National College of Ireland

Opinion Mining From Book Reviews Using Sentiment Analysis and Topic Modelling

MSc Research Project
Data Analytics

Balaji Pari
Student ID: 21217394

School of Computing
National College of Ireland

Supervisor: Abubakr Siddig

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Balaji Pari
Student ID:	21217394
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Abubakr Siddig
Submission Due Date:	18/09/2023
Project Title:	Opinion Mining From Book Reviews Using Sentiment Analysis and Topic Modelling
Word Count:	XXX
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	18th September 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Opinion Mining From Book Reviews Using Sentiment Analysis and Topic Modelling

Balaji Pari
21217394

Abstract

Books are considered as the bank of knowledge and learning, and they are emerging as an important habit among people. As interest among the people increases, it is important for the author and the publisher to understand the preferences of readers for different genres. A novel approach is introduced to extract the opinions from the reviews provided by the readers using topic modelling. The methodology involves sentiment analysis to predict the rating for each review using classification machine learning algorithms like Decision Tree, Naive Bayes and Random Forest. After rating prediction, reviews are labelled as positive and negative reviews based on the ratings predicted and Topic Modelling is performed using LDA (Latent Dirichlet Allocation) to extract the opinions for positive and negative reviews for different genre so that it helps the authors and publishers to understand about the opinions of readers for different genres. It works as an application framework to know about the opinions of the readers about different genres of books.

1 Introduction

1.1 Background and Motivation

In an era of growing digital world, Reviews from people on online platforms about hotels, movies, tourist places, applications, games etc. provides a valuable insights and sentiments which helps the business to understand about their flaws and to take necessary actions to grow their business. In extracting the opinions from the reviews machine learning plays a major role to take business driven solutions. Various business platforms who considers customer reviews for their decision making incorporate machine learning approaches to extract the opinions from the reviews and have proven to be successful in recent years.

Other platforms like movies, hotel and restaurants have applications that can help to know about the opinions but when it comes books, author and publishers there is no such application frame work and this proposed methodology will help as application framework to know about the opinions of the readers about different genres.

1.2 Research Objectives

- To predict the sentiment of book reviews
- To Implement the topic modelling using LDA to extract the positive and negative opinions based on different genre
- To visualize the topics contribution on review using cosine similarity

1.3 Research Question

- How can topic modeling and sentiment analysis be effectively combined to extract reader opinions from book reviews and understand about readers preferences?

The following sections covers about the related works based on this project, methodology that involves data preprocessing, Feature extraction, Model building, Hyperparameter tuning and Opinion mining followed by Design specifications, Implementation and Evaluation and finally Conclusion and Future work.

2 Related Work

This section explains about the similar works that has been done based on natural language processing, sentiment analysis and topic modelling. By investigating these papers paved the way for a efficient approach that combines sentiment analysis, topic modeling, and machine learning algorithms to extract opinions from book reviews and understand the readers' preferences and expectations.

2.1 Impact of Vectorization and Classification Machine Learning Algorithms

Machine learning algorithms are widely used for classifying the sentiments and proves to be successful in sentiment predictions with better accuracy and model performance. In this paper, (Badarneh et al.; 2023) classification machine learning algorithms were used to predict the sentiments from tweets. As a part of preprocessing step stemming is used to retain the original word from its different form. In some cases, lemmatization works better because there is possibility for stemming to change the meaning of the base word by removing the characters at the end. The results obtained that the classification models are efficiently predicting the sentiments with much better accuracy. Similarly Asha et al. (2023) used RNN to predict the people sentiment about covid-19 in India and Data is scrapped from twitter and both stemming and lemmetization was performed. NLP techniques such as named entity recognition and part of speech tagging also used in this research. Vectorization is an important feature extraction technique that is used to convert the text to numerical representation.Kumar et al. (2023) used TF-IDF and Bag-of-Words vectorization methods as feature extraction techniques. Classification machine learning algorithms were used to predict the sentiment of the product reviews from amazon. The research shows that performance of the models were drastically increased with much better accuracy and Kumar et al. (2023) also states that classification machine learning algorithms along with the feature extraction techniques works well for the prediction of sentiment. Similarly Styawati et al. (2022) used support vector machine to

understand about the attitude of the public towards online transportation and word2vec text embedding model as feature extraction technique to convert the words into vectors and the architecture of the word2vec is based on the skip gram model. Patel and Meehan (2021) also proposed a comparative study about TF-IDF and count vectorization to identify the fake news on reddit using classification algorithms. It states that data decides the best suitable vectorization techniques for the algorithm. There are many kinds of data that sentiment analysis can be done and one among them is social media data which are a complex data and as it involves lot of preprocessing steps based on the complexity. According to Satya et al. (2022) classification algorithms performs better even in social media data to predict the sentiment on data from Sestyc which is social media platform.

Hybrid model seems to be effective in classifying the sentiments based on the data. Kaushal and Chadha (2023) used hybrid model to predict the sentiments on what's app data which is a complicated data as it involves very deep language, slang and emojis. Data preprocessing would be a challenge in this type of data and use of hybrid model elevated the performance of model. Vectorization was used in this study and hybrid model was implemented using KNN, SVM and Decision tree. Hashing vector technique was used to create the sparse matrix for the TF-IDF and Bag-Of-Words vectors and reason for using the hashing vector technique is that there is no need for storing the dictionary in memory as it less scalable in large data set. Use of hybrid model resulted in accurate classification of sentiments than using them as individual models because of the data complexity.

2.2 Significance of Opinion mining

Opinion mining techniques are widely considered as effective method to know about others' opinions in large corpus of reviews.Kastrati et al. (2023) employed topic modelling using LDA and BERTopic to learn what individuals believed about the increase in energy prices. The attitude of the people through tweets on energy cost increase is tracked by BERT. After categorizing the tweets into good, negative, and neutral tweets, LDA is implemented to determine the underlying meaning of the tweets and it seems to an effective opinion mining technique by combining sentiment analysis and topic modelling. During the pandemic time OTT platform subscriptions significantly increased and partially replaced the conventional entertainment platform. Yawalkar et al. (2022) seeks to understand subscriber preferences and content consumption habits across OTT platforms. In this study, sentiment analysis with Liu Hu, topic modelling with LDA, and thematic analysis were all used. Following the pre-processing of the data, Topic modelling is carried out to comprehend the subtopics of the comments, and Liu-Hu sentiment analysis is used to classify the comments into positive, negative, and neutral ones. Finally, thematic analysis is done to understand the topics that are most prevalent in the comments. Opinions can be extracted in many different approaches and Yawalkar et al. (2022) proposed a method that uses machine learning algorithms to extract the opinions from amazon product reviews that classifies the sentiment of the text by predicting the sentiment score.

