Exploring authors and readers in the digital age with AI and data: Analyzing book reviews

Abhishek Subedi
Universität Passau
subedi01@ads.uni-passau.de

Anusha Bora
Universität Passau
bora01@ads.uni-passau.de

Upender Shana Gonda
Universität Passau
shanag01@ads.uni-passau.de

Keerthi Sagar Reddy Aereddy
Universität Passau
aeredd01@ads.uni-passau.de

Abstract

Books are a great source of inspirations to people. Readers are searching for better books and organizations like online book stores, government organizations are curious in understanding users interest and opinions. Book reviews in the form of text and rating are a good source of informative data that can be used for research. Our work in this research is deeply focused in understanding book reviews through text reviews and star ratings given to the book by readers. We explore growth and decline, sentiments, relation between review-rating and prediction of ratings from review for two specific genres i.e. Comics & Graphics and Mystery, Thriller & Crime. Our research contributes in understanding the dynamics of book reviews that can be useful for academic institutions, writers, readers, online book stores, governments organizations.

CCS Concepts: • **Information System → Computational Social Science**.

Keywords: books review, analysis, goodreads

ACM Reference Format:
Abhishek Subedi, Anusha Bora, Upender Shana Gonda, and Keerthi Sagar Reddy Aereddy. 2023. Exploring authors and readers in the digital age with AI and data: Analyzing book reviews. In *Proceedings of Computational Social Science Lab.* ACM, New York, NY, USA, 12 pages.

1 Introduction

Online book stores are widely used by readers and sellers to have access to huge amount of books and customers. Goodreads, Amazon are the most popular online platforms, where number of registered users in between 2011-2019 in goodreads was 90 million[10] and 2.3 billion books by 2018[21]. These online platform gives users freedom of rating and writing reviews on books as per their reading experiences. Online ratings and reviews can help provide genuine insight, give greater visibility to books, help understand readers, boost author profile and in overall is a good means of communication in between readers and authors.

Books in these platforms are categorized in various genre e.g:. Children, Comics & Graphics, Mystery, Thriller & Crime, History & Biography, Poetry, Romance etc. All of these genre books have specific information like rating, reviews text, publication year, book id etc. that can be utilized by online companies, research organizations, academic institutions to explore human behaviour, interest and trends.

This research is focused on goodreads datasets, where two genre comics & graphics and mystery, thriller & crime are considered. Goodreads allows users to rate a book in between 5 (highest star rating) to 0 (lowest star rating), write reviews about the books and rate the reviews. This helps gain reputations for the books and writers. Research into this information can help figure out facts, figures, patterns, analyse interest of users etc. and discover knowledge and insight from data to understand the ongoing scenario that is essential to business and policy makers.

2 Research Objective

As discussed, understanding users, their emotions towards books, books trend are crucial to provide quality books to users. This research is focused on analysing and exploring four research questions related to book reviews based on two genre (comics & graphics and myster, thriller & crime). Research questions are mentioned as below.

2.1 RQ1: Is there a growth or decline of book genre?

2.2 RQ2: What text reviews says about genre: Are readers positive, negative or neutral towards genre?

2.3 RQ3: Did the books with more text reviews receive higher ratings?

2.4 RQ4: How to predict ratings from review text?

3 Related Work

There are some of the efforts[6, 11, 16, 17] faced by combining both collaborative and content-based filters when considering the advantages of the text review. In the previous work[17], for classification problems, the authors considered the content-based filtering, in which comparison of the text review with the respective ratings and the dense matrix was

Table 1. Metadata of Datasets

no.	columns
0	language_code
1	is_ebook
2	average_rating
3	publication_year
4	book_id

Table 2. Book Review Datasets

no.	columns
0	book_id
1	rating
2	reivew_text

generated by filling and collaborative filtering were applied. The author proposed the unified model technique[14] based on the content-based and collaborative filters for a better understanding of the insights of ratings and text reviews. The modeling was applied to the text review in order to adjust the rating dimension. Also, the information provided to the model to learn latent topics and can predict the cold start problem. The author was performed on 27 classes of datasets and then which leads to the proper comparison of the rating and the reviews[14].

Previous work of reviews and ratings impact the sales of books, movies, restaurants, video games[8]. These all previous works created a relationship between the economic characteristics of a particular company and review characteristics. But, in this paper, we focused on the interconnection between the customer ratings and review texts.

Other related works are mostly concentrated on identifying fake reviews. Multiple authors[5, 7, 9] have identified fake reviews in online review websites, Marketplaces. These previous works also addressed the various kinds of problems such as Data sparsity, Cold start, and noise[18].

4 Dataset & Pre-processing

4.1 Data Collection

Goodreads datasets acquired in 2017 [22] [23] in which comics & graphics and mystery, thriller & crime are two genre used in this research. All data that can be viewed publicly without login are scrapped and redistributed for academic purpose only[12]. We follow a Purposive data sampling strategy, where we gather data according to predefined and relevant criteria to the research question[15]. Our predefined criteria is to get only columns that are required to the research.

Table 1, shows metadata of the datasets, where *language_code* is used to retrieve data with English language, *average_rating* and *publication_year* are used for research analysis and *book_id* to get distinct list of books from the datasets. Table 2, contains *rating* and *review_text* for the books that are available in Table 1.

4.2 Pre-processing

Pre-processing is required such that datasets is ready to use for input to the algorithms and visualizations. Therefore the following steps are done to pre-process data.

- The metadata of datasets has null values in the columns *average_rating* or *publication_year*. Since this columns are important features, the rows that does not have this information are discarded.
- Books that are not in English are removed.
- In book review datasets, only those rows whose *book_id* are in metadata are taken and everything else are discarded.
- The columns having text values are converted into lower case.
- Punctuation's and non alphabetic characters are removed
- Multi-space are reduced to single space
- Finally the working datasets are exported to .csv file format.

4.3 One-hot encoding

Many Machine learning algorithms don't work properly with categorical data. To do work with this algorithms should convert the data into a suitable form. Precisely, One hot encoding is a pre-processing step in machine learning, which can convert the categorical data into a suitable form for input to the machine learning model. Mostly, one hot encoding is useful when the data is not balanced or dependent on the other attributes in the dataset. It can be used to create the dummy variables for the given data. When you create the dummy variables for the categorical data, the categorical data is balanced and that can feed into the machine learning model.

