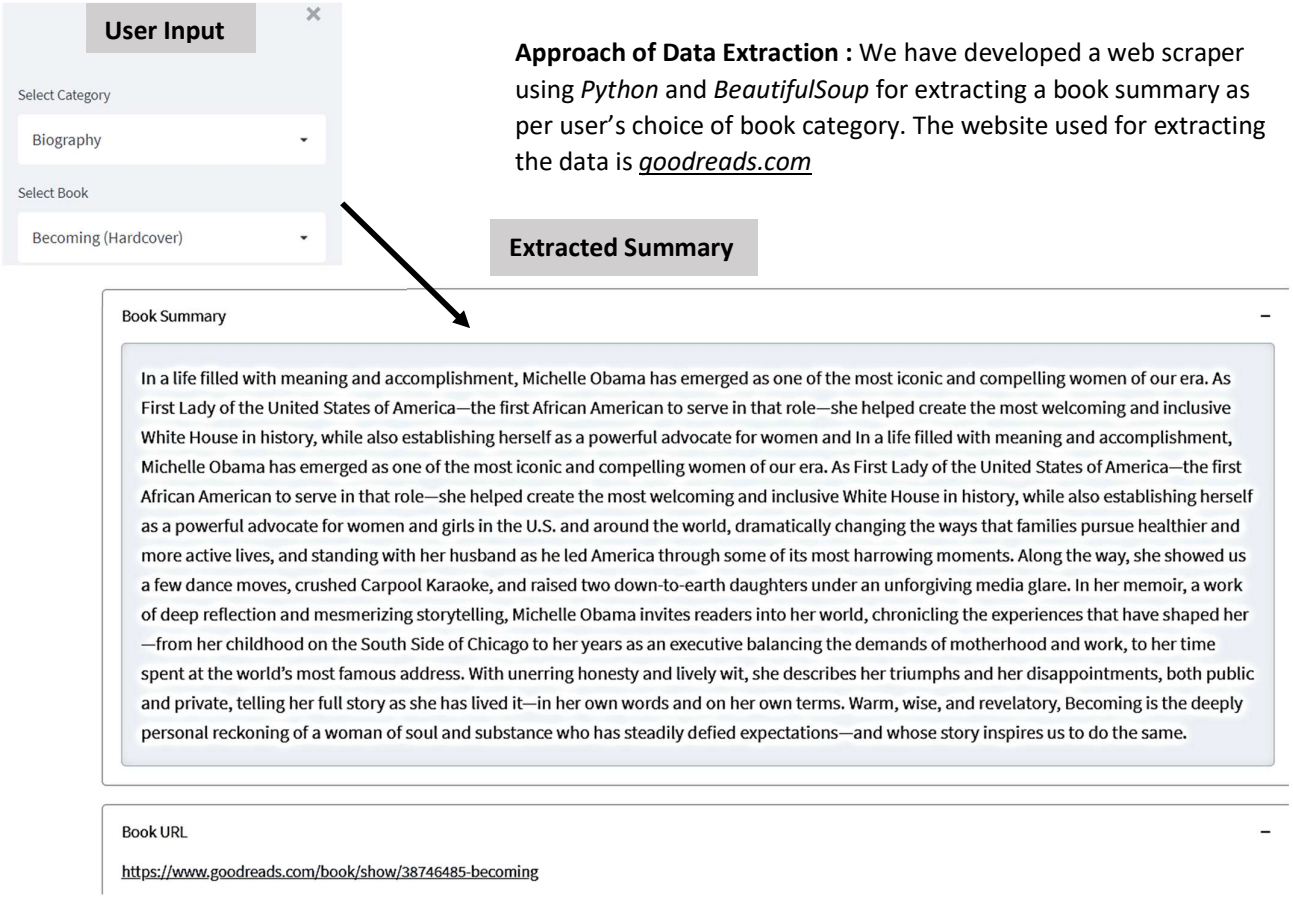
Book Summary Extraction and Sentiment Analysis

Data Exploration

2. Summary Extraction



Approach of Data Extraction : We have developed a web scraper using *Python* and *BeautifulSoup* for extracting a book summary as per user's choice of book category. The website used for extracting the data is *goodreads.com*

