Sentiment Analysis:

A comparison of “Alice’s Adventures in Wonderland” and “Moby Dick”

Essex ID: ad20322

Introduction

Sentiment analysis and text mining helps us get an insight into the type of sentiments or emotions expressed in the text by the author. This kind of analysis also aids in understanding the similarities and differences in sentiments among texts belonging to same or different categories of books. For this study, two texts are used from two different categories of audience, namely children and adult groups. The title, “Alice’s Adventures in Wonderland” is chosen from “The Child list” and the title “Moby Dick” from the list “The Adult list”. The text “Alice’s Adventures in Wonderland”, belongs to the genre fantasy and is aimed at a younger set of audience. On the other hand, the book “Moby Dick”, is aimed at adult readers and is an adventure fiction.

In this study, we will analyse both the texts to get the sentiments in them. We will see how prevalent different sentiments like positive and negative are in each text. We will also compare the two texts to see if sentiments like anger, joy, sadness, and fear are present. We will see up to what extent these sentiments are evident in both the texts. Often, we would expect the texts or books written for children to not to be too long and consist of less negative emotions, when compared to adult texts. Similarly, it would also be expected to have more of emotions like joy and surprise, rather than anger, sadness, and fear. We will compare the two texts in this study to investigate if the assumptions mentioned above are true. We will investigate if the sentiments in both the texts are similar or different.

Methods

In order to perform sentiment analysis, firstly, both the datasets and the required packages are loaded. In this analysis, the packages for text analytics and sentiment analysis, namely stringr, tidytext, and textdata are being used. We are going to perform the same analysis on both the datasets and compare the outputs.

After the preliminary investigation of data, we perform data cleaning. We retain only the columns and the rows required for the analysis. The columns other than the text and the rows other than the main body of the text are removed. Further, we are going to find any other patterns which may not add much significance to the analysis and filter them. Patterns like chapter numbers, titles and empty rows are detected in the data and filtered using stringr package. Further, it is an essential part of the process to extract vector of words or tokens from the text. In this stage, all the punctuations are removed and the letters are converted to lower case to maintain uniformity among the words, throughout the data. Although, the data was cleaned initially, it is important to again investigate the tokens created. Any white spaces, digits or unnecessary characters still present in the data are removed. The next steps in the analysis involve getting the count of all the unique words in the data and removing all the stop words among them. These are the words which are commonly used in the English language and may not add significance in sentiment analysis. All the above steps are applied to both the datasets, from child and adult list. The two datasets are cleaned and prepared for sentiment analysis.

Firstly, we will analyse the quantitative statistics of the words in both the datasets. We will refer to the dataset “Alice’s Adventures in Wonderland”, as child dataset and to “Moby Dick” as adult dataset to draw easy comparison between them. The child dataset has a total of 2192 words in it. On the other hand, the adult dataset has 16,466 words. The maximum frequency of words in the child data is given by 386 and by 1028 in the adult dataset. The median word frequency is the same for both the datasets, which is 2.