For example, when we are trying to predict the salary of four different positions in the company, we can use one-hot encoding. There are four different positions, Manager, Data Engineer, Data scientist, HR. When we are trying to predict the salary for these positions, we can convert this categorical data into suitable input form for the Machine learning model to erect the salaries of the positions. The one-hot encoding for these 4 four positions is given below[19].

$$salary = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Table 3. Datasets Statistics

Book Genre	Comics	Crime
Unique books	23937	66315
Is ebook	1886	17517
Is not ebook	22293	48798
Publication year	1986-2018	1946-2018
No. of reviews	330052	1056240
Avg. length of reviews	46.0	6.1
Avg of Avg. rating of books	3.91	3.87

4.4 TF-IDF

Term frequency-inverse document frequency is mostly used for the term weighting scheme in information retrieval systems. TF-IDF term consists of two terms Tf and IDF. Tf gives the probability of occurrence of a term and this term is normalised by the total frequency in the document. IDF is the amount of information, given as the log of the inverse probability[3].

NLP(Natural Language Processing) is dealing with human language understanding. In machine learning, NLP is using for so many applications such as chatbots, translation and candidate filtering, etc.

The machine learning algorithms don't work with the text directly, the text should be converted into vectors before feeding to the machine learning model. In NLP, the most commonly used approach is Bag of words(BoW). But, there are multiple drawbacks are there for BoW. In BoW, It will take every word that has the same preference and it does not account for noise also.So, to overcome the drawbacks of the Bag of words model will use TF-IDF.

TF-IDF gives frequency for every word, it will understand the text and removes all stop words. It weighs all the words and sorts the words according to their occurrence in the human language. The most frequent words sorted as less weight and less frequent or very rare words will get high weight in the document.

4.5 Preliminary Descriptive Analysis

Table 3, shows general statistics about the datasets. In comparison Mystery crime & thriller have two and half time more books than comics & graphics. Due to huge amount of text reviews average length in crime & thriller happened to be very low in comparison to comics & graphics. From Table 3 statistics we can see more readers tends to prefer hard copy of a book rather than ebooks. Figure 1 and Figure 2, shows the distribution of review length in words with respect to review count. In both the genre review with length between 0 to 500 are high number.

Books of various genres could be a determinant of popularity. Some genres are immensely popular, whereas others require specific preferences. Books receive ratings and reviews

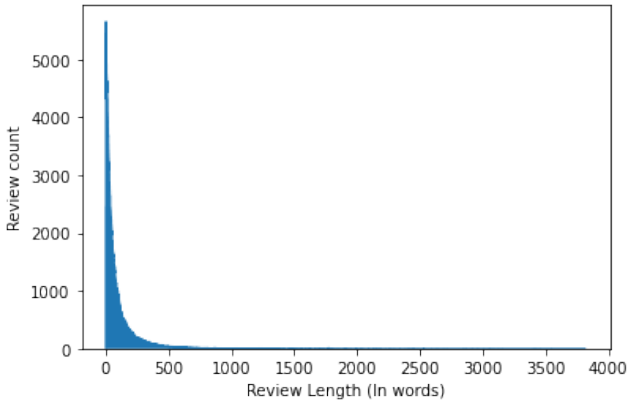

Figure 1. Distribution of review length for Comics & Graphics

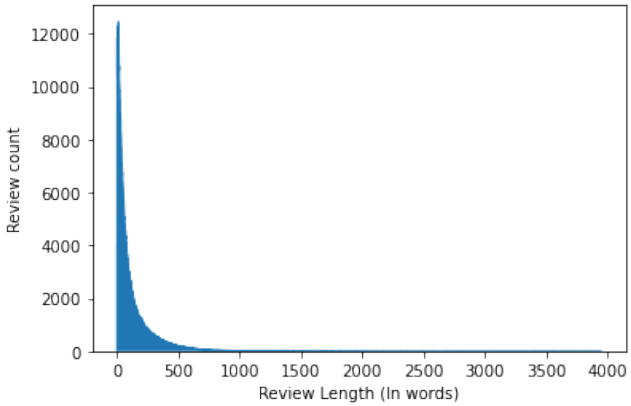

Figure 2. Distribution of review length for Mystery, Crime & Thriller

based on a variety of factors that reflect aspects of readers such as content, quality, book price, applicability, and other elements. In Figure 3 We estimated the average rating distribution of the Comics & Graphics genre and in Figure 4 the average rating distribution of the Mystery, Thriller & Crime genre. On average there does not exist a huge difference in rating for the genres. However, both genres receive the highest ratings in larger proportions i.e; Comics & Graphics genre have the highest number of books with ratings between 4 to 4.5, and Mystery, Thriller & Crime have ratings of 3.5 to 4.

In Figure 5 and Figure 6, We computed the number of books per publication year for both genres. Both the genres have the raise in highest number of books in recent years. we observed that from the year 2005, the largest number of books are published and from 1940 to 2005, there were remarkably few books published. This could be due to several factors, including the less availability of online bookstores and social reading platforms and also the availability of the

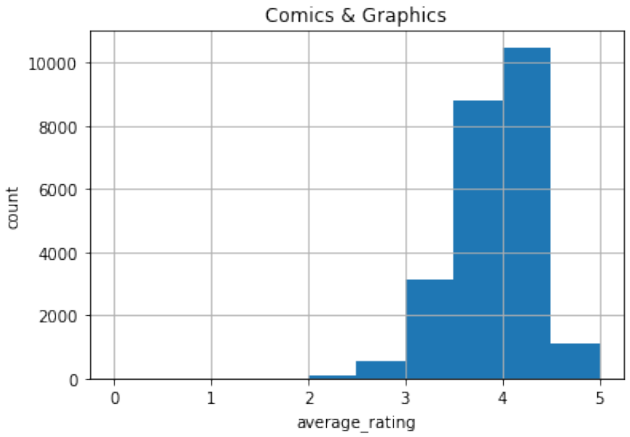

Figure 3. Distribution of book ratings. Comics & Graphics

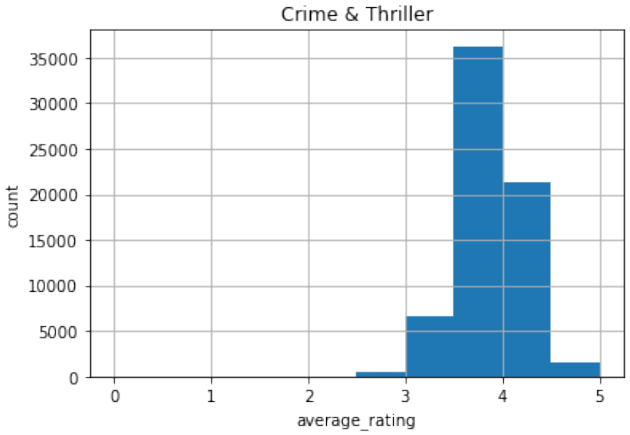

Figure 4. Distribution of book rating. Mystery, Thriller & Crime

internet at different times and the decreasing number of online readers as well as fewer books published.