LDA is considered as a go-to algorithm for topic modelling as it is very effective in extracting the key term from large corpus of unstructured data. There is an evaluation

metric called coherence score which helps to choose the number of topics based on the data. Paul et al. (2022) used LDA algorithm to extract topics from Bangle news corpus and coherence score was used as an evaluation metric to determine the number of topics and this proposal suggest that high coherence value works better for large data corpus. As LDA seems to have an advantage on extracting sentiments from unstructured data, Ishmael et al. (2023) used it effectively to extract sentiment of the students on their assessment feedback for their module. The proposed methodology involves extraction of sentiments based on the sentiment score of the keywords under each topic. As LDA algorithm assigns keywords for each topic based on probability scores which seems to be effective on applying it to real time data.

2.3 Joint analysis using Sentiment Analysis and Topic Modelling

Sentiment analysis along with topic modelling on reviews is always effective to understand the audience opinions and Sindhu et al. (2021) proposed a joint sentiment topic modelling analysis to understand about the important aspects of a particular product. Each keyword under a topic have sentiment scores so that it explains how much weight age it has on the whole review corpus and it will be easy for the customer who is checking on reviews to buy that particular product. There is an additional layer added to this approach called credibility check that explains how close the reviewer who is reviewing the product has expertise or familiarity about the product, ensuring that their opinions and assessments are reliable. To achieve this, it uses occupation and interest of the reviewer along with their reviews. Table 1 shows the comparative analysis of key findings.

Table 1: Comparative Analysis of Relevant Studies

Study	**Key Findings**
(Badarneh et al.; 2023),(Asha et al.; 2023),(Ishmael et al.; 2023)	Improved sentiment prediction with different preprocessing techniques.
(Kumar et al.; 2023),(Styawati et al.; 2022),(Patel and Meehan; 2021),(Satya et al.; 2022)	Enhanced model performance in sentiment prediction using vectorization techniques.
(Kaushal and Chadha; 2023)	Hybrid model for complex data like WhatsApp chats.
(Kastrati et al.; 2023), (Yawalkar et al.; 2022)	Integration of sentiment analysis and topic modeling.
(Paul et al.; 2022), (Sindhu et al.; 2021)	Calculating the Coherence scores for topic modeling to find the relevant topics.

3 Methodology

Opinion mining can vary as positive and negative opinions from the readers and in this proposed methodology to extract the opinions from the readers based on different genre, Sentiment analysis is performed on the reviews given by the readers to predict the sentiment using classification mission learning algorithms. Once the sentiments are predicted, Topic modelling is performed based on different genre to extract the positive and negative

opinions expressed by the readers. This proposed methodology involves these following steps.

3.1 Data Collection

Data Collection Data is collected from Kaggle which is an open data set repository, and it is based on the amazon book reviews contains two data sets, one with information about the books and the other contains reviews and ratings about the books from the readers.

3.2 Data Preprocessing

There are two data sets, one data set with details of the book like name of the book, author name, published date and description of the book, etc.. and the other data set with review text, summary of the review, rating from readers, name of the book, price, etc.. The review data set has null values and duplicates so it is processed to remove the null values and duplicate values from the review and review summary columns and it has reviews from different years starting from 2000's and for this project reviews after the year 2020 are considered to know about the recent opinions of the readers. Once the data is filtered, it is then joined together using title column and unnecessary columns are removed.

Title	Description	User	Rating	Review Summary	Review	Author	Genre	publisher
The Gods of Mars	The Barsoom series continues: John Carter retu...	A1TE3EQMT442R2	1	Great story sorriest printing and binding Ive ...	The story , of course , is fabulous but this i...	['Edgar Rice Burroughs']	['Fiction']	Open Road Media
Swan	Laat je verrassen door New York Times-bestsell...	A8D3KS7II35N9	1	It was ghostwritten, duh!	Hello people. First of all, Naomi Campbell did...	['Frances Mayes']	['Fiction']	Lindhardt og Ringhof
Color me healthy	To combat the epidemic of weight gain, improve...	A2BN5BNK1HYGPG	1	RIP OFF, STOLEN WORKS	Don't waste your hard earned money on this. In...	['Daniel Dolgin Ph.D']	['Health & Fitness']	Page Publishing Inc
Color me healthy	To combat the epidemic of weight gain, improve...	AVMDFFEOOXZL	1	Reminds me of a joke...	What do you call the medical school student wh...	['Daniel Dolgin Ph.D']	['Health & Fitness']	Page Publishing Inc
Que Tal (Spanish Edition)	BESTSELLER #1 DEL NEW YORK TIMES Mark R. Levin...	A1H4I32GYHRCSU	1	What were they thinking?	This is the first review I have ever written, ...	['Mark R. Levin']	['Political Science']	Simon and Schuster

Figure 1: Overview of data set

Before performing sentiment analysis it is important to clean the review text and in sentiment analysis a set preprocessing steps are involved to clean the text properly before using that respective column to predict sentiment. Following steps are performed to clean the review text.

1. Converting upper case to lower case: There is a possibility to consider two occurrences of the same word different due to case sensitivity. So, all the reviews are converted into lower case.

2. Removing Special Characters: For Sentiment analysis and topic modelling it is important to remove the special characters. A function is used to remove special characters from review.

3. Removing Stop words: Stop words like is, it, the etc. are meaningless for a model and do not represent any sentiments and it is always better to remove them from corpus. A function is used to remove the stop words using English language model.

4. Lemmatization: Lemmatization helps to recover the word's base form to obtain the actual meaning of the word irrespective of how it used by the reviewer and helps to understand core semantics of the text. A function is used to lemmatize the words from the reviews.

5. Removing unwanted words: Topic modelling is all about extracting correlated key words in the form of topics and to extract the meaningful topics from the corpus commonly occurring words are removed and only nouns are processed in the model and function is used to remove the commonly occurring words.

Once the data is preprocessed, it is then split into test and train data to implement feature extraction techniques.

3.3 Feature extraction

Feature extraction is performed to convert the words into vectors as in numerical forms, so that it is easy for the machine learning models to understand and predict the ratings rather than directly giving the words to the model. In this project feature extraction is performed using count vectorization and TF-IDF vectorization methods.

- Vectorization : Initially the text is breakdown into individual words also known as tokens and then an unique vocabulary list is created using the distinct tokens in the corpus and each assigned an index value. A vector is created for each document equal to the size of the vocabulary and each position corresponds to the words in the vocabulary and value indicates the frequency. Finally, all the document merged together to form a matrix where rows represents the document and columns represents the word in vocabulary. The values indicates the number of occurrences of word in the entire corpus. Figure 2 shows an example for vectorization.

	about	bird	heard	is	the	word	you
About the bird, the bird, bird bird bird	1	5	0	0	2	0	0
You heard about the bird	1	1	1	0	1	0	1
The bird is the word	0	1	0	1	2	1	0

Figure 2: Vectorization

3.4 Model Building

Rating values are labeled as positive, negative and neutral and for the prediction of sentiment three classification machine learning models are used.

