MA331_coursework_Project

Text analytics of the TED talks by Sherwind Nuland and Ben Saunders

2111168_Ahmad_Cheema

18/11/2022

Introduction:

This course work includes dataset of "ted_talks" including speakers with their given talks on different times, every student is assigned of two speakers and our task is to present and compare the word frequency and sentiment analysis of transcripts of both ted speakers. The talks of ted speakers are purified from data set "ted_talks" and in this course work my allocated two speakers are "Sherwin Nuland" and "Ben Saunders". Ted is a platform or a hub where we can find information of the Teds where world leading thinkers are asked to give their talks of their life experiences. We have two ted talks of each speaker, "Sherwin Nuland" topics are "The extraordinary power of ordinary people" given in February 2003 and "How electroshock therapy changes me" given in February 2001. Likewise, "Ben Saunders" also has two topics which are "Why did I ski to the North Pole" given in February 2005 and "To the south pole and back-the hardest 105 days of my life" which was given in March 2014.

Word frequency is done by using tidy data theory which is used to handle data easier and more effective, and it includes packages such as dplyr, ggplot, tidy verse and tidytext. Tidy text package does not allow us to keep text data in tidy form during an analysis. In Sentiment analysis, there are so many methods exist for estimating the emotion in text. We are using three lexicons which are briefly explained in methods, these three lexicons are based on single words and contain many English language words assigned to positive and negative sentiments possibly like fear, anger, surprise and anticipation. We will see how far these sentiments are evident in the dataset of two talks of both speakers. The objective of this study is to examine whether the assumptions mentioned above are true by comparing the two-ted speaker's dataset.

Let's find out if both texts express similar or different sentiments.

Methods:

As we are asked to perform word frequency and sentiment analysis, at first step we are going to load required packages in R studio of dsEssex and tidyverse which are here used to load dataset ted_talks and then we will filter our two ted speakers and then load packages and libraries of dplyr, textdata, ggplot2, stringr, tidyverse and tidytext. We will perform analysis on both ted talks of speakers and then present and compare outputs of frequency of words and sentiment analysis. We are naming here "ted1" as "Sherwin Nuland" and "ted2" as "Ben Saunders" for the better compatibility in results. Then using stop words lexicon which is a function used to return vector of stop words in a language and here to filter stop words using snowball lexicon among tokens in ted1 & ted2 dataset to visualise frequent occurring words. We used slice max function on both ted1 and ted2 speaker's dataset to select the top rows ordered by n. In string check, Using mutates here to override the existing columns or bars in both ted1 & ted2 speakers' dataset. After getting summary of both speakers, we will head straight to ggplot2 function because we are using tidyverse tools here. Both ted1 and ted2 datasets are prepared for sentiment analysis by applying these all above steps.