To show the relationship between ratings count, average ratings, and text reviews count a 3D plot image was created for both comics and graphics, thriller and crime datasets which represents the clustered datasets as shown in figures 7 and 8. For a clear understanding of the both datasets, we have divided it into 3 groups namely purple, yellow and green. Where the purple group had most of its data points across the average ratings axis but with the least text reviews count and ratings count. Whereas the yellow group with a smaller range of average ratings but had a higher number of text reviews count and ratings counts compare to the purple group. Finally, in the green group, average ratings are similar to a yellow group but overall it had a higher range of both text reviews and ratings count than a yellow group.

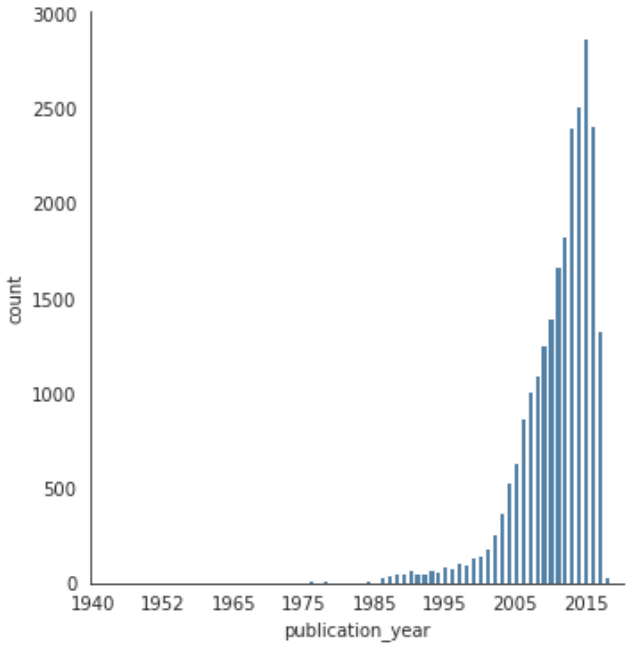

Figure 5. Frequency distribution of books per publication year. Comics & Graphics

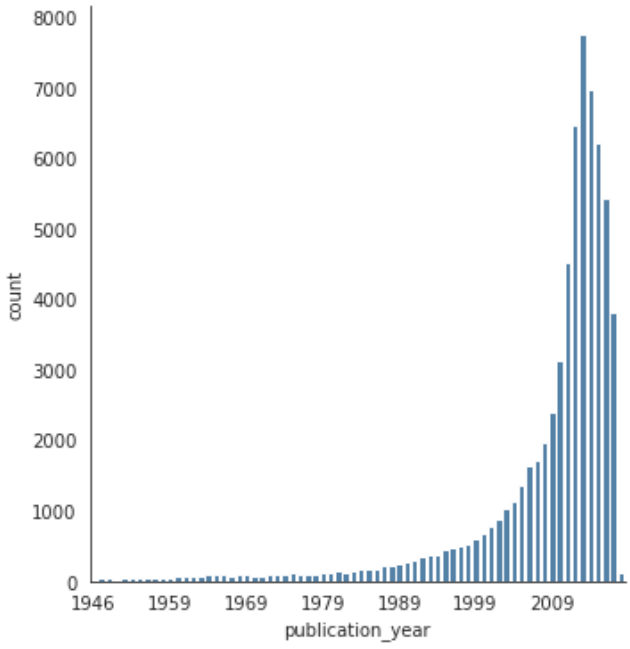

Figure 6. Frequency distribution of books per publication year. Mystery, Thriller & Crime

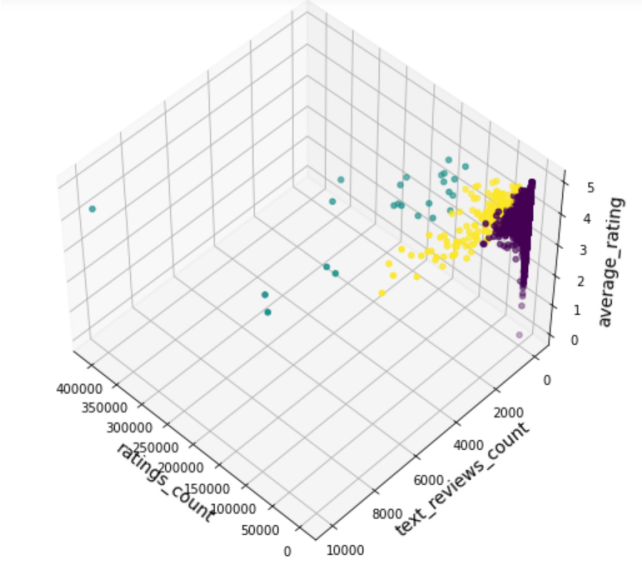

Figure 7. 3D visualization of Comics and Graphics datasets

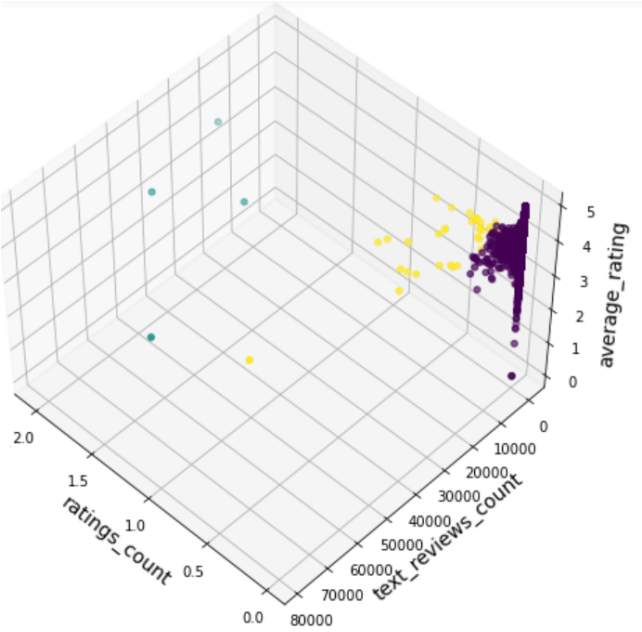

Figure 8. 3D visualization of Mystery, Thriller & Crime datase

5 Research Results

5.1 RQ1: Is there a growth or decline of book genre?

5.1.1 Approach. To solve RQ1, we took metadata of datasets as in Table 1. *Average_rating* and *publication_year* for corresponding *book_id* are the features considered for analysis. *Average_rating* are grouped on the basis of *publication_year* and mean 1 [24], standard deviation 2 [24] of grouped *Average_rating* are calculated for all corresponding year.

$$Mean = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}} \tag{2}$$

Validation of data is important step to ensure the reported measurement are in right direction and very close to the truth i.e. error or uncertainty of our result is minimum. For this purpose we count the number of samples for each *publication_year* and calculate standard error 3 [4] for each *publication_year*. Now we have all the necessary calculations for plotting and analyzing RQ1.

$$Error = \frac{\sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n-1}}}{\sqrt{n}} \tag{3}$$

5.1.2 Analysis. Our answers for RQ1, are as follows:

- Comics & Graphics is declining with time.
- Mystery, Thriller & Crime is growing with time.