1. **Decision Tree:** Decision tree provides a clear and understandable way of decision making by ranking the features that determines the sentiment. It is important to understand the relationship between the words and phrases and decision tree helps

to capture the nonlinear relationship among the words and phrases effectively. It can handle numerical vectors well and understands which feature is more influential in predicting the sentiment.

2. **Random Forest:** Random forest combines multiple decision tree and predict the final sentiment by average or voting. It protect the model from over fitting and easily the understands the influential features in predicting the sentiment. There are possible chances for noise in a text even after the preprocessing steps and random forest effectively handle noisy data and provides a stable prediction because of its ensemble nature. It uses a technique called out-of-bag (OOB) that estimates then model performance without any additional validation set.

3. **Naive Bayes:**It is based on Bayes theorem which is works on the basis of probabilities and gives the observed evidence. Because of it's probability nature, it reduces the time consumption in training the model and it effectively handles the large text data in case of sentiment prediction. it helps to capture the important relationships and pattern in the text data which is important in sentiment analysis and effectively handle high dimensional data. Probabilities calculated by Bayes theorem provides insights on effectiveness of an individual word in predicting the sentiment which helps in the model decision. It supports incremental learning so that new data can be added at any time.

3.5 Hyperparameter Tuning

Hyperparameter Tuning is implemented on all the three models to optimize the performance and to avoid over fitting of models. Sometimes text data in sentiment analysis may have imbalance in way that a sentiment class can have more number of samples than the other classes which leads to biased model. Hyperparamter tuning can handles the data imbalance by adjusting parameters related to class weights, sampling techniques to improve the performance on minority classes. It increases model robustness to perform well on noisy data with right combination of parameters.

3.6 Opinion Mining

Once the sentiments are predicted, Topic modelling is implemented on the negative and positive reviews separately to understand about the underlying opinions of the readers about different genres. LDA (Latent Dirichlet Allocation) algorithm is used as topic modelling technique in this project.

LDA initially assigns the words randomly to different topics from the document and carry out this process repeatedly to adjust the words to the topics based on the probability and further it refines the probabilities of the words occurring under a topic by iteratively carry out the process until a stable state is reached where topics are discovered and words are assigned to the topic coherent manner. Figure 3 shows the simple visualization on how LDA works.

Figure 4 shows step by step visualization of methodology involved in this project. It starts with data preprocessing which involves text cleaning, stopword removal and lemmetization followed by feature extraction and then it leads to model building and it

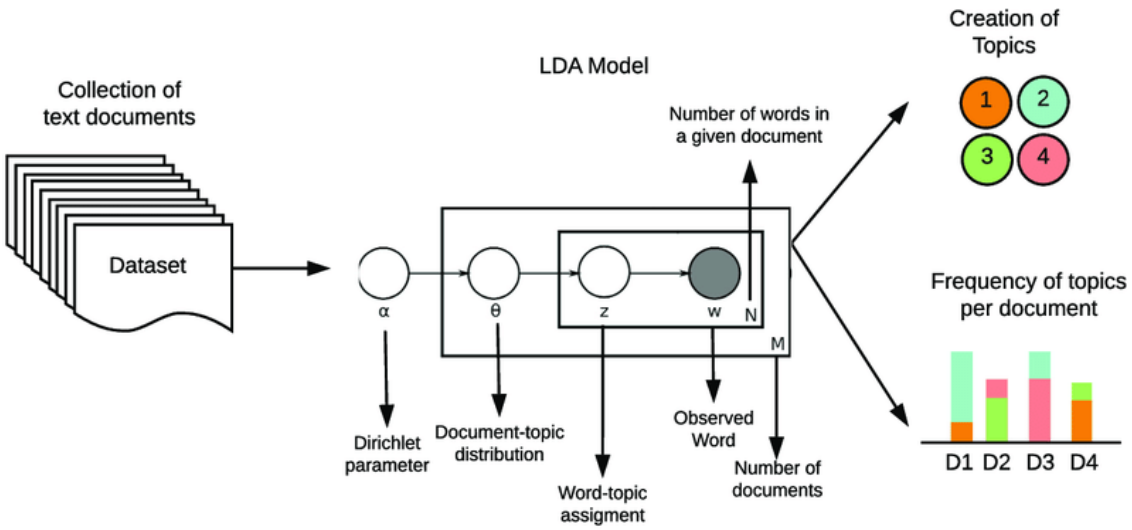

Figure 3: LDA

is optimized by doing hyperparameter tuning to find the best model for this data set and once the sentiments are predicted, LDA is implemented to extract the postitive and negative opinions from the reviews.

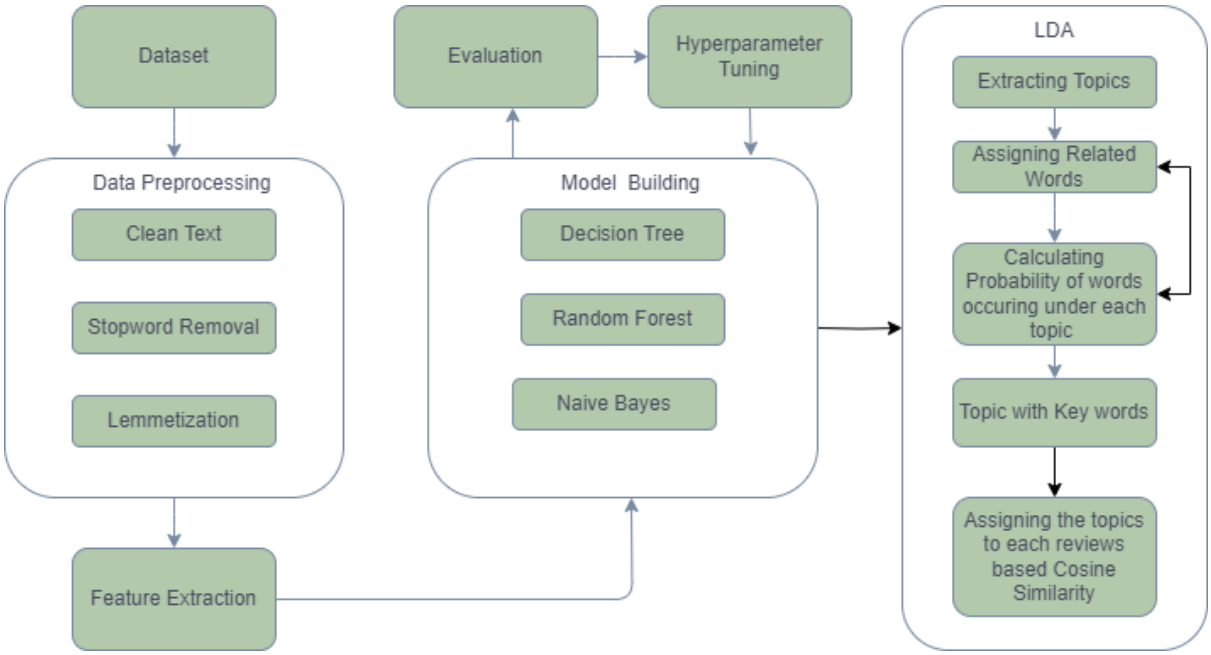

Figure 4: Methodology

3.7 Cosine Similarity

Cosine similarity is the process of calculating the similarity between two vectors and a similarity score will be generated at each iteration. This similarity score defines relationship between the two vectors. After the topics are generated with key words, a look up