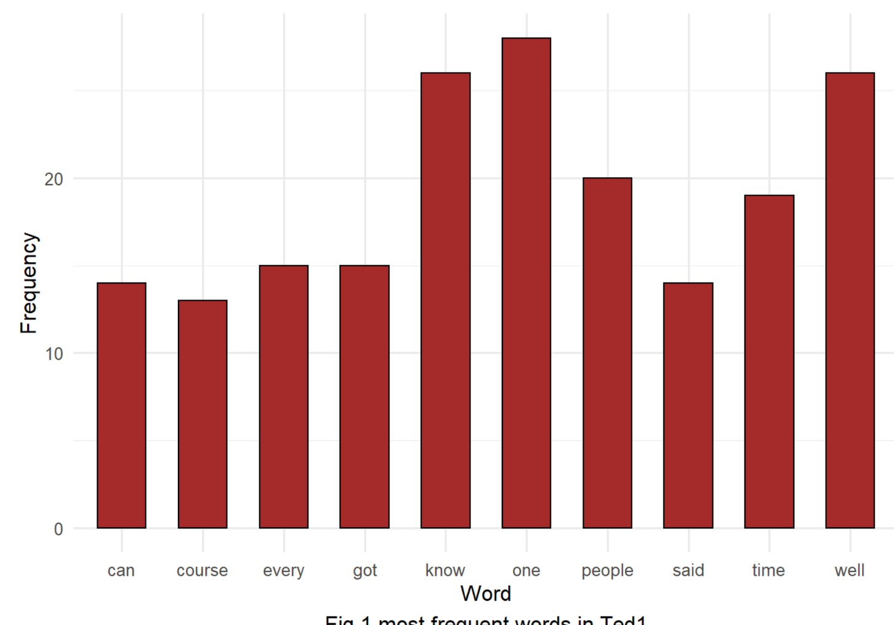

Fig 1 most frequent words in Ted1

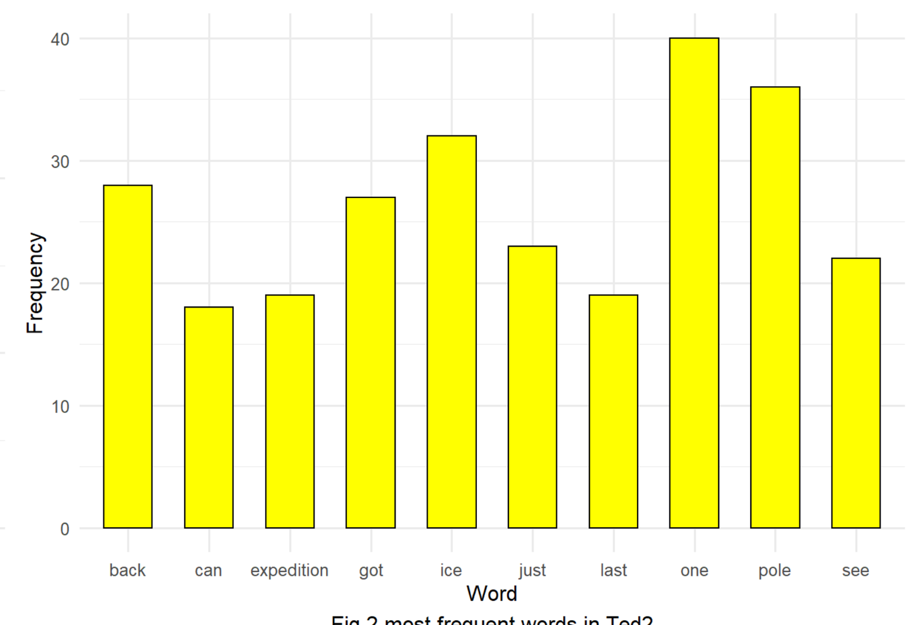

Fig 2 most frequent words in Ted2

By visualising, we got 10 most frequent word of ted1 and ted2 as we can see in Fig 1 and Fig 2. We are analysing number of words in both ted1 and ted2 datasets. We are naming here ted1 as Sherwin Nuland and ted2 as Ben Saunders for the better compatibility in results. Ted1 has total of 1128 words and ted2 has 1448 words. Fig 1 and Fig 2 shows the 10 most frequent words in ted1 and ted2 datasets. The most frequent words in ted1 dataset are can, course, every, got, know, one, people, said, time and well. The frequency of the word one is notably high compared to other words in the data, which is 28 times. Likewise, in the ted2 dataset frequent words are back, one, pole, ice, got, just, see, expedition, last and can. The higher frequency word also in this dataset is one which is 40 times comparing to other word in ted2 dataset. These frequent words mention to main events on which talks are based on both speakers.

Sentiment Analysis

Now performing sentiment analysis on both ted1 and ted2 speaker's dataset using some different lexicons which are loaded within the packages are bing, afinn and nrc. Initially downloading the bing which is used to divide words into positive and negative sentiments. Then afinn lexicon which used to assign the score between -5 and 5, like -5 negative and 5 positive sentiments. Eventually to provide several sentiments like to extract sentiments in words we will download and use third lexicon nrc.