Figure 9 shows mean of *average_rating* for each *publication_year* for Comics and Graphics. We can see from 1986 to 2005 the mean rating are bit high but at the same time there are high fluctuations in the graph. After 2005 there is a steadily decay of graph which indicates the genre is declining with time. Similarly Figure 10 shows *average_rating* for each *publication_year* for Mystery, Thriller and Crime. As before, we can see similar pattern where there are more ups and downs in the graph in between 1946 to 2000. After which there is rapid growth of the graph, that indicates the genre is growing.

Since there is always a chance of uncertainty about our research answers, we analyzed the standard error for both genre. The fluctuations occurred in the Figure 9 and Figure 10 might be because of the number of data analyzed may vary each year. Due to which there is a possibility of high error in our analysis.

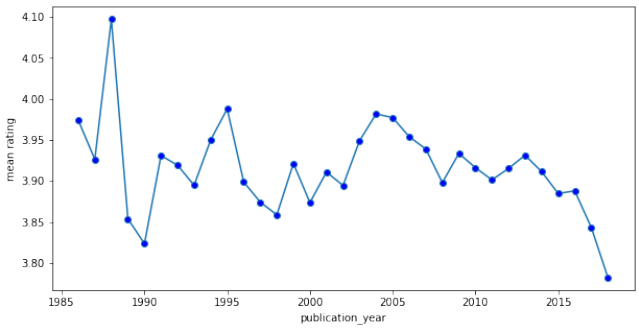

Figure 9. Comics & Graphics Average review rating

Figure 11, depicts standard error for comics & graphics in the form of error plot. It is clearly seen that in the beginning

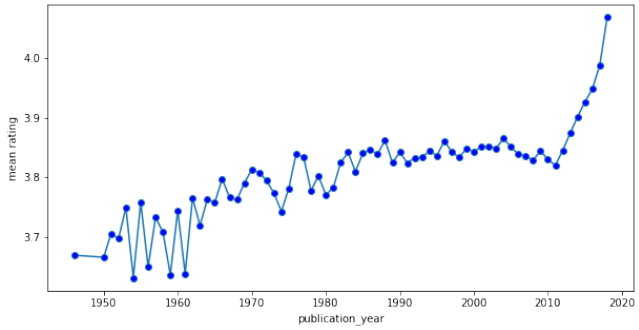

Figure 10. Mystery, Thriller & Crime Average review rating

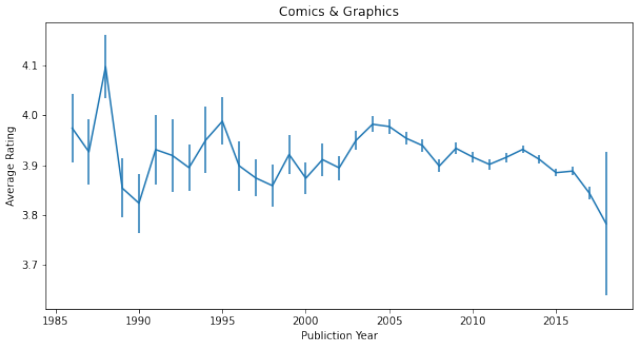

Figure 11. Comics & Graphics Average error bar

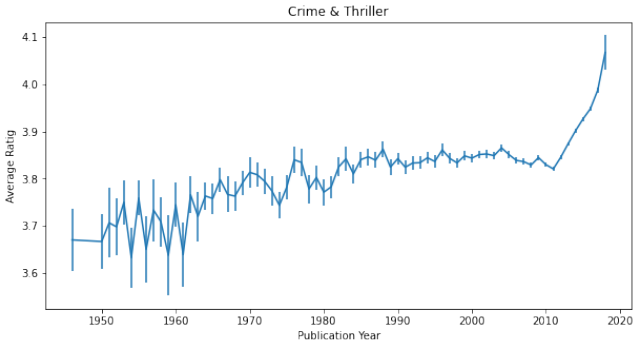

Figure 12. Mystery, Thriller & Crime Average error bar

of publication years the error are quite high, which indicates high error in the reported measurement that will finally reflect in our research answers. But in the recent years, i.e. in between 2005 to 2017, errors are minimum which gives us an strong basis to support our research answers. Similarly, in Figure 12, there is a similar pattern as before, where errors are high in the previous years of publication but after 2000 the uncertainty have decreased with low errors. This also provides us with a concrete base to claim our research answers are very close to truth.

Hence, from the above research analysis considering the publication year which have minimum amount of standard error, we strongly come to a conclusion that Comics & Graphics is declining and Mystery, Thriller & Crime is growing with time.

5.1.3 Contributions. There are various use cases where we can benefit from the RQ1. Online book stores, government organizations, academic institutions, writers are some actors who could make use of such research analysis. It can contribute to explore interest of users, customers, citizens, students to provide them with meaningful books of various genre as per their interest and feedback. Society is an amalgamation of peoples with various interest and to keep the diversity of interest alive academic institutions and government organizations could play an important role to inspire and motivate readers into various genre by analyzing RQ1 into more genre.

5.2 RQ2: What text reviews says about genre: Are readers positive, negative or neutral towards genre?

5.2.1 Appraoch. To refer to our RQ2, we have considered the review data of two Goodreads datasets i.e; Comics & Graphics and Mystery, Thriller & Crime. We fetch the columns *Publication_year* and *review_text* corresponding to given *book_id* as shown in Table 1 and Table 2 . We analyzed the sentiment of the reviews as positive, negative, neutral and compound for each datasets. Standard error with respect to compound value of sentiment are calculated to examine the uncertainty of our research answers. The reviews from the two datasets demonstrates that the language is an informative data to analyze. This study shows the polarity of reviewer's opinions based on compound score towards the genres, in order to improve the reading choices of other individuals.

To address RQ2, we first combined two datasets, i.e. metadata (Table 1) and book review data (Table 2) with respect to *book_id*. *Publication_year* and *review_text* are two main features used for analysis. Based on computational techniques to analyze subjectivity in text, sentiment analysis can determine the attitudes, emotions, and feelings of an individual. We implemented sentiment analysis to detects polarity using a sentiment lexicon called VADER(Valence Aware Dictionary for Sentiment Reasoning) which is available in the NLTK package [13]. VADER includes a combination of a sentiment dictionary, which is a list of lexical features which are generally labeled as positive or negative views, depending on their semantic intensity, which reflects the sentiment scores. [13] In our research, we analyzed the polarity scores as positive, neutral, and negative which represents the proportion of review text that falls under these categories.