table is created with a name for each topic based on the keywords present in it. This look up table is used to calculate the cosine similarity between each review and topic. Topic with high similarity score would be assigned to that particular review and this goes on for each review until all the reviews are assigned with topic.

The cosine similarity measures the similarity between two vectors, **A** and **B**, by calculating the cosine of the angle θ between them:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \cdot \|\mathbf{B}\|}$$

Where:

- **A** and **B** are the vectors to be compared.
- $\cdot$ denotes the dot product of the two vectors.
- $\|\mathbf{A}\|$ represents the Euclidean norm (magnitude) of vector **A**, and similarly for vector **B**.
- θ represents the angle between the vectors.

4 Design Specification

In this section design specification of this project involved in each phase of the project is explained clearly.

- Data downloaded and imported from kaggle using jupyter notebook
- Data Preprocessing involves Text cleaning, Stopword removal and Lemmetization
- Feature extraction using vectorization methods to represent the text in numerical format. Vectorization methods used are Count vectorization and TF-IDF Vectorization
- Model building to predict the sentiments. Models used are Decision Tree, Naive Bayes and Random Forest.
- Hyperparameter tuning to find the best model to predict the sentiment
- Implementing topic modelling using LDA to extract the topics and its related keywords on positive and negative reviews
- Assigning the topics for each review based on Cosine Similarity
- Visualizing the results to know about the contribution of each topic in reviews

Figure 5 represents the architectural design involved in each phase of this project and the generic steps involved in each phase.

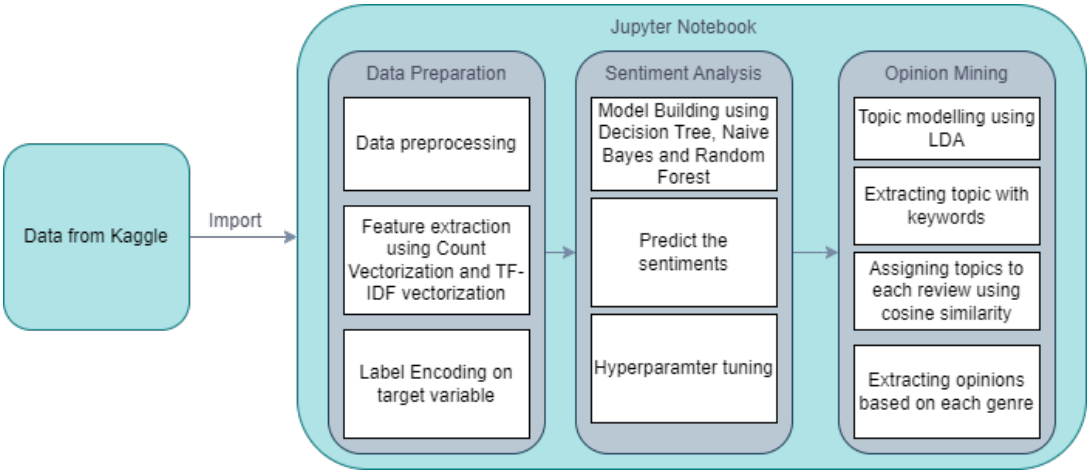

Figure 5: Implementation Design

5 Implementation

5.1 Data Preprocessing

Once data is prepared for preprocessing it is checked for the count of sentiment class for each rating range from 1 to 5. Figure 6 shows that rating are not biased towards particular rating and there is no imbalance in the data set

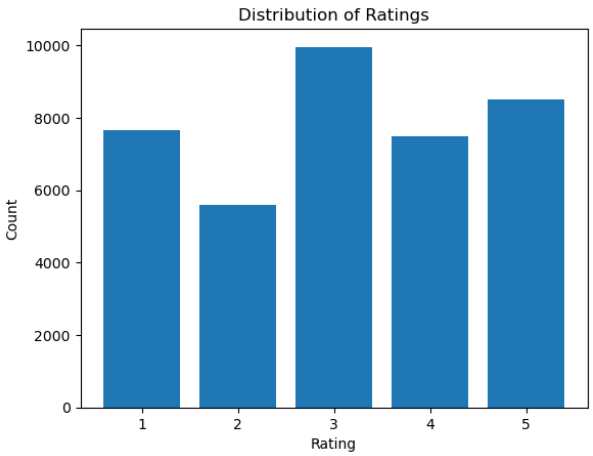

Figure 6: Distribution Of Rating Column

Data is then preprocessed to clean the text by removing the stopwords and noise then lemmatize the words to return it's base form so that same meaning could be obtained from two different occurrence of a same word. Figure 7 shows the exmaple for before and after the preprocessing of review column. It can be seen that review text is cleaned by removing the stopwords and the tense forms of the words are removed and returned with base word.

Review	Clean_Review
The story , of course , is fabulous but this i...	story course fabulous worst bind see year read...
Hello people. First of all, Naomi Campbell did...	hello people naomi campbell write book write a...
Don't waste your hard earned money on this. In...	waste hard earn money instead buy real thing I...
What do you call the medical school student wh...	medical school student graduate class doctor
This is the first review I have ever written, ...	review write say book agree reviewer book diso...
This is the course book assigned by my college...	course book assign college professor find book...
This is an incomplete 1902 translation of Prus...	incomplete translation prus classic historical...
This author knew very little about ancient Egy...	author know little ancient egypt people histor...
I bought this book because of the good review ...	buy book good review read preview kind confuse...
I purchased this book based on the praise give...	purchase book base praise give review pretty t...

