MA331 Coursework

Sayali Kambli

30/03/2021

Introduction:

Main purpose of this report is to perform sentiment analysis on two novels from Project Gutenberg Collection,each one from Child and Adult Category[1].Following are the novels I have selected for analysis.

1.Child Category : *Rebecca of Sunnybrook Farm* - It is a children's novel by Kate Douglas Wiggin, published in 1903, which tells the story of Rebecca and her aunts[2].

2.Adult Category : *Great Expectations* - It is the thirteenth novel by Charles Dickens,published in 1861 ,which depicts the education of an orphan called Pip[3].

Sentiment analysis is expected to show some similarity or difference between how adult and child novels keep their audiences engaged.

Methods:

1. Analysing data

Both data-sets have columns Id and text.Text column has text data from novel including Name of the novel,Author and Chapters.I have used *stringr* package to filter chapter count.

Great Expectations from Adult Category has `20024` rows and `59` chapters. Rebecca of Sunnybrook Farm from Child category has `8033` rows and `31` chapters.

2. Cleaning data

I have removed Id column as it is not required.Then I have found out where both data-set's actual text starts and ends and extracted that part only.As sentiment analysis will be performed,I have also removed empty lines, numbers and non-characters.

3. Extracting Words

Data-sets need to be divided into words to perform sentiment analysis. I used *unnest_tokens* function from *tidytext* package to extract word tokens from both novels. Following table shows top 4 used words from both novels.

ChildBookWords	ChildBookcount	AdultBookWords	AdultBookcounts
the	3993	the	8145
and	2679	and	7098
to	1897	i	6483
of	1798	to	5154

As most commonly used words are stopwords and not needed.

4. Removing StopWords

I used *stop_words* database from *tidytext* to remove stopwords from my dataset. After removing them,I found most commonly 4 words used.

ChildBookWords	ChildBookcount	AdultBookWords	AdultBookcounts
rebecca	571	joe	691
jane	276	miss	383
miss	176	time	373
aunt	173	pip	326

These most common words from both novels are character's names or what they are called.

5. Sentiment Analysis

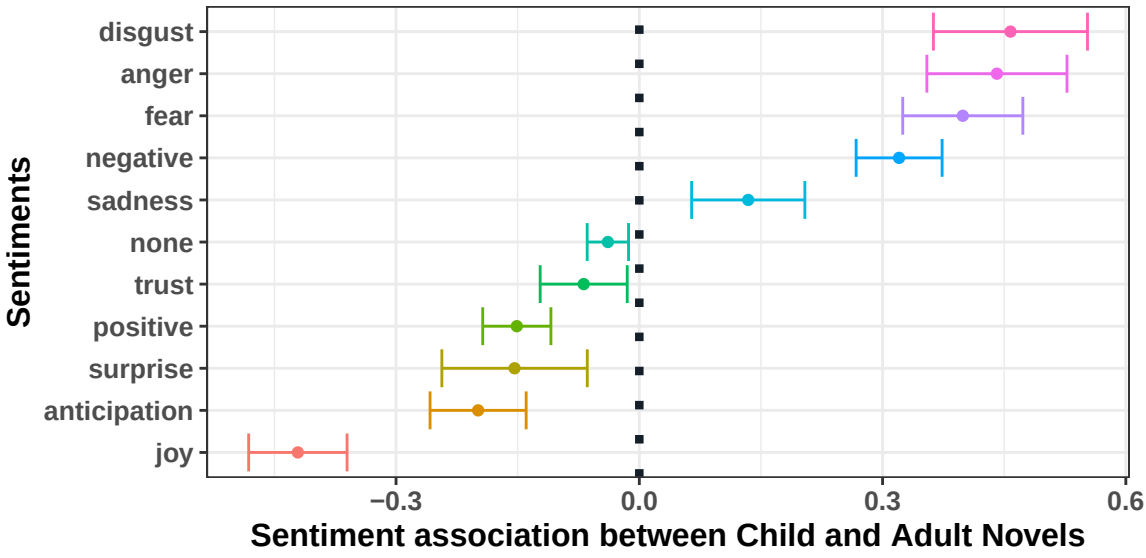
In sentiment analysis each word is assigned a sentiment like anger,joy or positive.It helps analyze the emotions from the text for human understanding. For sentiment analysis, I have used following lexicons from *textdata* package.

1. *nrc lexicon* to give various sentiments: Here I have compared the frequencies of each sentiment appearing in each novel[4].Then I have calculated a confidence interval for each sentiment in both novels(Adult and child) using log(OR) Where OR is Odds Ratio.[5]

2. *bing lexicon* to assign words into positive and negative sentiments. Here I have used net sentiment(positive-negative) to plot these sentiment values over the time period of each novel[6].