Further, analysis of the Compound score is based on the average of all lexicon ratings which are normalized between -1 i.e; extreme negative, and +1 i.e; extreme positive. Table 4

Table 4. Sentiment Scoring

Sentiment	Compound Scores : cs
Pos	cs>=0.05
Neu	cs>=-0.05 & cs<0.05
Neg	cs<=-0.05

shows the compound score metric range. We, are considering compound score for the analysis of our RQ1, as it is widely used for sentiment analysis. Now we have complete data, i.e. compound sentiment and standard error for each publication year to begin with analysis.

5.2.2 Analysis. Our answers for RQ2, are as follows:

- Comics & Graphics is moderately positive.
- Mystery, Thriller & Crime is moderately positive.

In our research, we obtained a compound score value that is around 0.4 for both genres as shown in Figure 13 and 14. Therefore, we concluded that for both genres, the reviews are moderately positive.

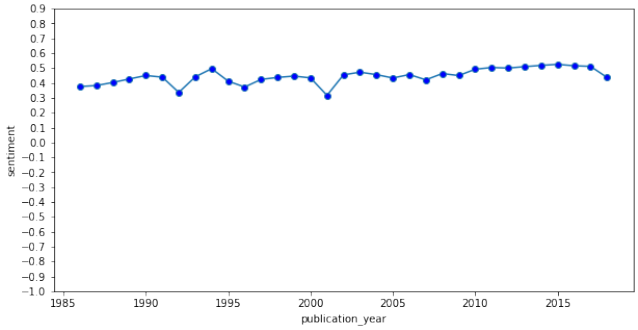

Figure 13. Compound Sentiment. Comics & Graphics

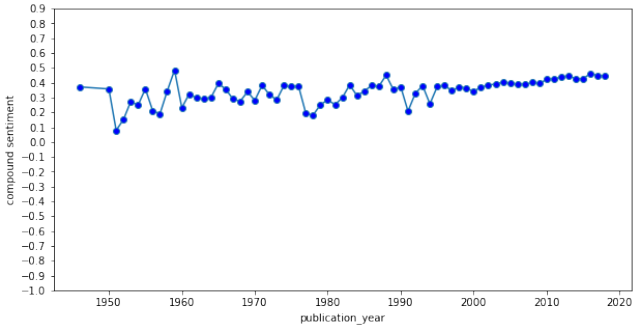

Figure 14. Compound Sentiment. Mystery, Thriller & Crime

However, we cannot justify our research answers is appropriate based on several reasons that may contribute to uncertainty of our analysis. The lack of online bookstores and social reading platforms as well as the availability of the internet at different times might affect the amount of data available. The number of online readers and number of reviews helps us to get opinions of people, but due to uneven data in each year results in uncertainty of measurements. Due to these reasons, we have measured the standard error for both genres. From Figure 15 and 16, we observed that, the error rate in the Comics & Graphics genre is relatively high from the year 1986-2000, which results in uncertainty. However, it is very low from 2000-2017. Whereas in Mystery, Thriller & Crime the error rate is relatively high from the year 1946-2000 and from the year 2000-2017 is very low. Considering the standard error after 2000 we have a strong basis to claim the research answers about RQ2 of both genres are legitimate i.e; the reviews of Comics & Graphics and Mystery, Thriller & Crime are moderately positive.

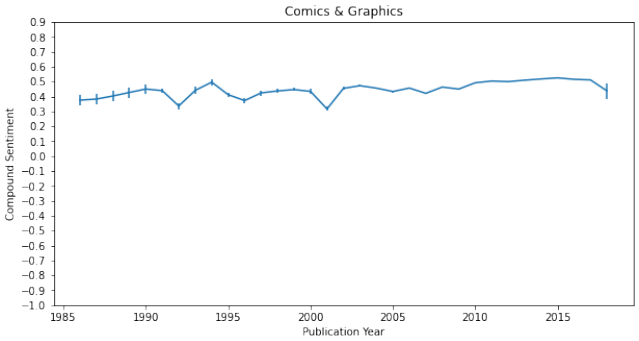

Figure 15. Error bar of sentiment for Comics & Graphics

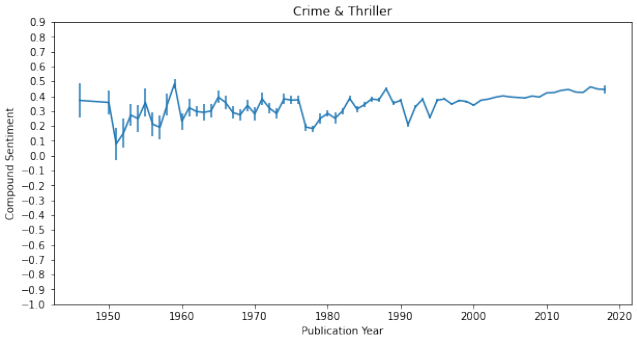

Figure 16. Error bar of Sentiment for Mystery, Thriller & Crime

There are challenges and shortcomings to be addressed. For example the inability to perform well in different areas, the poor accuracy and performance in sentiment analysis due to a lack of labeled data, the inability to handle complex sentences requiring more than sentiment words, and the simplicity of the analysis. Usually, Positive sentiment will be correlated with good or good words, and negative sentiment will be correlated with negative words. While the

challenge here is that different people express their opinions differently, some people will write their thoughts straight while others will inject sarcasm into their writing. Some of the other challenges faced are related to negation detection, word ambiguity. These Challenges will be addressed in future work.

5.2.3 Contribution. RQ2 and its answer can help find the sentiment towards the genres and improve books as per users feedback and motivate writers. Understanding users interest is always an asset to online book stores for providing meaningful resources.

5.3 RQ3: Did the books with more text reviews receive higher ratings?

5.3.1 Approach. We are going to analyze both genre datasets and trying to find how the behavior of the ratings is changing with the respective increasing order of the text reviews count. For this, initially checking any correlation exists between attributes of both genres and then considered the text reviews count and average ratings to know the behavior of ratings.

5.3.2 Analysis. Our solution for RQ3, are as follows:

- The text reviews and the ratings are inline to each other in the both genres.