Figure 7: Before and After Preprocessing of Review Column

5.2 Feature Extraction

Label encoding is performed on rating column and a new column is created called the Sentiment column and it has values 0, 1 and 2 for Negative,Neutral and Positive respectively. clean review text column which is an independent variable is assigned to a separate variable called 'X' and Sentiment which is dependent variable is assigned to a separate variable called 'Y'. Data is then split into test and train with eighty percentage of data in train and twenty percentage of data in test and feature extraction is implemented using count vectorization and TF-IDF techniques on the clean review column in order to find the best extraction technique for the proposed methodology. Figure 8 visualizes the distribution of sentiments after performing the label encoding and It shows that 0, 1 and 2 are the negative, neutral and positive sentiments respectively.

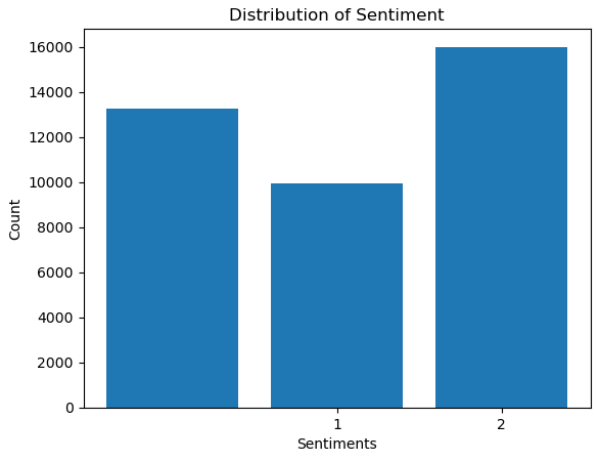

Figure 8: Label Encoding of Rating Column

5.3 Model Building

As discussed earlier in the methodology, three classification model are implemented to predict the sentiments namely Decision Tree, Naive Bayes and Random Forest. Figure 9, 10 and 11 shows the base models for all the three classification algorithms. It can be

seen that models have only minimum accuracy in predicting the sentiments.

```
Confusion Matrix for Decision Tree:
[[683 219 308 148 198]
 [272 191 338 161 154]
 [306 275 719 380 351]
 [165 145 433 382 382]
 [155 120 358 314 684]]
Score: 33.91
Classification Report:              precision    recall  f1-score   support

           1       0.43      0.44      0.44      1556
           2       0.20      0.17      0.18      1116
           3       0.33      0.35      0.34      2031
           4       0.28      0.25      0.26      1507
           5       0.39      0.42      0.40      1631

    accuracy                           0.34      7841
   macro avg       0.33      0.33      0.33      7841
weighted avg       0.33      0.34      0.34      7841
```

Figure 9: Base Model For Decision Tree Without Any Hyperparameter Tuning

```
Confusion Matrix for Multinomial Naive Bayes:
[[ 811   94  553   36   62]
 [ 189   84  744   58   41]
 [ 143   75 1428  195  190]
 [  63   12  595  467  370]
 [  51   16  320  167 1077]]
Score: 49.32
Classification Report:              precision    recall  f1-score   support

           1       0.65      0.52      0.58      1556
           2       0.30      0.08      0.12      1116
           3       0.39      0.70      0.50      2031
           4       0.51      0.31      0.38      1507
           5       0.62      0.66      0.64      1631

    accuracy                           0.49      7841
   macro avg       0.49      0.45      0.44      7841
weighted avg       0.50      0.49      0.47      7841
```

Figure 10: Base Model For Naive Bayes Without Any Hyperparameter Tuning

```
Confusion Matrix for Random Forest Classifier:
[[ 920   18  433   28  157]
 [ 298   32  627   37  122]
 [ 253   22 1259  132  365]
 [  79    5  629  244  550]
 [  78    6  301  104 1142]]
Score: 45.87
Classification Report:              precision    recall  f1-score   support

           1       0.57      0.59      0.58      1556
           2       0.39      0.03      0.05      1116
           3       0.39      0.62      0.48      2031
           4       0.45      0.16      0.24      1507
           5       0.49      0.70      0.58      1631

    accuracy                           0.46      7841
   macro avg       0.45      0.42      0.38      7841
weighted avg       0.46      0.46      0.41      7841
```

Figure 11: Base Model For Random forest Without Any Hyperparameter Tuning

5.4 Hyperparameter Tuning

Hyperparameter tuning is performed on Naive Bayes and Random Forest algorithm to improve the model accuracy and it seems like decision tree doesn't works well for this approach as the accuracy is not even close to 50 percentage so, no hyper parameter

tuning is performed on decision tree algorithm. Table 2 shows the Mean Test Score for each hyperparameter values of Naive Bayes model and it can be seen that for alpha value 1.0 the model have the highest Mean Test Score. Similarly table 3 shows the the accuracy for each hyperparameter values of Random Forest model it can be seen that for estimator value 300 the model have the highest accuracy.

Hyperparameter (alpha)	Mean Test Score
0.1	0.636578
0.5	0.652743
1.0	0.655549
1.5	0.649714
2.0	0.642158

Table 2: Hyperparameter Tuning Of Naive Bayes

Hyperparameter (n estimators)	Accuracy
100	0.636578
200	0.652743
300	0.655549

Table 3: Hyperparameter Tuning Of Random Forest

5.5 Over Sampling After Finalizing The Naive Bayes Model

From the Hyperparameter tuning, Naive Bayes model decided as the appropriate model for the proposed methodology. To further improve the accuracy of the Naive Bayes model over sampling using SMOT is implemented to balance the sentiments. Figure 8 shows that the distribution of sentiments after the label encoding seems to be slightly imbalance. Using SMOT over sampling is performed to balance the sentiments and again Hyperparameter tuning is performed to find the highest Mean Test Score. Table 4 shows that after over sampling alpha value of 0.1 have the highest Mean Test Score value.

Hyperparameter (alpha)	Mean Test Score
0.1	0.691172
0.5	0.691146
1.0	0.685156
1.5	0.678906
2.0	0.672474

Table 4: Hyperparameter Tuning Of Naive Bayes after Over Sampling

5.6 Topic Modelling Using LDA

Topic modelling is implemented using LDA after the sentiment prediction and Figure 12 shows that contribution of each genre books in the data set. Fiction genre contributes

more than 50 percentage.So, the base LDA model is implemented on fiction genre to extract the negative and positive opinions.

