103090316
8. Ozsarfati, E.; Sahin, E.; Saul, C.J. & Yilmaz, A. Book genre classification based on titles with comparative machine learning algorithms. IEEE *In* 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 2019, pp. 14-20, doi: 10.1109/CCOMS.2019.8821643
9. Biradar, G.R.; JM, R.; Varier A. & Sudhir, M. Classification of book genres using book cover and title. *IEEE* International Conference on Intelligent Systems and Green Technology (ICISGT), Visakhapatnam, India, 2019, pp. 72-723. doi: 10.1109/ICISGT44072.2019.00031
10. Ostendorff, M.; Bourgonje, P.; Berger, M.; Schneider, J.M.; Rehm, G. & Gipp, B. Enriching BERT with knowledge graph embeddings for document classification. 2019, doi: 10.48550/arXiv.1909.08402
11. Panchal, Y., Book genre categorisation using machine learning algorithms (K-nearest neighbor, support vector machine and logistic regression) using customised dataset, 2021, available at SSRN: doi: https://ssrn.com/abstract=3805945
12. Krishnan, A., Antony A., Jayachandran, D. & Shoba T, A survey on book genre prediction methods from summary, *IJIRE*-V3I03-250-253, 250-253.
13. Rasheed, A.; Umar, A.I.; Shirazi, S.H. *et al.* Cover-based multiple book genre recognition using an improved multimodal network. IJDAR, 2023, **26**, 65–88.

doi: 10.1007/s10032-022-00413-8

14. Zaki, M.J. & Meira, Jr. W. Data mining and machine learning: Fundamental concepts and algorithms. 2nd edn. Cambridge University Press, London, 2020, ISBN: 978-1108473989.
15. Bird, S.; Klein, E. & Loper, E. Natural language processing with python. O'Reilly Media, Inc. 2009, ISBN: 9780596516499
16. Xu, S.; Li, Y. & Wang, Z. Bayesian multinomial naïve bayes classifier to text classification. In: Park, J.; Chen, SC. Raymond Choo, KK. (eds) Advanced multimedia and ubiquitous engineering. Future tech MUE, 2017. *Lecture Notes in Electrical Engineering*, **448**. Springer, Singapore. doi: 10.1007/978-981-10-5041-1_57
17. Kleinbaum, D.G. & Klein, M. Logistic regression: A self-learning text. Springer New York. 2010, ISBN: 9781493936977
18. Zhang, Y. Support vector machine classification algorithm and its application. In: Liu, C.; Wang, L.; Yang, A. (eds) Information computing and applications. ICICA 2012. Communications in computer and information science, **308**. Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-34041-3_27
19. Genuer, R. & Poggi, J.M. Random forests with R. Springer nature switzerland AG 2020, ISBN : 978-3-030-56484-1, doi: 10.1007/978-3-030-56485-8
20. Deng, F.; Huang, J. & Yuan, X. Performance and efficiency of machine learning algorithms for analysing rectangular biomedical data. Lab Invest 101, 2021, 430–441. doi: 10.1038/s41374-020-00525-x

CONTRIBUTOR

Dr. Jhimli Adhikari received Master of Computer Application and PhD in Computer Science from Jadavpur University, Kolkata and Goa University respectively. At present she is working as Professor in the Department of Computer Science, Narayan Zantye College, Goa, India. Her areas of research interest include Data mining, Machine learning and Artificial intelligence. She conceptualised the study, designed the methodology, conducted the experiments, performed the analysis, interpreted the results, and prepared the manuscript.