Firstly, checking the **Correlation** between the attributes. The main purpose of finding the correlation between the variables or attributes in a dataset is to explore the associative relationship between the independent and dependent variables or attributes in a dataset[20]. The correlation coefficient is measured on a scale that varies from +1 through 0 to -1[1]. If the correlation value between the two attributes is close to +1 then that implies both attributes are highly correlated to each other and in the same way the correlation value between both attributes is close to -1 then that represents a negative correlation between them. If a correlation between both attributes is 0 then there is no correlation, which concludes both variables or attributes in a dataset are independent of each other.

Correlation between variables can be represented mathematically. For instance, In general, Pearson's correlation coefficient is used in applications. Pearson formula to quantify the degree of relationship(R) between variables A and B, can be given as[20]:

$$R = \frac{n(\sum AB) - (\sum A).(\sum B)}{\sqrt{n(\sum A^2) - (\sum A)^2}\sqrt{n(\sum B^2) - (\sum B)^2}} \quad (4)$$

where,
n = Number of observations
A= Measures of Variable 1
B = Measures of Variable 2
$\sum$ AB = Sum of the product of respective variable measures
$\sum$ A= Sum of the measures of Variable 1
$\sum$ B = Sum of the measures of Variable 2
$\sum$ A2 = Sum of squared values of the measures of Variable 1
$\sum$ B2 = Sum of squared values of the measures of Variable 2

Table 5. Correlation of Comics and Graphics

	Text reviews count	Ratings count	Average ratings
Text reviews count	1.000	0.759	0.066
Ratings count	0.759	1.000	0.101
Average ratings	0.066	0.101	1.000

Table 6. Correlation of Mystery, Thriller & Crime

	Text reviews count	Ratings count	Average ratings
Text reviews count	1.000	0.732	0.005
Ratings count	0.732	1.000	0.013
Average ratings	0.005	0.013	1.000

From both the tables 5 and 6, we can observe that in the graphics and comics dataset, the correlation between the text review count and the ratings count is very high its around 76 percent. In the same way, the thriller and crime dataset attributes namely text review count and ratings count are also highly correlated to each other with 73 percent. But, we can see that there is no influence of ratings and text review count on average ratings count in both datasets because the correlation value between aver-age ratings count and the other two are almost 0. So, we can visualize the relation between average ratings count and the text review count in order to observe the behavior of the average ratings count with respect to the increasing order of the text review count.

Here, from the figure 17 and 18 we can observe that , we have considered the text review count on the X-axis and average ratings count on the Y-axis, The blue data points represent the text review counts of a particular book with the respective X-axis and in the same way to the Y-axis it indicates the respective Average ratings of the same book. We also considered the confidence interval for the predicted output. Where The confidence interval describes a certain range of the estimated plot[2].

If we look into the comics and graphics dataset figure 17 we can observe that as the number of text review count

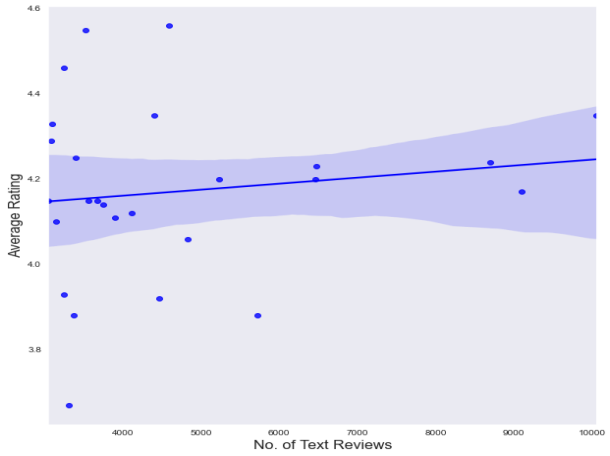

Figure 17. Text reviews count VS Average ratings of Comics and Graphics

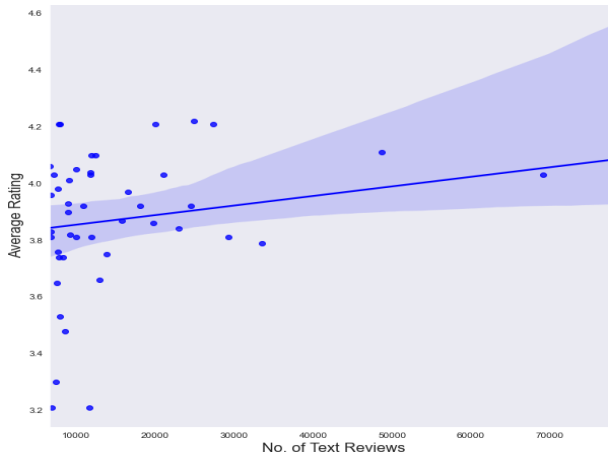

Figure 18. Text reviews count VS Average ratings of Mystery, Thriller & Crime

increasing respectively the average ratings also gradually increasing from around just above 4.1 ratings to the 4.2 ratings. Similarly, when we observe the thriller and crime dataset figure 18 as the number of text review count increasing respectively the average ratings also gradually increasing from around just above 3.8 ratings to almost 4.0 rating.

So, from both genre output plots figure 17 and 18 , we can conclude that both the text reviews and the ratings are approximately inline in both datasets.

Now, we want to check whether the same relation exhibits or not when we consider small portions or percent of data from the datasets. For instance, let us consider the Top20 rated books from each genre.

Here we have considered the only top 20 rated books from both the genres and plotted between text review count and the average ratings.

Interestingly, from the comics and graphics dataset figure

Figure 19. Top20 rated books of Comics and Graphics

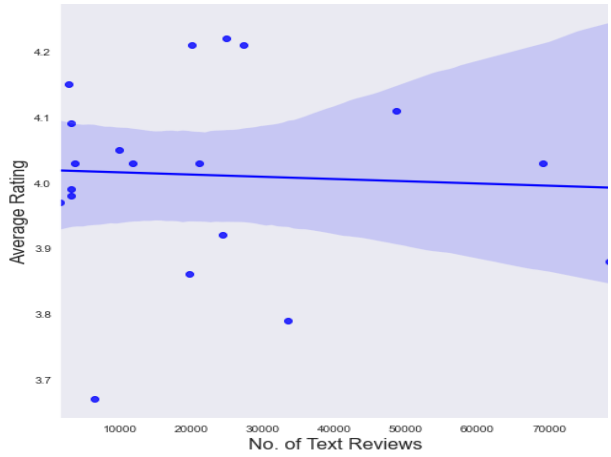

Figure 20. Top20 rated books of Mystery, Thriller & Crime

19 we can notice that the number of text reviews count on the X-axis increases respective average ratings on the Y-axis started declining slowly from around 4.4 ratings to almost between 4.2 and 4.3 ratings. In the same way, if we have seen at the thriller and crime dataset figure 20 the output line almost remains constant at around just above the 4.0 rating, between text reviews count and average ratings.