1. Basic Exploratory Data Analysis

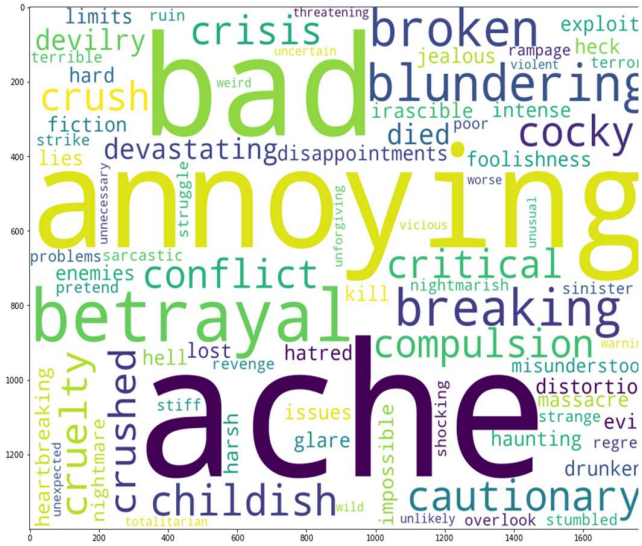
Approach:

- User has around 29 different book category or genre options and each category has 50 different books and 1450 books in total. It is not feasible and useful to run EDA in entire dataset. Hence, we have taken 4 different categories (**Fiction, Sport, Horror, Biography**), each distinct from one another.
- We have selected these categories to check the efficacy and quality of data that we have extracted.
- We have carried basic EDA initially just to check the variety of words and parts

