Report on Comparative Sentiment Analysis of ‘Heidi’ and ‘Emma’

Introduction

Sentiment analysis is the contextual exploration of text , to identify patterns and extract subjective information related to text [1]. This report aims for the same between two classic literary works , and see differences or similarities between them. The key difference between the selected titles are the intended audiences, i.e one is a children’s book and the other ,adults. For the children’s book , it is ‘Heidi’ , by Johanna Spyri , in 1881, and the adult’s title, is ‘Emma’ , written by Jane Austen published in 1815.

Heidi is about an orphan girl who is sent to live with her Grandfather , who is described as cold and distant , living alone on top of a mountain. Heidi’s innocent and lively aura , gradually melt’s the old man’s heart, She has a tumbling journey during her time away for work, and is only healed when she returns home to him and the mountain hill. Although a children’s writer , Spyri’s characters were constantly challenging notions. She designs her characters out of stereotypes. Heidi too has a simple life, charming yet notorious child.Notable themes are Discussions about Christianity ,Restoration and healing through nature.
Emma ,is set in the Regency period in London, and it follows Emma , the beautiful , witty and wealthy , and who belongs to upper classes of rural London. The story follows the toils and turns of human relationships, as Emma embarks on her journey of matchmaking.
Austen, heroines were head-fast , strong , independent women ,often living exciting adventurous lives , much like her. Notable themes are intimate real human relationships of all kind , discussions about importance of marrying wisely and witty characters.

Why these titles? The primary reason is the similarities of the author’s general themes in their work. Both the authors have written feminist literatures that challenged norms but are kept lighthearted. The report aims to compare the sentiments of both and also to compare the words used for few sentiments , for different audiences. The report takes into context three main things ,the audience of each title, i.e , adults and children ,the literary styles of the authors and the general themes in their works.

Methodology

Data - Language for analysis is R .Data is maintained by Project Gutenberg , for both the titles. Note, only the text of the actual literature, that is, the story , and not the others (example chapter names , preface etc) is used. The cleaning and extraction of text is done using the ‘stringr’ package. Data is then converted to a tidy format for ease of analysis.

Sentiment Analysis - The part can be split into following three steps

1. Tokenization - convert text into recognisable , independent unit, here , words
2. Count words - the words are counted for the number of times they occur in each work .This step also involves removal of stop-words, i.e, commonly occurring words that does not contribute meaning on their own , example , the, and etc
3. Sentiment analysis - Libraries of words and the sentiments attached to them are used. Here we use the “nrc” lexicon , that categorises sentiments into 10 categories- anger, anticipation, disgust , and so on. Note the same word could have different sentiment throughout the text depending on the context of the text. We also use lexicon “bing” lexicon that categorises words into negative or positive and a score between -5 to +5 .

Examples of “nrc” sentiments :

word	sentiment	word	Sentiment
abandon	fear	immediately	anticipation
Bouquet	Joy	ungrateful	Anger

Result and Visualisations - The report will use simple column graphs for analysis of sentiments in each work , and further of two opposing sentiments , joy and fear. We look at the kind of words used to express each sentiment and comment at the choice respective to both the works. We also look at the positive too negetive ratios in both the works , assessing through a column graph after a “bing” lexicon analysis.

Results

The data present in the tidy format with the word count gives us the most frequent words from both the works. A sample of top ten words from each works is as following :
Child (Heidi)

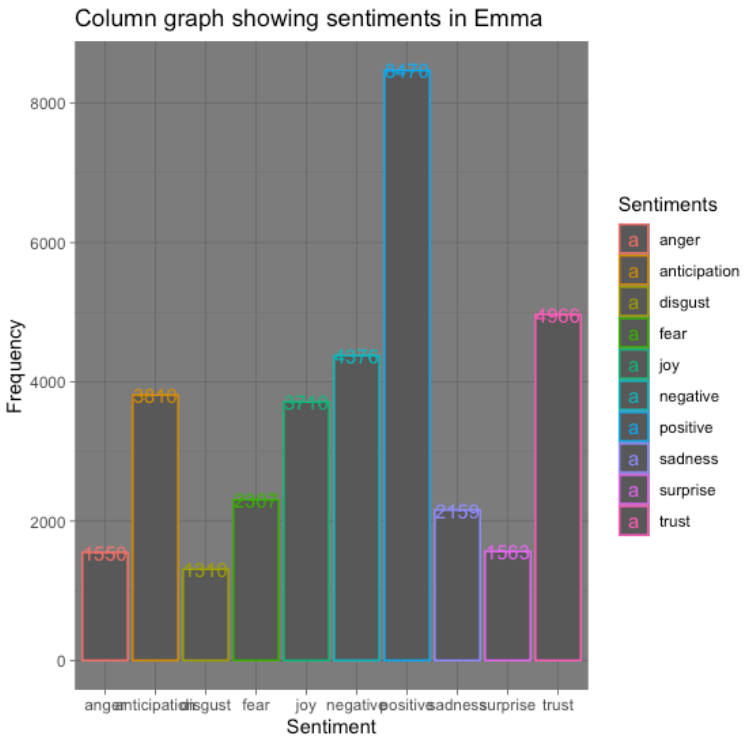
	Word	Frequency	Word	Frequency
1	Heidi	917	grandmother	238
2	Peter	335	day	237
3	child	290	time	195
4	grandfather	279	Mountain	175
5	Clara	270	Looked	149