So, from both plots figure 19 and 20, we can assume that for a small amount of data from datasets the behavior of the ratings is not inline with the text reviews count. But, these changes in the behavior of the ratings are very minute which is almost negligible.

Supportive Statement: To support the above statement, we have considered the various subsets from the dataset of both genre's like considering the top 100 rated books, taking 500 to 1000 rated books, bottom 500 rated books, and between 18000 to 20000 rated books. In all the cases we have

noticed that almost negligible or increasing order behavior of the ratings with respect to the increase in text reviews count.

Reason: The one of the main reasons might be that the considered subsets of data are very less percent of the whole datasets.
But, When we considered more percent of datasets or whole datasets at a time we observed that both the text reviews count and the ratings are inline.

Finding: Therefore, finally we can conclude that, as text reviews count increases respective average ratings also increasing which implies both are inline to each other.

5.3.3 Contributions. This approach can be helpful, When there is a need of knowing the behavior of ratings with respect to increasing order of text review counts in a dataset.

5.4 RQ4: How to predict ratings from review text?

5.4.1 Approach. We considered the ratings and text review attributes from both genres and implemented a model to predict the appropriate ratings from respective text reviews of a book from both genres.

5.4.2 Analysis. Our solution for RQ4, are as follows:

- The implemented model had predicted the ratings with an accuracy of above 50 percent from the text reviews for both genres.

Firstly, we took a randomly sampled dataset of a total of 15 percent from the original dataset reviews. But, in the training, we took an almost very small quantity of data to train the model easily and to get the overall accuracy trained the model with the overall dataset later.

Secondly, We split the data into train and test data. When we tried to visualize the dataset Rating vs Test reviews count from both datasets, the rating 4 is a little skewed towards the rating 5 as shown in figure 21 and 22. This means, there is an imbalance in the dataset. So, to balance the data we did one-hot encoding to balance the data. We created the dummy variables for each rating, this had created the binary columns for each level of rating which can be seen in the table 7.

Table 7. sample creation of dummy ratings using one hot encoding

Actual Rating	Dummy Ratings					
	0	1	2	3	4	5
2	0	0	1	0	0	0
4	0	0	0	0	1	0
3	0	0	0	1	0	0
5	0	0	0	0	0	1

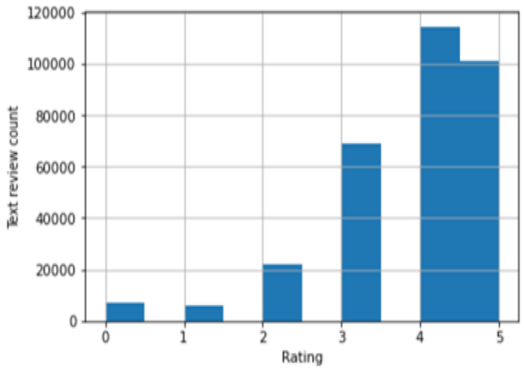

Figure 21. Rating VS Text reviews count of Comics and Graphics

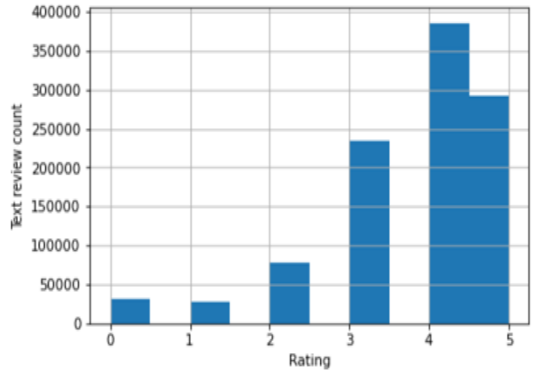

Figure 22. Rating VS Text reviews count of Mystery, Thriller & Crime

After that, we took the portion of sample data(0.1) randomly to train the model easily. When we found the better model, retained the model with total data. In the final preprocessing step, TF-IDF (Term Frequency inverse document frequency) weighs the words rather than Creating word vectors with Bag of Words(BoW). TF-IDF creates the weights for every word and it gives less weight for the high-frequency words and More weight for low-frequency words.
For example, there are some words that will come very frequently in the English language are they, the an, a. For these words, it gives very little weight as we can negligible. If there is a word good or bad or something similar words will get higher weight because these words will decide the sentence is tilting towards the positive or negative side in the review.

We have used the hybrid model by combining the Naive Bayes and Logistic Regression. Firstly, the Naive Bayes Classifier is works on the principle of the Bayes theorem. It gives the conditional Probability for each review that can be classified as rating 1 - 5.
For instance, we can take one small review “This book is terrible”. Then we calculate what is the probability of this

review is classified as rating 1 or rating 2 or rating 3 or rating 4 or rating 5. Otherwise, we can ask what is the probability of a 1 rating given by the Review "This book is terrible". Then based on the conditional probability we can calculate as:

$$P\left(\frac{this}{1rating}\right)*P\left(\frac{book}{1rating}\right)*P\left(\frac{is}{1rating}\right)*P\left(\frac{terrible}{1}rating\right) \tag{5}$$

After calculating this, we predict the rating of which probability is highest and classify accordingly. It gives a 1 rating if negative or a 5 ratings if it is positive.

To get better results, we have used the Naive Bayes calculations or features in the Logistic Regression model. We took Naive Bayes probability calculations and multiply with the original feature matrix and then learn the parameters with the Logistic Regression model. This model has explained in Jeremy Howard's explanation.

After doing this, we take sklearn pipeline to combine the steps. First, we took the data and tokenises and calculates the TF-IDF. These TF-IDF features are fed to the NB features method to create the NB features and these NB features are fed to the Logistic Regression to learn the parameters.

Table 8. CV score of Comics & Graphics and Mystery, Thriller & Crime

CV Score of ratings	Comics & Graphics	Crime & Thriller
Rating 1	0.982	0.974
Rating 2	0.932	0.926
Rating 3	0.787	0.780
Rating 4	0.646	0.641
Rating 5	0.741	0.778

To predict how accurate our results we used 3 cross-fold validation as shown in table 8. We should evaluate that, the rating for the Particular text review is predicted correctly or not. Whenever we predict the ratings for the Particular text Review, it does not mean that it is an accurate one. From table 8 we can observe that, the Cross Validation score is more for rating 1 than 2 and 3 and also 5 compare to 4 respectively.