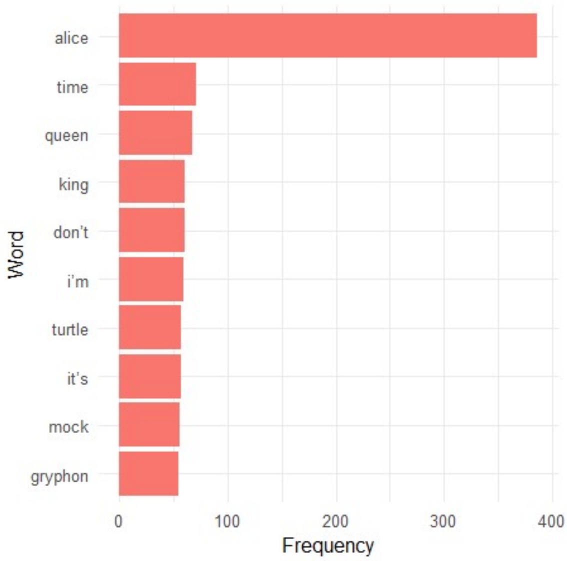


Fig.1. 10 most frequent words in child data

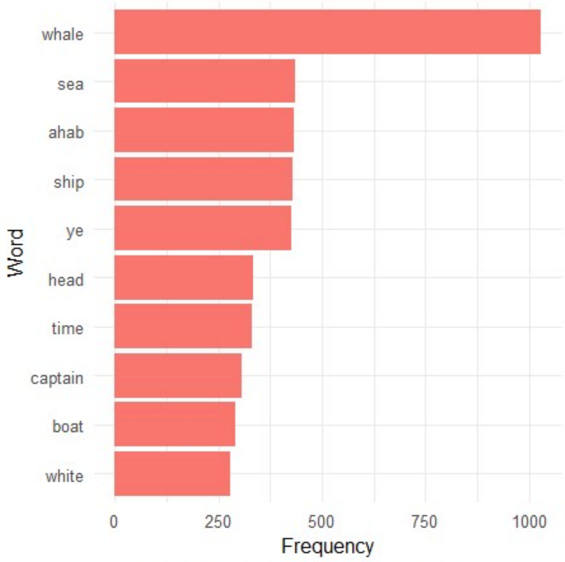


Fig.2. 10 most frequent words in adult data

Fig.1 and 2 show the 10 most frequently occurring words in each dataset. As observed from the figure, some of the top occurring words are in child data are “alice”, “time”, “queen”, and “king”. The frequency of the word “alice” is significantly high compared to other words in the data, which is 386 times. Similarly, in the adult data, the words “whale”, “sea”, “ahab”, and “ship” have the highest number of occurrences. These words refer to the main characters and objects, around which each story is framed.

Furthermore, we will analyse the sentiments in both the texts. There are different lexicons within the packages loaded initially to identify various sentiments for the words in the texts.

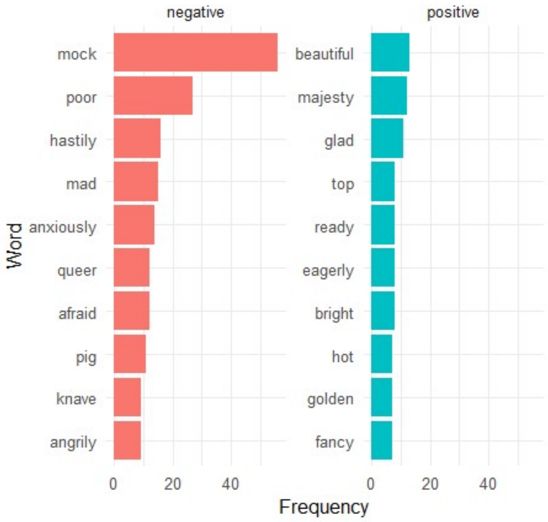


Fig.3. Top 10 positive and negative sentiments in child data

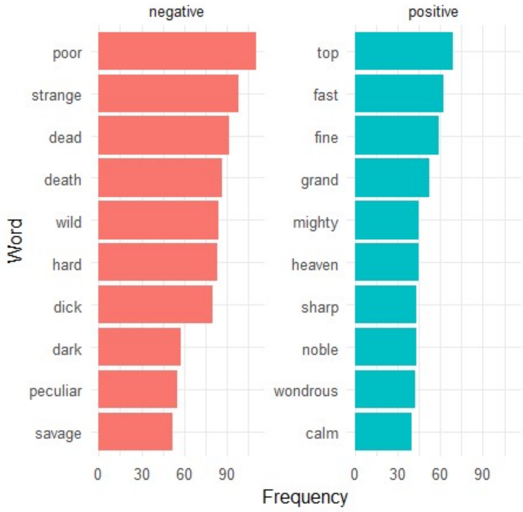


Fig.4. Top 10 positive and negative senitments in adult data

Firstly, we will use the lexicon called bing, which gives the words with positive and negative sentiments in the data. It will give the sentiments for only those words which are in the downloaded dictionary and match with the words in the dataset used for analysis. Fig.3 and 4 show 10 most frequent words with negative and positive sentiments in both the datasets. It is evident from the plot that the words, poor and top are present in both the datasets. While mock, hastily, mad, and anxiously are the other most frequently occurring negative words in child data, strange, dead, death and wild are the most frequent negative words in adult data. On the other hand, beautiful, majesty, glad, and ready are among the positive words in child data and fast, fine, grand, mighty, and heaven are part of the positive words in adult data. The frequency of negative and positive words is greater in adult data; however, it is notable that the size of adult dataset is also significantly greater than the child dataset.

The lexicon named afinn gives the sentiment scores ranging from -5 to 5, where negative values imply negative and positive values imply positive words, in addition, -5 being the most negative and 5 being the most positive. The boxplots in fig.5 and 6 show the distribution of positive and negative sentiment scores for both child and adult datasets. The maximum and minimum sentiment scores for child data are given by -3 and 4 respectively. On the other hand, these scores are given by -5 and 5 for adult data. The median score is observed to be -2 and -1 in child and adult data, respectively. As per the sentiment scores given by the afinn lexicon, adult data has stronger positive and negative words, compared to child data.

Fig.5. Sentiment scores for child data

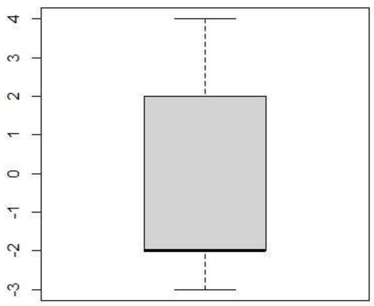
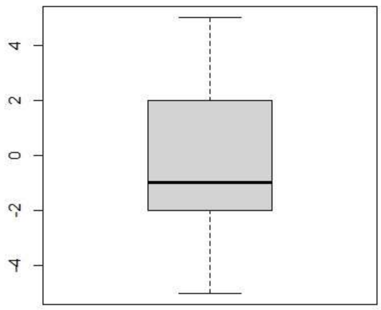


Fig.6. Sentiment scores for adult data



Finally, the lexicon nrc is used to analyse other sentiments like anger, joy, fear and so on in both the texts. The lexicon dictionary gives the sentiments for 855 and 4120 words in child and adult datasets, respectively.