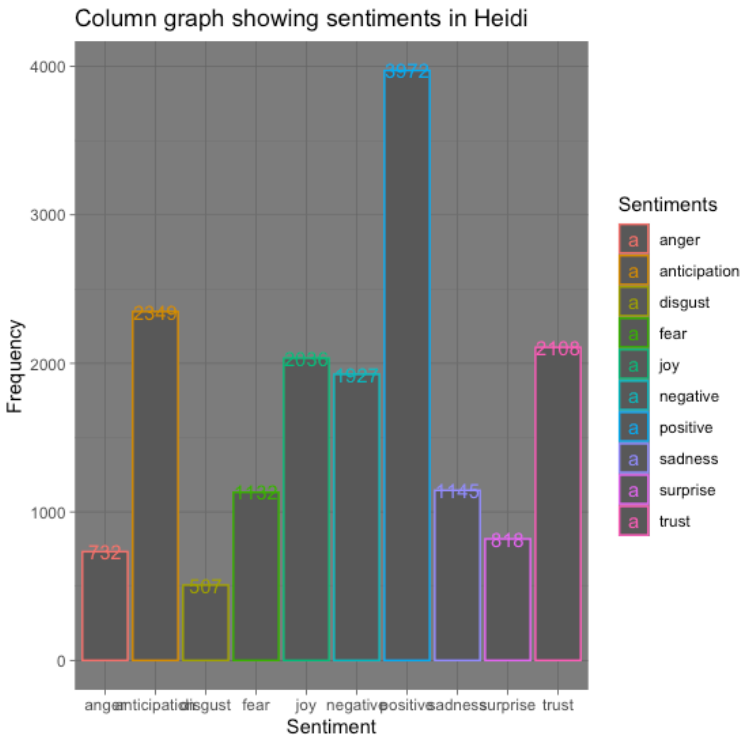
Adult (Emma)

	Word	Frequency	Word	Frequency
1	Emma	785	Elton	319
2	miss	599	Jane	281
3	harriet	415	time	279
4	Weston	389	woodhouse	277
5	Knightley	356	Dear	241

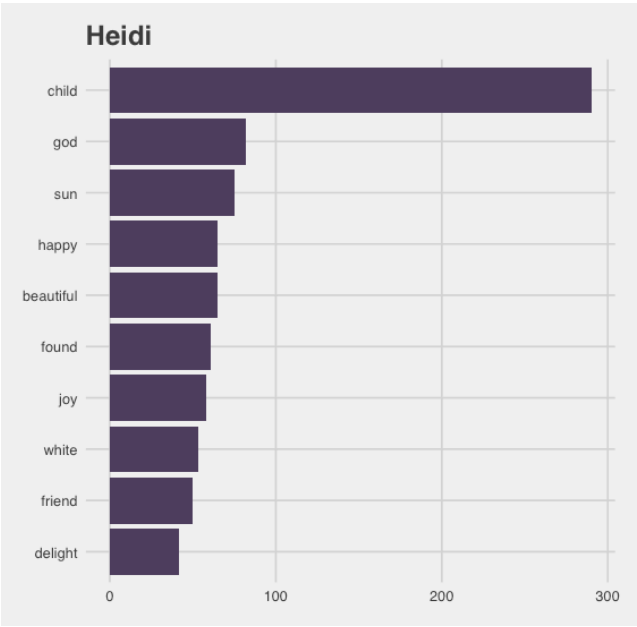
The words in both the works are not a bit surprising as most of them are character names , which is expected to be used frequently. Also the words give a picture about the kind of scenery the stories are set in. Mountain ,day and time along with the character names , say a lot about the story itself. In Emma , Dear and Woodhouse gives an idea about the period of the setting , i.e , Regency period , where words like ‘Dear’ were used for addressing someone, and well , wood houses are no longer present ! Quite indicative of the time! Something fun to notice is ‘time’ is present in both the tables.