A. Frequently used words in the

B. Positive Words in the summaries

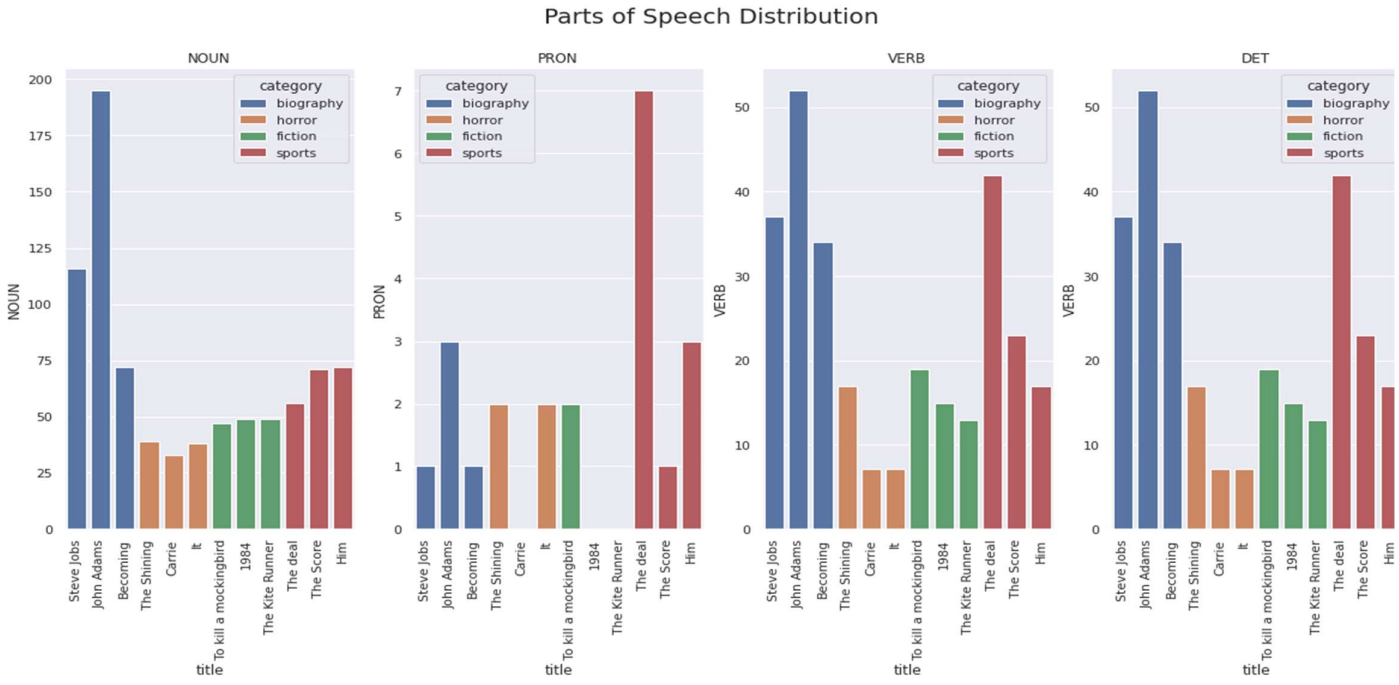
C. Negative Words in the summaries



3. Parts of Speech (POS) Tagging

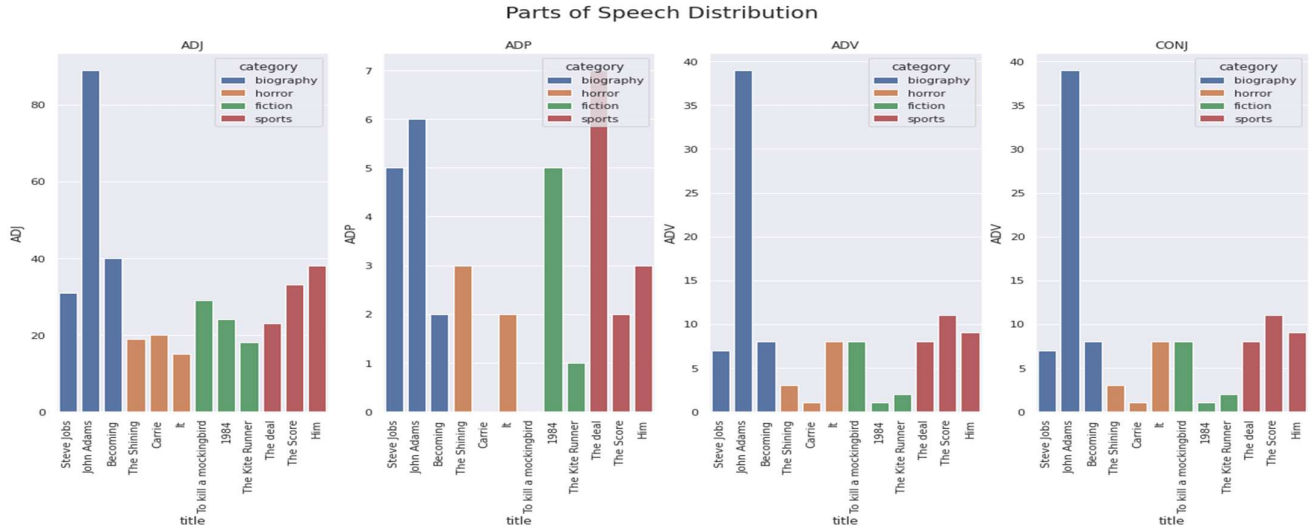
- As we are dealing with diverse categories of book, sometimes it becomes difficult for a user to just decide which book to read based on sentiment of the corpus.
- Part of speech tagging used in Entity Recognition is one method which will quickly help user to scan the book summary quickly.
- Just by a glance at the summary will help user understand what are the major characters, organizations, places, and even the plot of the Book.
- Hence, for the later part of our NER tagging we carried of POS tagging for our extracted data.
- Have a look at book-wise parts of speech charts.

D. Parts of speech (Useful POS: Noun, Verb)



As shown in above figure D, Biography books have generally highest number of nouns, off course it is subject to number word counts, but here the books we have selected have almost similar summary word-count, except for the book John Adams.

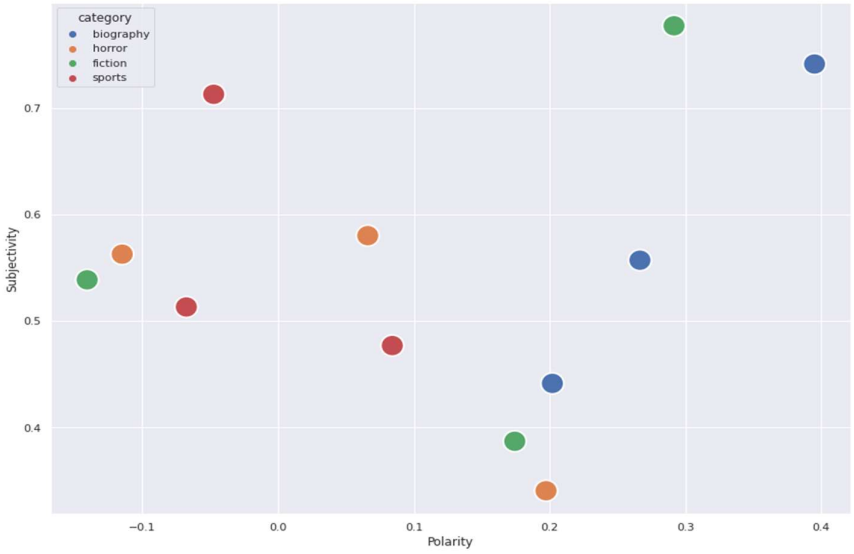
E. Parts of speech (Useful POS: Adjective)