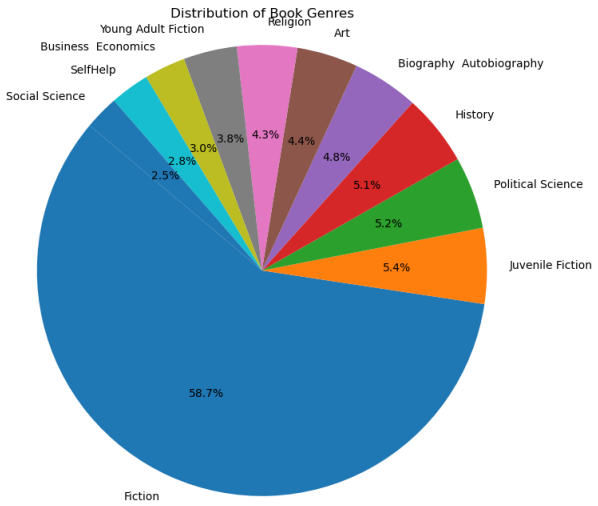

Figure 12: Contribution Of Genre

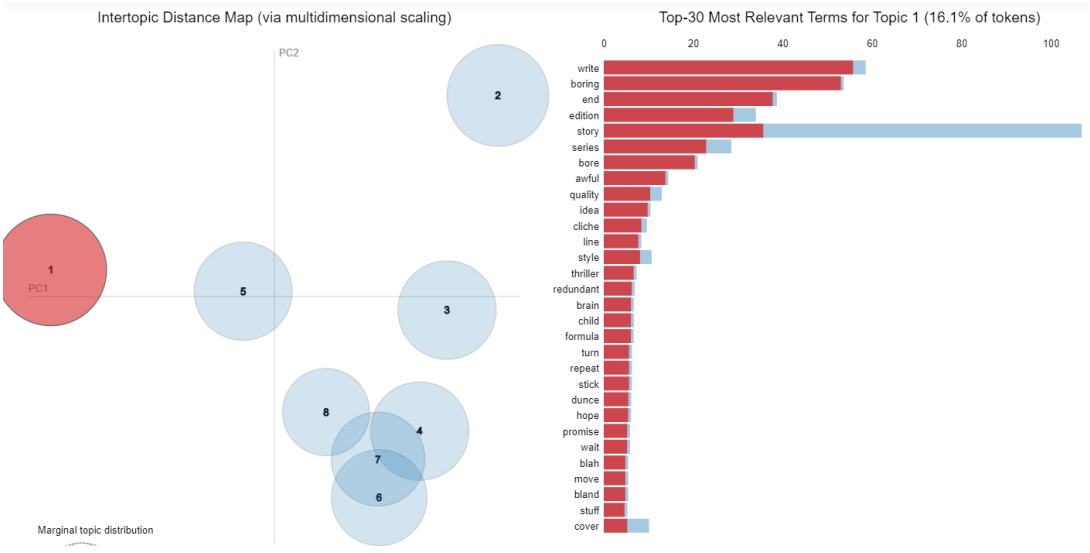

Figure 13: Base LDA Model For Fiction Genre

Figure 13 shows the LDAvis plot for base topic modelling for fiction genre on negative reviews. It can be seen that how frequently the key words are used by the readers to express their negative opinions. In this base model 8 topics are generated and under each topic the frequently occurring words are shown in the LDAvis plot and frequency of each word is also mentioned in rec colour and blue represents the overall frequency of that word.

Figure 14 shows the most frequent negative keywords that are expressed by the readers on fiction genre. Size of the words defines the frequency of the word occurring in the reviews.

Figure 14: Most Frequent Topic Keywords On Fiction Genre

5.7 Calculating the Cosine Similarity To Assign Topics For Each Review

After the topic extraction, a look up table is created given a name for each topic related to the keywords in it. This look up table is used to assign the topic for each review based on the cosine similarity scores. Topic with high similarity score would be assigned to that review.

5.8 Summary Of Implementation

Table 5 shows the finalized methods for the proposed methodology to predict the sentiment

Methods	Finalized approach
Preprocessing	Text Clean, Stop-words,Lemmetization
Feature Engineering	TF-IDF Vectorization
Model	MultiNomial Naive Bayes
Hyperparameter Value (alpha)	0.1
Topic Modelling (alpha)	LDA
Number of Topics (alpha)	10

Table 5: Finalized approach

6 Evaluation

The proposed methodology involves classification problems and the evaluation metrics for classification problems are confusion matrix, precision, recall and f1 score. The finalized model is Naive Bayes model with alpha value of 0.1.

- Confusion Matrix

  Figure 15 shows the confusion matrix which is an evaluation metrics for classification problems and it explains true positive, true negative, false positive and false negative for actual and predicted values. Using this metrics accuracy, precision, recall and F1 Score are calculated.

- Accuracy, Precision, Recall and F1 Score

  Figure 16 shows the classification report that explains the recall, precision , accuracy and F1 Score. It can be seen that the finalized model have an accuracy of 68.85 percentage.

- Actual vs Predicted

  Figure 17 shows the actual value vs predicted value that explains how close the model is predicted the sentiments compared to actual values.

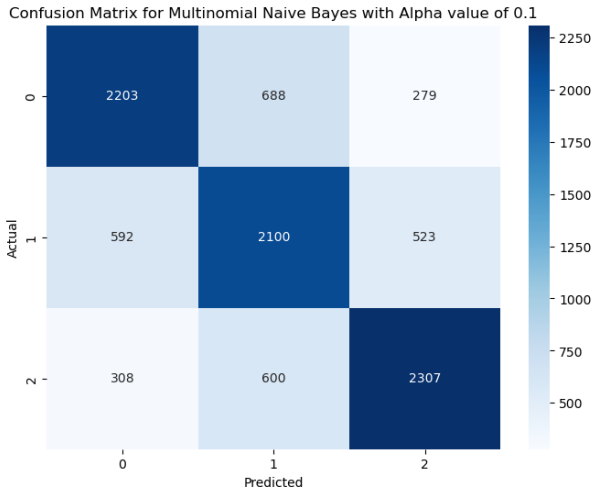

Figure 15: Confusion Matrix for Multinomial Naive Bayes

```
Confusion Matrix for Multinomial Naive Bayes:
[[2203  688  279]
 [ 592 2100  523]
 [ 308  600 2307]]
Score: 68.85
Classification Report:              precision    recall  f1-score   support

           0       0.71      0.69      0.70      3170
           1       0.62      0.65      0.64      3215
           2       0.74      0.72      0.73      3215

    accuracy                           0.69      9600
   macro avg       0.69      0.69      0.69      9600
weighted avg       0.69      0.69      0.69      9600
```

Figure 16: Classification Report for Multinomial Naive Bayes

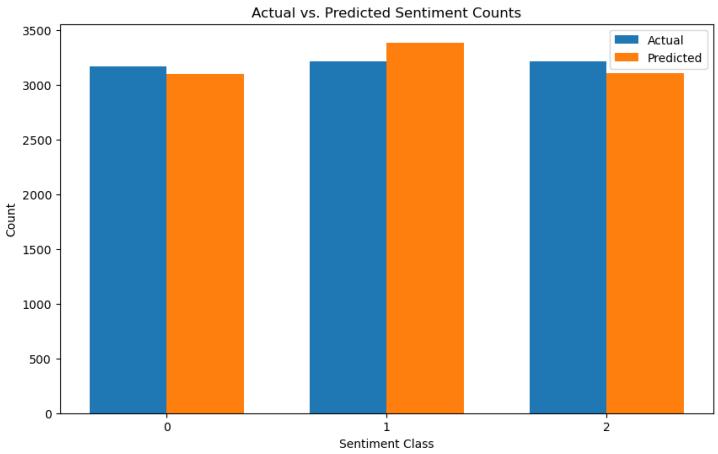

Figure 17: Classification Report for Multinomial Naive Bayes

- Coherence Score for Topic Modelling

  Table 5 shows the coherence score for each number of topics and for the fiction genre the number of topics choose is 10. Coherence score should too high and too low for LDA topic modelling as it leads to over fitting or under fitting.