For a better comparison of ratings, we have calculated the F1 score as shown in table 9. When we look into F1 score results, Rating 1 and Rating 5 are classified as approximately accurate with 66 and 68 percent for rating 1 in comics & graphics and crime & thriller genre respectively. Similarly, for rating 5 model had predicted with 73 and 76 percent of accuracy for both datasets respectively. But, the model registered with very low accuracy for Ratings 2, 3, and 4 respectively for both comics and crime genres. Because these ratings are more ambiguous.
For example, we have a review of rating 2 or 4 when our model tries to classify these ratings it may predict wrongly

Table 9. F1 score of Comics & Graphics and Mystery, Thriller & Crime

F1 Score of ratings	Comics & Graphics	Crime & Thriller
Rating 1	0.66	0.68
Rating 2	0.30	0.29
Rating 3	0.42	0.46
Rating 4	0.47	0.49
Rating 5	0.73	0.76
Average	0.51	0.53

because these rating text reviews may be similar to the rating text reviews of 1 or 5 respectively. Then, it's very hard for the model to predict the actual rating from the respective test reviews.
Finally, if we observe our F1 score table once again we can conclude that the average accuracy of both genres is more than 50 percent which implies our model is predicting or working moderately for both datasets.

Especially, for this research question, we have tried a couple of other approaches like predicting the rating directly from the text reviews using various regression models, SVM . But, we didn't get good results. So finally, we end up with these approach in which we couldn't get a higher accuracy rate but comparatively with other approaches we had a good result with this.

5.4.3 Contributions. This approach can be helpful, when there is a scenario of finding the ratings from the Text reviews of a respective book.

6 Conclusion

In our research, we explored dynamics of book reviews. From our research questions, we found that Comics & Graphics is declining but Mystery, Thriller & Crime is an increasing genre. On the other hand sentiment analysis of text reviews associated with books shows readers are moderately positive about both genre. The two statement above contradicts in case of Comics & Graphics and the reason behind this might be the fact that text reviews are more informative than star ratings, at the same time there are limitations of sentiment analysis as well. We further worked to see the correlation between text reviews and star ratings, and found out that as one is increasing so is the other, which means they are inline to each other for both genre. This helps to find out the relation between dependent and independent variables. Finally, predicting ratings from text review was interesting to explore and our model had an accuracy of above 50 percent for both genre. Here 50 percent sounds quite small number but considering the fact that text reviews are opinions of

each individual and may contains sarcasm and problems of negation, word ambiguity which make the problem quite challenging. Hence we take above 50 percent of accuracy as a good achievement. We believe our research will contribute to organizations and society to their best interest.

Individual Contributions

Upender Shana Gonda and Keerthi Sagar Reddy Aereddy: We both Contributed to data acquisition, preprocessing, Related work and 3D visualization of both genres and involved in research questions discussion. In our 4 RQ's, We were involved in the implementation and report writing for 2 RQ's completely namely, RQ3 and RQ4. Moreover, in our weekly group discussions, We also participated and exchanged our ideas regarding the other 2 research questions as well for better approaches and getting to know the regular updates from other 2 teammates. Also, contributed to the phase-wise presentations as well in making and presenting the slides. Overall, its a teamwork, and everyone contributed their part in the successful completion of the Lab.

References

[1] [n.d.]. 11. Correlation and regression. https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression. (Accessed on 07/19/2021).

[2] [n.d.]. Confidence Intervals. https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Confidence_Intervals/BS704_Confidence_Intervals_print.html. (Accessed on 07/19/2021).

[3] Akiko Aizawa. 2003. An information-theoretic perspective of tf–idf measures. *Information Processing & Management* 39, 1 (2003), 45–65.

[4] Douglas G Altman and J Martin Bland. 2005. Standard deviations and standard errors. *Bmj* 331, 7521 (2005), 903.

[5] Xavier Amatriain, Neal Lathia, Josep M Pujol, Haewoon Kwak, and Nuria Oliver. 2009. The wisdom of the few: a collaborative filtering approach based on expert opinions from the web. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 532–539.

[6] Justin Basilico and Thomas Hofmann. 2004. Unifying collaborative and content-based filtering. In *Proceedings of the twenty-first international conference on Machine learning*. 9.

[7] Zhiyong Cheng, Ying Ding, Xiangnan He, Lei Zhu, Xuemeng Song, and Mohan S Kankanhalli. 2018. Aˆ 3NCF: An Adaptive Aspect Attention Model for Rating Prediction.. In *IJCAI*. 3748–3754.

[8] Judith A Chevalier and Dina Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of marketing research* 43, 3 (2006), 345–354.

[9] Cristian Danescu-Niculescu-Mizil, Gueorgi Kossinets, Jon Kleinberg, and Lillian Lee. 2009. How opinions are received by online communities: a case study on amazon. com helpfulness votes. In *Proceedings of the 18th international conference on World wide web*. 141–150.

[10] Statista Research Department. 2021. Number of registered members on Goodreads from May 2011 to July 2019. https://www.statista.com/statistics/252986/number-of-registered-members-on-goodreadscom/. (Accessed on 07/10/2021).

[11] Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: improving rating predictions using review text content.. In *WebDB*, Vol. 9. Citeseer, 1–6.

[12] Goodreads. 2019. Goodreads datasets. https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/home?authuser=0. (Accessed on 07/11/2021).

[13] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8.

[14] Guang Ling, Michael R Lyu, and Irwin King. 2014. Ratings meet reviews, a combined approach to recommend. In *Proceedings of the 8th ACM Conference on Recommender systems*. 105–112.

[15] Natasha Mack. 2005. Qualitative research methods: A data collector's field guide. (2005).

[16] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*. 165–172.

[17] Prem Melville, Raymond J Mooney, Ramadass Nagarajan, et al. 2002. Content-boosted collaborative filtering for improved recommendations. *Aaai/iaai* 23 (2002), 187–192.

[18] Tiago Santos, Florian Lemmerich, Markus Strohmaier, and Denis Helic. 2019. What's in a Review: Discrepancies Between Expert and Amateur Reviews of Video Games on Metacritic. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–22.

[19] Cedric Seger. 2018. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing.

[20] Samithamby Senthilnathan. 2019. Usefulness of correlation Analysis. *Available at SSRN 3416918* (2019).

[21] Craig Smith. 2021. Goodreads Statistics, User Counts and Facts. https://expandedramblings.com/index.php/goodreads-facts-and-statistics/. (Accessed on 07/11/2021).

[22] Mengting Wan and Julian J. McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan (Eds.). ACM, 86–94. https://doi.org/10.1145/3240323.3240369

[23] Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. 2019. Fine-Grained Spoiler Detection from Large-Scale Review Corpora. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28-August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 2605–2610. https://doi.org/10.18653/v1/p19-1248

[24] Xiang Wan, Wenqian Wang, Jiming Liu, and Tiejun Tong. 2014. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC medical research methodology* 14, 1 (2014), 1–13.