5. Subjectivity and Polarity Analysis of the book summaries

Before, going ahead with building a sentiment analysis model for the extracted summary, we wanted to check the whether the extracted summary represents the ethos of an entire book. Hence, we decided to do Subjectivity and Polarity analysis to check the emotional and factual context of the text corpus.

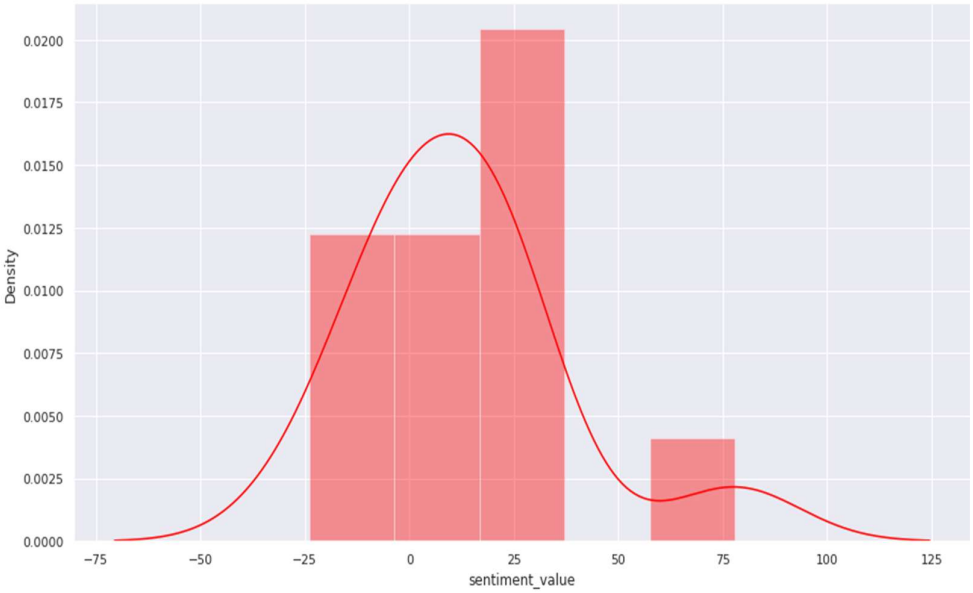
Objectivity [-1 to 1]
Subjectivity [0 to 1]



- Books with the category Horror best depict the subjectivity and objectivity relation in this chart.
- 2 books from the horror category show the negativity, that possibly because of the choice of negative words in the book summary corpus.
- Books with the Biography category have higher polarity, and are showing expected behaviours.
- Category of Sports and fiction shows no common pattern, because of its contextuality.

7. Calculating sentiment values based on Affinity scores

We cannot really define positivity or Negativity of the text based on the Polarity and the subjectivity of the text because it just decides the values best on independent word. We need to take more factors into account such as what is the semantic relation between the words inside the corpus of text. For that we decided to calculate the sentiment value of the entire summary based on the affinity score of each word present in the text corpus