Topics	Coherence Score
8	0.7294
10	0.7333
12	0.7460
14	0.7277

Table 6: Coherence Score For Number Of Topics LDA- Fiction Genre

6.1 Experiment 1 - Count Vectorization as feature extraction technique and ratings as target column

In Experiment 1, Count Vectorization technique is used as feature extraction technique with all the other general preprocessing steps. Rating column acted as target column and models are built to predict the rating. Turns out that the model are under performed with an accuracy of 49.32 for Naive Bayes algorithm, 33.91 for Decision Tree algorithm and 45.87 for Random Forest algorithm.

6.2 Experiment 2 - Count Vectorization as feature extraction technique and Sentiments as target column

In Experiment 2, Count Vectorization technique is used as feature extraction technique with the necessary preprocessing steps. Rating columns converted to sentiments as positive, negative and neutral based on their rating value and label encoding is performed to change it 0,1 and 2 and set as target column and models are built to predict the sentiments. The accuracy of the models increased with 66.22 for Naive Bayes algorithm, 48.46 for Decision Tree algorithm and 62.57 for Random Forest algorithm.

6.3 Experiment 3 - TF-IDF Vectorization as feature extraction technique and Sentiments as target column

In Experiment 3, TF-IDF Vectorization technique is used as feature extraction technique with the necessary preprocessing steps. Rating columns converted to sentiments as positive, negative and neutral and label encoding is performed and set as target column and models are built to predict the sentiments. The accuracy of the models decreased with 58.50 for Naive Bayes algorithm, 47.88 for Decision Tree algorithm and 62.47 for Random Forest algorithm.

6.4 Experiment 4 - Hyperparameter Tuning after finalizing the Naive Bayes Model

In Experiment 4, Hyperparameter tuning is performed for Naive Bayes with alpha value of 1.0 but there is no change in the accuracy of the model and Count vectorization as feature extraction technique and sentiment as target column

6.5 Experiment 5 - TF-IDF Vectorization as feature extraction technique and Sentiments as target column along with over sampling of data

In Experiment 5, TF-IDF Vectorization technique is used as feature extraction technique with the necessary preprocessing steps. Sentiments are used as as target column and models are built to predict the sentiments. Additionally over sampling is performed to balance the sentiments as there is slight imbalance in data when it is converted into sentiments from rating. SMOTE is used to perform the over sampling and then again Hyperparameter tuning is performed to find the best possible alpha value and it seems to be 0.1. The final accuracy for Naive Bayes algorithm with alpha value 0.1 is 68.85.

6.6 Discussion

The main aim of this application is to understand about the readers opinions about different genre from book reviews and to achieve that sentiment prediction is carried out using classification algorithms and LDA is implemented to understand about the preferences of readers based on different genres. The finalized model for the application is Naive Bayes with alpha value of 0.1. Two types of vectorization techniques are used and finally TF-IDF vectorization is chosen after performing several experimentation on vectorization. Sentiment prediction carried out in two different ways one with predicting the rating and other with predicting the sentiments and the models worked well in predicting the sentiments. Over sampling is also performed to balance the data set and in future methods data augmentation can works better to balance the data set. There are limitation for this application as the accuracy is 69 percentage and still works needs to be done in order to increase the accuracy but it can be useful to understand about the basic expectation of readers. With the help of this application, the author or the publishing companies may have an idea about some of the key areas that needs to be taken care while publishing or writing a book.

7 Results

After topics are assigned to each review using cosine similarity scores it is then visualized to see which opinions contributes the most among the readers for different genres. For example some visualization plots are shown in order to understand about the topics contribution. Figure 18 shows the contribution of topics on positive reviews for fiction genre and it can be seen that the lot of readers happy about the story,writing and inspiring them too. Figure 19 shows contribution of topics on negative reviews for fiction genre and it is seen that readers are not happy with the story, plot, content and they feel like characters are misleading to somewhere all these are the negative opinions of readers about fiction genre.Figure 20 shows contribution of topics on Positive reviews for History genre and it shows that readers are feel like influencing, insightful and good writing. Figure 21 shows contribution of topics on negative reviews for Juvenile Fiction genre. Many readers feel that the story is not good and some like its a waste of time and not engaging.