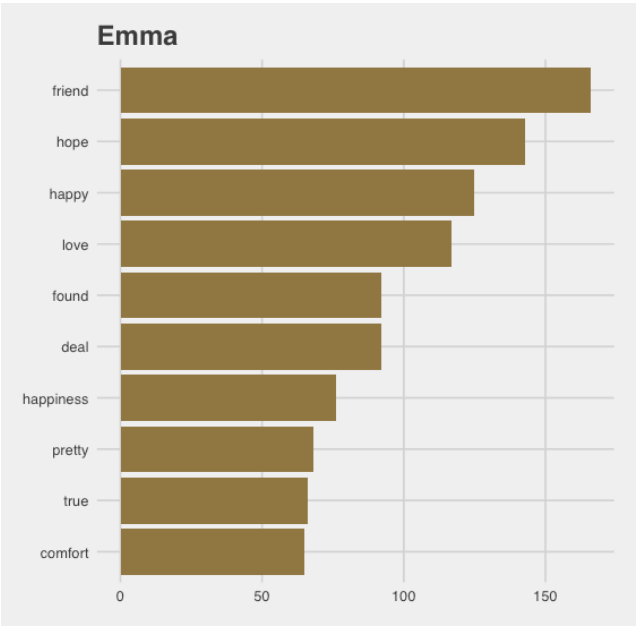
The sentiment analysis for the words , matched against words in the lexicon “nrc” gives us the following column graphs(next page). A very obvious feature from the visuals are that they almost look identical , however they are not. This can be owed to Austen’s literary style , that is witty , humorous , and dealing with the the topic of romance and human relationships , much like the themes in a children’s book , joy , adventure etc . Although the numbers are highly skewed in terms of the frequencies of each sentiment , this can be owed to the differences in the sizes of the respective books. Heidi has 26,761 tokenized words (stopwords excluded) while Emma has 46,666 words! However the graphs are better ways to look at the proportion of each sentiment. ‘Positive’ is the highest in both. ‘Joy’ ,’trust’ and ‘anticipation’ slightly higher in Heidi. The ‘negative’ is slightly higher in Emma , however clear margins cannot be drawn.
Owing to the high similarity in the sentiments of both the works , we next take a look at the top words used in few of the sentiments .



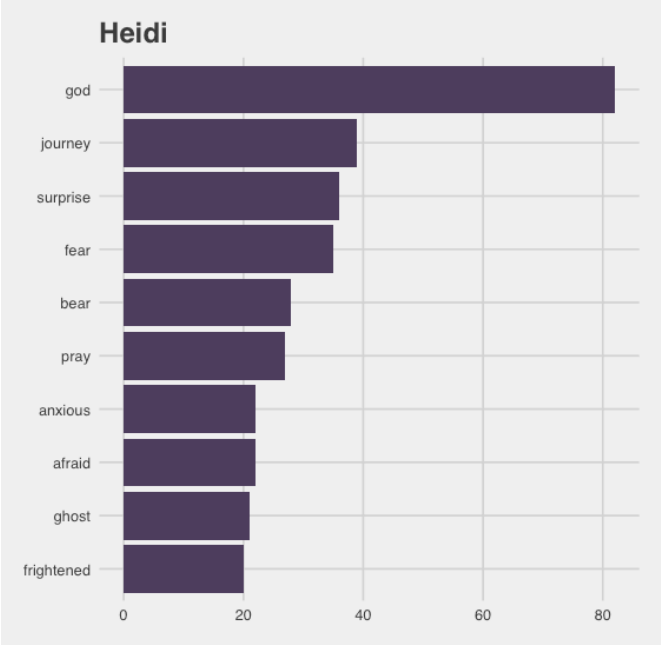


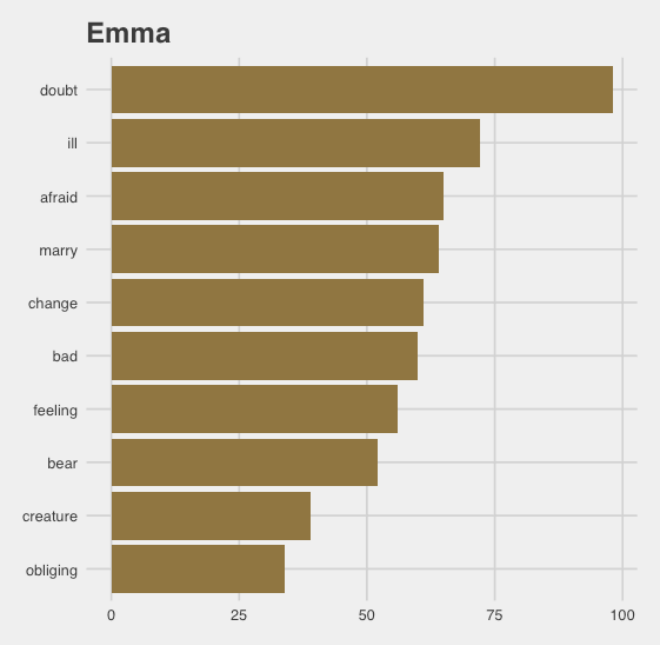
Joy : The sentiment 'joy' seems reasonably high in both the works . The most commonly used words in both the works are shown below .For Heidi , the words are indicative to the central theme of a child's book and also to the story specifically. Words like 'sun' , 'happy' , 'delight' , 'joy' , 'friend' , are words expected in a work meant for children , and words like , 'god' , and 'child' are indicative to the story itself which revolves around these themes . Note how 'god' comes in sentiment joy , showing it has been used in a positive connotation throughout , through discussions about Christianity. In Emma however , the words take a more romantic notions of joy like , 'friend' ,'hope' , 'love' , 'pretty' ,'comfort' etc.