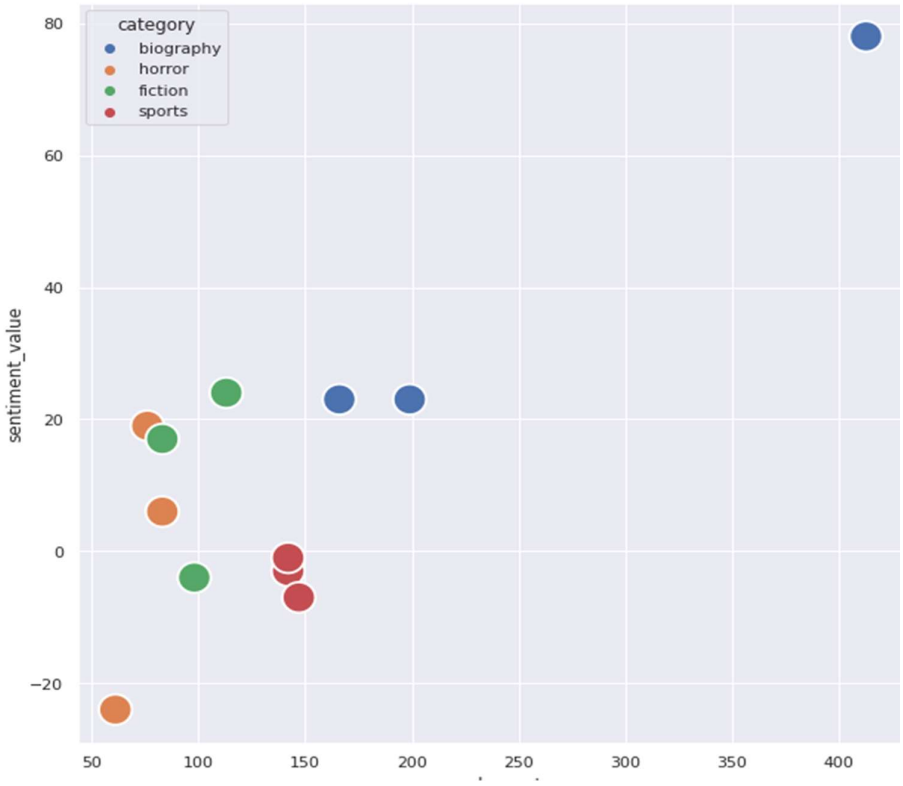
F. Distribution of Sentiment values across categories

- In this chart we can see the distribution of sentiment values for all 4 categories of books.
- We'll get the clear idea with the bar chart for every book just to see which categories are at the extreme ends of this chart

G. Sentiment values Books and Category wise

- Above chart is showing the expected behaviour for Biography books, User expects it to be inspiring and motivational hence the positive trend in these types of books is as expected.
- Books with Horror category are also showing expected behaviour, they are expected to be negative because presence of negative and haunting characters of the book.
- Fiction books are also showing expected trend.
- But the problem lies with Sports category, all of them are inclined towards negativity which is kind of questionable. It might be because of smaller word count. Hence, we will carry out word count to sentiment value analysis just to see the effect of number of words on sentiment value