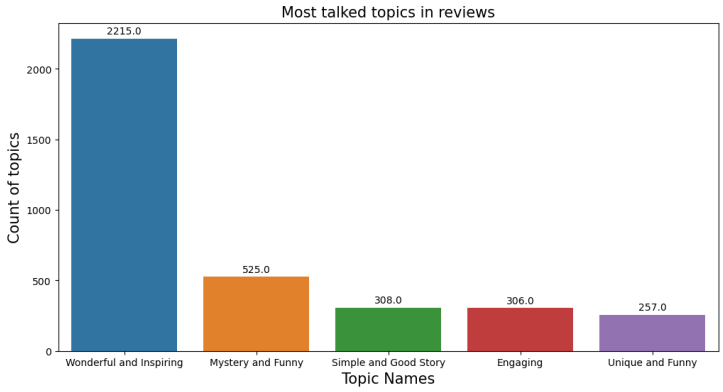

Figure 18: Contribution Of Topics for Positive Review - Fiction

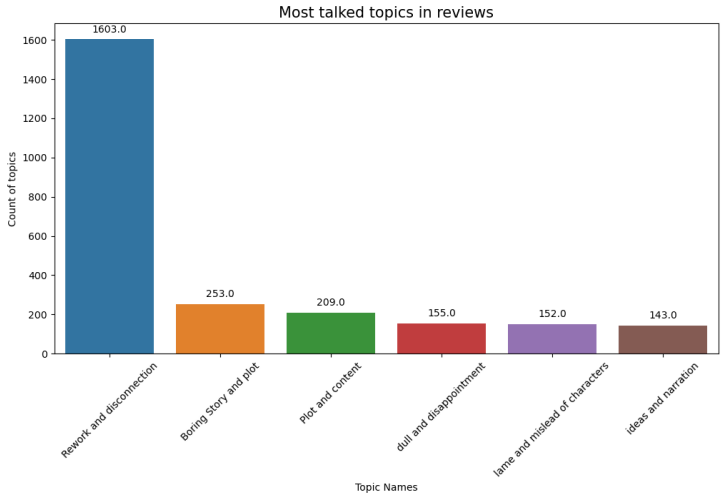

Figure 19: Contribution Of Topics for Negative Review - Fiction

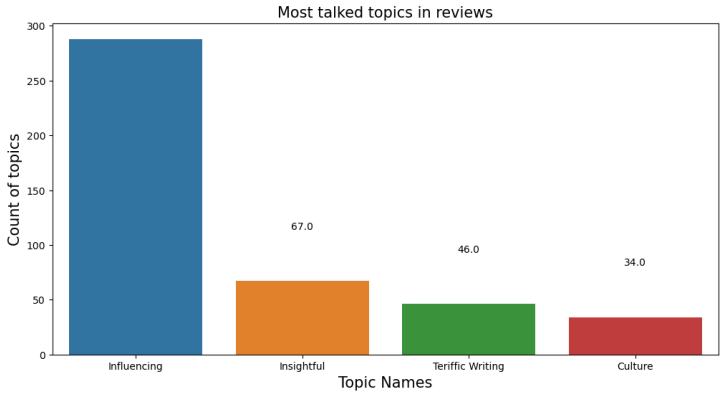

Figure 20: Contribution Of Topics for Positive Review - History

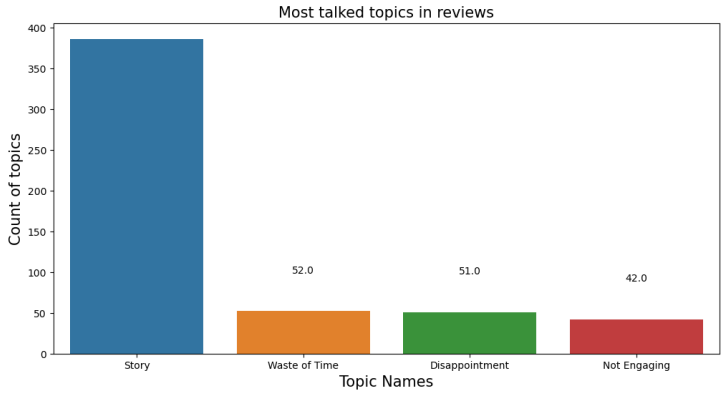

Figure 21: Contribution Of Topics for Negative Review - Juvenile Fiction

8 Conclusion and Future Work

In the application point of view this proposed approach will help the authors and publisher to know about their audience and some basic ideas about the expectations. There are several applications for other platforms like movies, restaurants, hotels to know about their audience and this proposed methodology helps to cover about books. In research point of view the findings are not up to the mark in case of predicting the sentiments as the proposed methodology have an maximum accuracy of 69 percentage by using Naive Bayes model because of the complexity in the data set and LDA worked well in extracting the opinions based on the sentiments. It may be a kick start for an application to be developed in future with some new additional methodologies to handle the data well. But still the authors and publishing companies can rely on this application to know some important highlights and outlines for different genres so that it can be fullfilled in future books.

As a part of the future work, instead of using classification machine learning algorithms for sentiment predication pre-trained model like BERT can be used as it is an pre-trained model that helps to understand the hidden pattern in review data and help in accurate prediction of sentiments. Use of other deep learning algorithms may also helps in this case due to the ability of understanding the hidden pattern in data. LDA performance can also be improved by trying to dig deep the topic extraction with minimum overlaps among the topics. Using deep learning algorithms for predicting the rating instead of sentiments a recommendation system can be built using the review text alone without help of any ratings.

References

Asha, V., Vishwanatha, C. R., Kumar, A., Shilpa, S., Shivaiah, G. E. and Kumar, A. (2023). Sentimental analysis of lockdown in india during covid-19, *2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC)*, pp. 1–6.

Badarneh, A. S., Al-Darwesh, S., Alzubi, O., Qassas, W. and ElBasheer, M. (2023). Sentiment analysis of tweets: A machine learning approach, *2023 IEEE Jordan In-*

ternational Joint Conference on Electrical Engineering and Information Technology (JEEIT), pp. 181–186.

Ishmael, O., Kiely, E., Quigley, C. and McGinty, D. (2023). Topic modelling using latent dirichlet allocation (lda) and analysis of students sentiments, *2023 20th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pp. 1–6.

Kastrati, Z., Imran, A. S., Daudpota, S. M., Memon, M. A. and Kastrati, M. (2023). Soaring energy prices: Understanding public engagement on twitter using sentiment analysis and topic modeling with transformers, *IEEE Access* **11**: 26541–26553.

Kaushal, R. and Chadha, R. (2023). Hybrid model for sentiment analysis of whatsapp data, *2023 3rd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*, pp. 215–220.

Kumar, A., Jain, T., Tiwari, P. and Sharma, R. (2023). Opinion mining on amazon musical product reviews using supervised machine learning techniques, *2023 11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON)*, pp. 1–6.

Patel, A. and Meehan, K. (2021). Fake news detection on reddit utilising countvectorizer and term frequency-inverse document frequency with logistic regression, multinominalnb and support vector machine, *2021 32nd Irish Signals and Systems Conference (ISSC)*, pp. 1–6.

Paul, P. C., Shihab Uddin, M., Ahmed, M. T., Moshiul Hoque, M. and Rahman, M. (2022). Semantic topic extraction from bangla news corpus using lda and bert-lda, *2022 25th International Conference on Computer and Information Technology (ICCIT)*, pp. 512–516.

Satya, B., S J, M. H., Rahardi, M. and Abdulloh, F. F. (2022). Sentiment analysis of review sestyc using support vector machine, naive bayes, and logistic regression algorithm, *2022 5th International Conference on Information and Communications Technology (ICOIACT)*, pp. 188–193.

Sindhu, C., Mukherjee, D. and Sonakshi (2021). A joint sentiment-topic model for product review analysis of electronic goods, *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 574–578.

Styawati, S., Nurkholis, A., Aldino, A. A., Samsugi, S., Suryati, E. and Cahyono, R. P. (2022). Sentiment analysis on online transportation reviews using word2vec text embedding model feature extraction and support vector machine (svm) algorithm, *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, pp. 163–167.

Yawalkar, P., Birari, A., Bharathan, G., Vakayil, S. and Sharma, R. (2022). Subscriber preference and content consumption pattern toward ott platform: An opinion mining, *2022 International Conference on Trends in Quantum Computing and Emerging Business Technologies (TQCEBT)*, pp. 1–7.