Results:

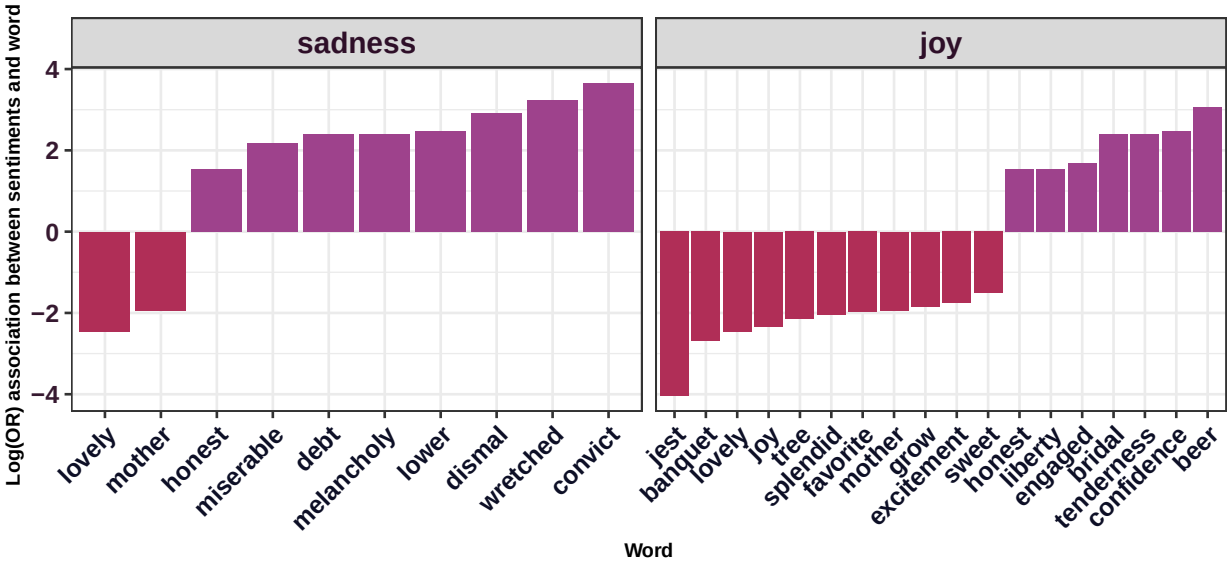
Sentiments association between both novels



Graph shows that anger,disgust,fear,negative and sadness sentiments are associated with adult novel. On the other hand, the joy,positive, surprise, and anticipation sentiments are associated with the child novel. Trust and none emotions are close to both novels.

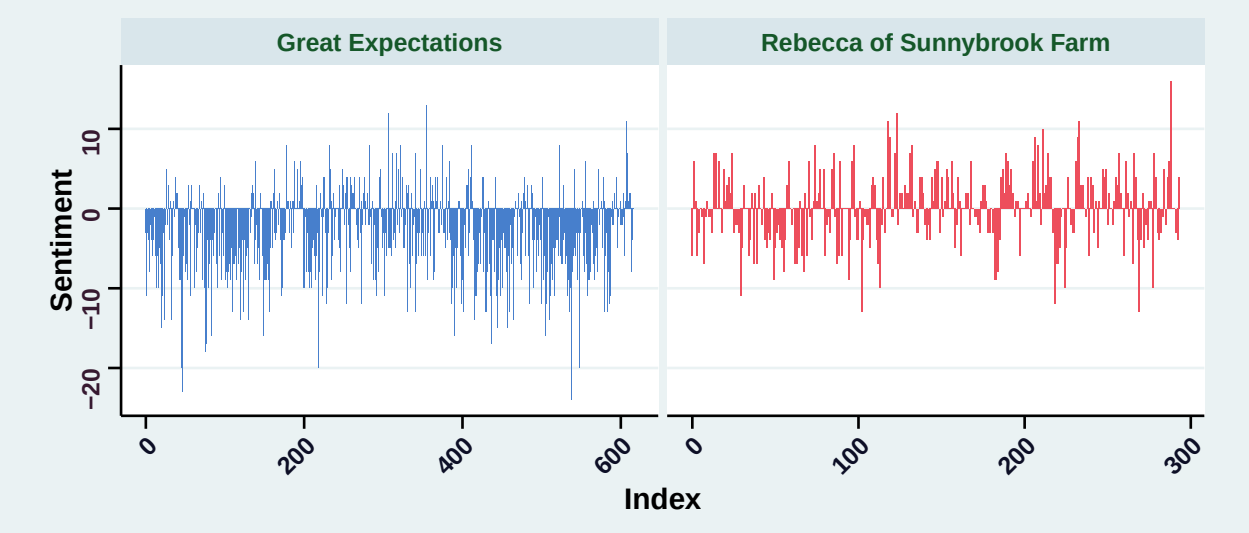
Comparing Sadness and Joy in novels

I have shown different words used for 2 sentiments in both novels. In the case of sadness more words are in adult novel. On the other hand more joyful words are in child novel.



Positive-Negative sentiments analysis over the plot of novels

I have created graph which shows sentiments(positive-negative) analysis over the time period of novels, Where Index keeps track of 90 words section of text.



As clearly seen, Adult category book has very few positive sentiments compared with negative sentiments. Whereas child category has more Positive sentiments than negative sentiments.

Discussion :

Main challenge of my work was that I was comparing 2 novels from different categories and also having different length of words.

My main findings are as follows:

* Author Charles Dickens has used more negative sentiments in Adult novel Great Expectations to show real world problems and tragedies which will make adults related with.

* Author Kate Douglas Wiggin has used more happy sentiments in her Child novel Rebecca of Sunnybrook Farm like Joy,Surprise to attract and motivate children.

* Trust and none emotions are close to both novels.

So I want to conclude that both novels are expressing well targeting their audiences.

References :

I have used following References in my report.

[1] - *http://data-science.essex.ac.uk:3838/MA331*

[2] - *https://en.wikipedia.org/wiki/Rebecca_of_Sunnybrook_Farm*

[3] - *https://en.wikipedia.org/wiki/Great_Expectations*

[4],[5] - *http://data-science.essex.ac.uk:3838/MA331/W9/#section-iii.-sentiment-analysis*

[6] - *https://www.tidytextmining.com/sentiment.html*