H. Sentiment values to word count analysis

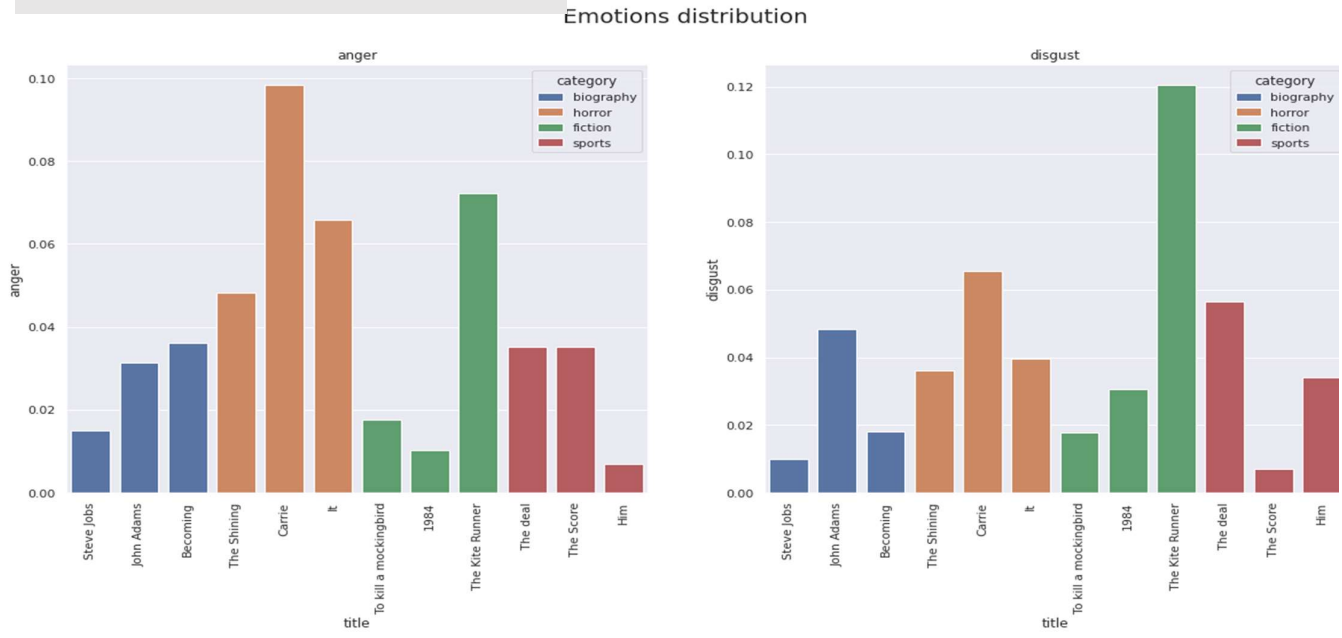


- Here, we can clearly see that in sports category word count is smaller than most of the books, and that it is affecting the sentiment value as well.
- To overcome this, we are taking word count into account , along with various other emotions such as joy, anger, disgust, anticipation, surprise.
- With raw emotions analysis we can increase the certainty of recommendation, and user will decide based on these emotions whether to read the book or not.

I. Raw emotions analysis

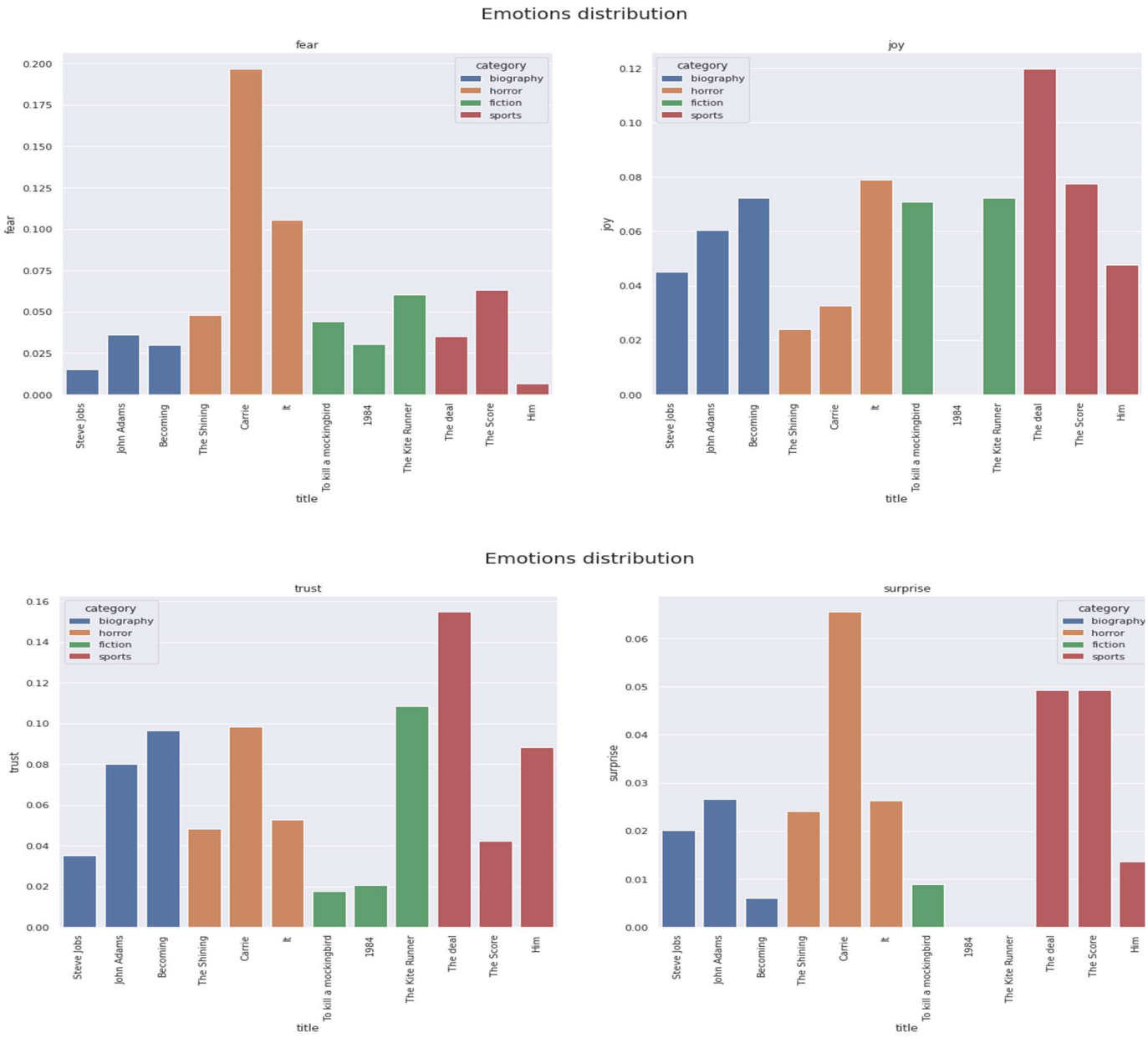
All the above EDA methods till now gives us the idea about How Positive, Negative, or Neutral the book summary is. But, when it comes to deciding the contextuality of the book such as Horror Category, it is expected that the book sentiment will be Negative, it doesn't tell user who is fan of Horror category books any useful information. Hence, in this case emotions other that positive , negative and neutral are necessary. Checkout the below chart depicting more such raw emotions from every book and category.