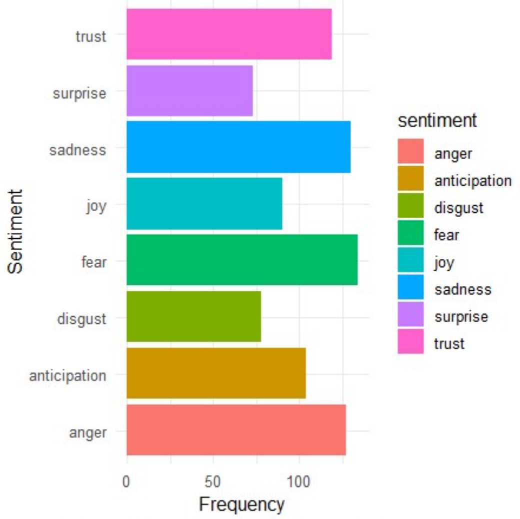


Fig.7. Frequency of sentiments in child data

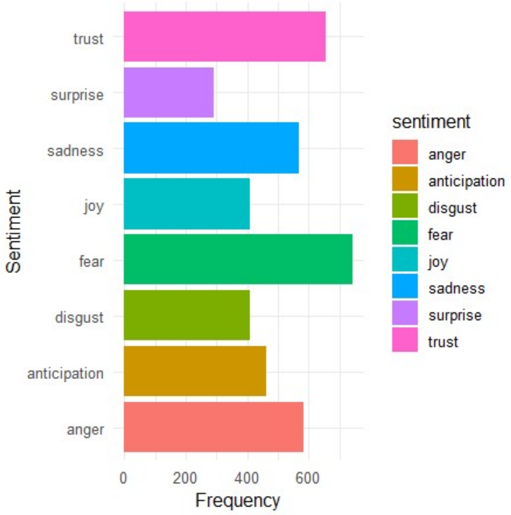


Fig.8. Frequency of sentiments in adult data

The fig.7 and 8 shows the frequency of each sentiment recognised by the nrc lexicon in both the datasets. It is evident that fear is the dominant sentiment in child data, followed by sadness, anger and trust. The sentiments, surprise and joy are the least in child data. It is notable that fear is also the predominant sentiment in adult data, followed by trust, anger and sadness. In addition, surprise and joy have the lowest occurrence compared to other sentiments in adult data as well.

Results

The length of the child data was as expected and found to be lower than the adult data, with only 2192 words after removing the stop words. Whereas, it was 16,466 for adult data. Further, from the sentiment analysis, it was observed that both the texts, child and adult have almost similar sentiments. As shown in table.1, the child data had 1% higher negative sentiment than the adult data. In addition, it was interesting to note from the sentiment scores that the adult data had stronger positive and negative words compared to child data.

Table.1 Count and proportion of positive and negative sentiments

Sentiment	Child data		Adult data	
	Count	Prop	Count	Prop
Positive	141	0.34	815	0.35
Negative	275	0.66	1532	0.65

Further, from table.2 we can see that the percentage of other sentiments extracted using nrc lexicon. The proportion for each sentiment was calculated as a fraction of frequency of each sentiment out of total number of words detected by nrc lexicon. It was notable that the largest 3 sentiments associated with child data are fear, sadness and anger. On the other hand, fear, trust and anger were the largest 3 sentiments associated with adult dataset. In addition, the percentage of joy and surprise was observed to be greater in child data, compared to adult data. In addition to that, anger and sadness is also high in child data.

Table.2 Percentage of sentiments in child and adult data

Sentiment	Child data	Adult data
Anger	14.85	14.13
Anticipation	12.16	11.21
Disgust	9.12	9.93
Fear	15.67	18.01

Sentiment	Child data	Adult data
Joy	10.53	9.95
Sadness	15.2	13.76
Surprise	8.54	7.06
Trust	13.92	15.94

Discussion

As assumed in the beginning of the analysis, the child text, “Alice’s Adventures in Wonderland”, has lesser words than the “Moby Dick”, adult text. On the other hand, the assumptions made, were proven to be partially true in this study. It was found that both the texts have almost same sentiments. In fact, child text has a slightly higher percentage of negative emotion, in comparison to adult text. While the child data was observed to have higher percentage of joy and surprise, it also had a higher percentage of anger and sadness.

The analysis of sentiments is limited to only the words in the dictionaries, which may not include all the words, which can restrict our study. Further, given more understanding of detailed sentiment analysis will aid in performing more meticulous study and producing better results and conclusions.

Bibliography

[1] https://www.tidytextmining.com/sentiment.html