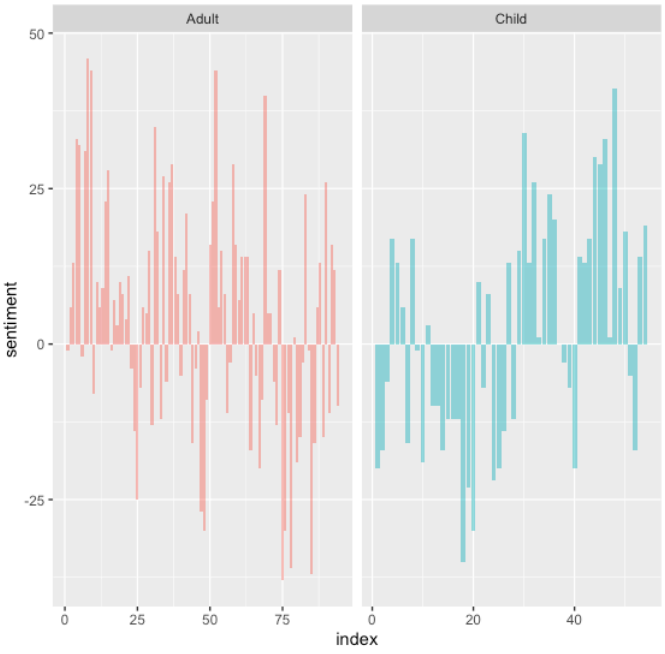
Fear : Fear seems to be almost in the same proportion in both the works. We look at the words used to describe them.





In Heidi , the word ‘god’ occurs the highest number of times with sentiment , which is again testament to Spyri’s discussion of religion. Other words are ‘ghost’ , ‘journey’ , ‘pray’ are indicative to the story. The words expressing are quite general and a range of emotions a child is likely to experience sometime through childhood. In Emma however, the words take a more serious tone with words like , ‘marry’ , ‘ill’ , ‘doubt’ , ‘change’ , ‘obliging’ , which describes the toils experienced by adults in later years of life. These words are both general to adult’s book themes and specific to ‘Emma’.

Postive vs Negative :
The sentiment analysis using “bing” lexicon gives us the following graph for the two books (at an interval of every 500 words-this is the index). No clear margins, but affirming the sentiment analysis with “nrc” , the positives seem to be slightly more in number. The numbers are as follows.
Adult : Positive - 5160 , Negative - 4729
Child : Positive - 1957 , Negative - 1849



Discussions
We looked at the sentiment analysis of two works intended for different audiences. Although initially I expected marginal differences in the sentiments of both, this was proven to be wrong , as both the books have similar sentiments. This is owed to Austen’s literary style , that talks about deep subjects but often is kept humorous and lighthearted , and also Spyri’s attempt to keep her characters as real as possible for an audience of children. The analysis however shows that as we transition to adulthood , the notions of the sentiments changes , like the difference between what was ‘fear’ for a child, differed to what was to an adult. However, it is commendable , the balance of sentiments in both the works , since they deal with the matter of life and human relationships.

References

[1] Sentiment analysis , Concept analysis and Applications
https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17
[2] Text mining with R : A tidy approach
https://www.tidytextmining.com/tidytext.html
[3] Sentiment Analysis of Harry Potter Book Series using R : Divyesh Patel
http://www.j-asc.com/gallery/66-november-1233.pdf
[4] Johanna Spyri :Heidi the Girl from the Alps
http://heidi-children-story-books.all-about-switzerland.info/
[5] Emma by Jane Austen | Summary and Analysis, by Course Hero (YouTube)
[6]Jane Austen -- an 18th century woman for the 21st century | JoAnne Podis | TEDxUrsulineCollege (YouTube)