J. Emotion analysis (anger, disgust)



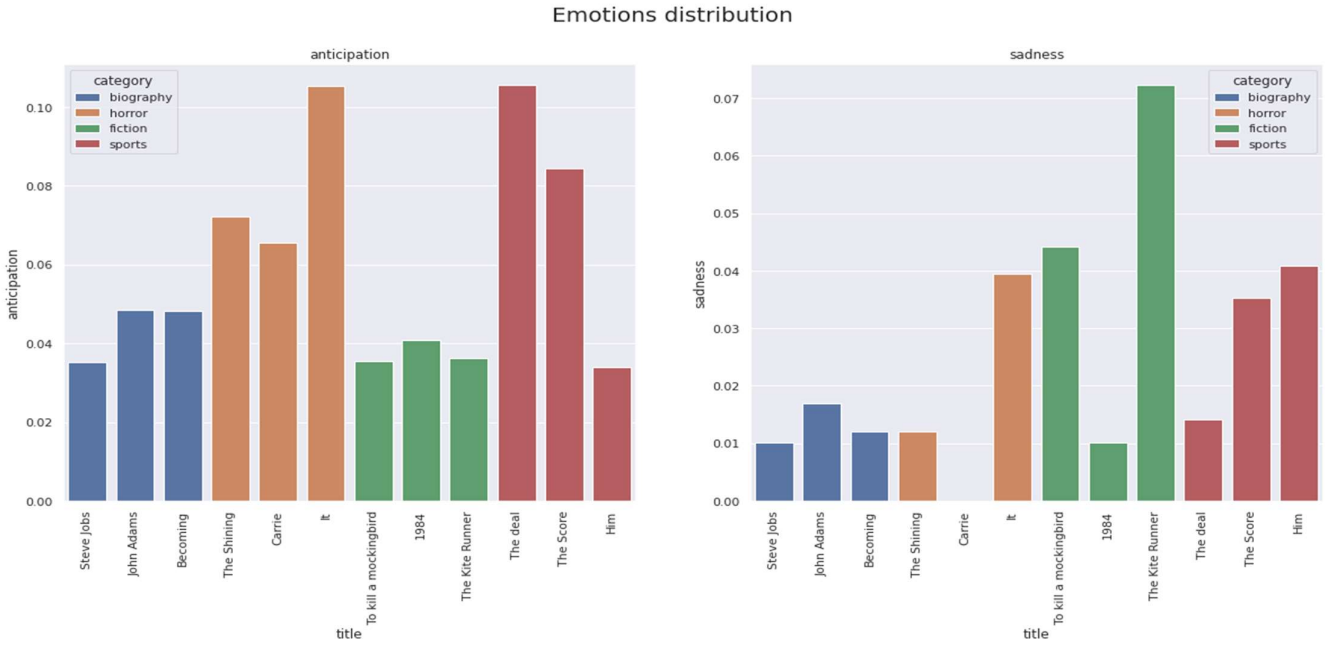
From above chart we get the clear idea about anger and disgust characteristics of a book summaries. Horror category books has maximum anger and disgust factors which is as expected. Kite Runner is also one such book which depicts the authors frustration and disbelief in the authoritarian regime at the time. Hence, the analysis is correct in this case also.

K. Emotion analysis (Fear, joy, trust, surprise)



- From above charts, we can see that Carrie which is a horror story book has a maximum fear and surprise element.
- The Trust chart solves our purpose of emotion analysis. Because in previous sentiment value analysis we were constantly getting Negative sentiments for sports category books. Here we get the maximum Trust and Joy values for the sports category books, and user gets the clear idea about what this book has to offer.
- If we take the example of Kite Runner which we saw in the previous chart, here the trust value is promising, because of the characters interpersonal relations with his family. Hence, the analysis trustworthy and capturing the ethos of an entire book.

L. Emotion Analysis (Anticipation, Sadness)

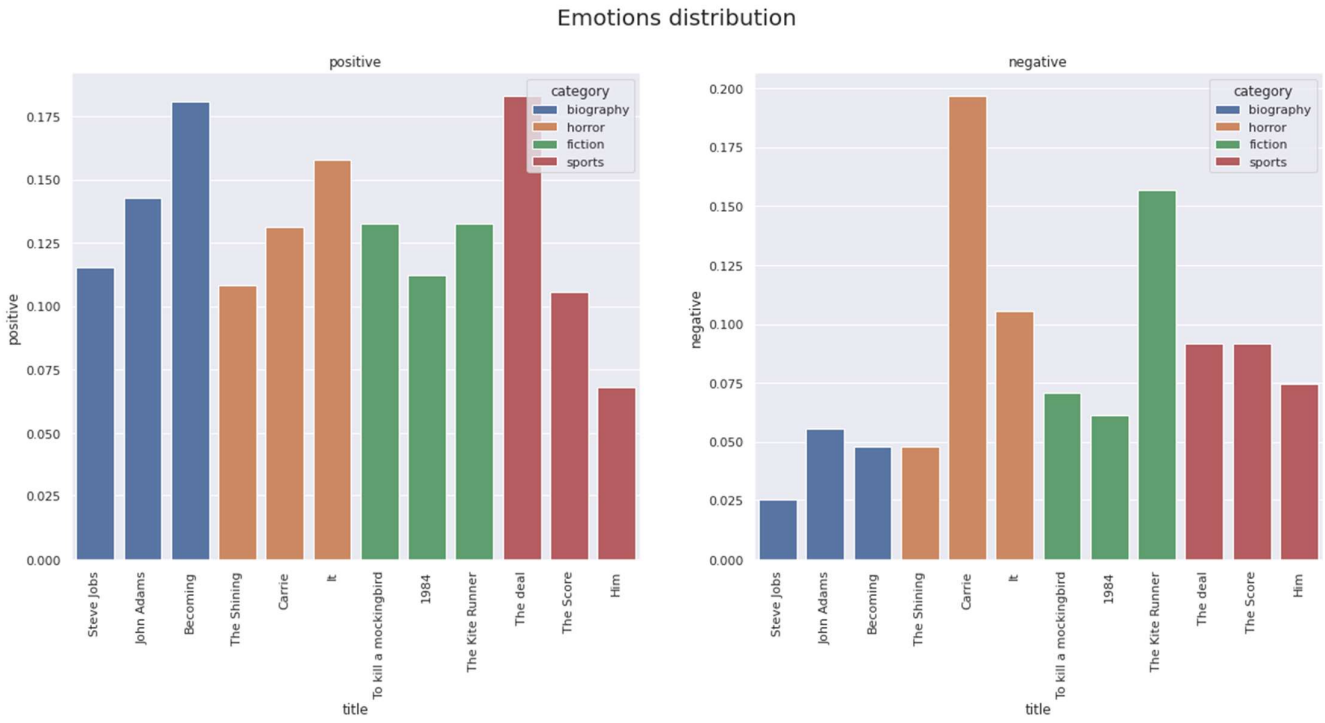


In the above chart for Anticipation, the sports category has higher values. Person reading sports category looks for hope and excitement and hence the value of anticipation is reflecting the content of sports category books correctly.

M. Emotion Analysis (Positive and Negative)

This analysis tells us the Positivity and Negativity of the book summary as per the Emotion analysis. Here, we have taken into account the word count of text corpus. Hence, we can compare this analysis with one that we carried out in Figure G using Sentiment Values

New Method with word count consideration and emotion analysis



By comparing above charts, we can clearly see the difference in sentiment values. Analysis in the second chart looks more promising and reflects the original sentiment of the book. We can see the negative sentiments in the first chart for Sports category and in the second chart how it has been corrected due consideration of various raw emotions and word count of the summary text.