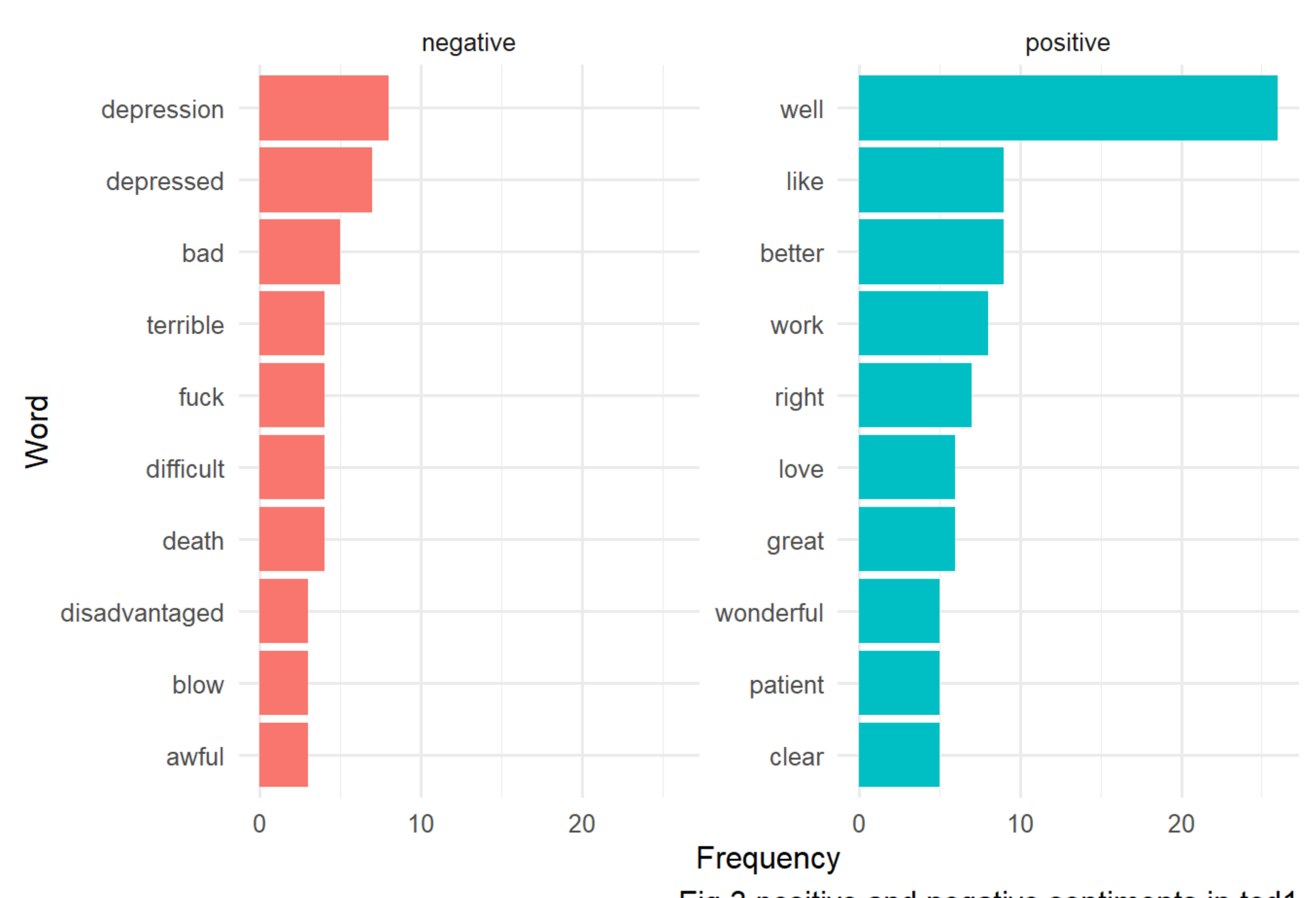

Fig 3 positive and negative sentiments in ted1

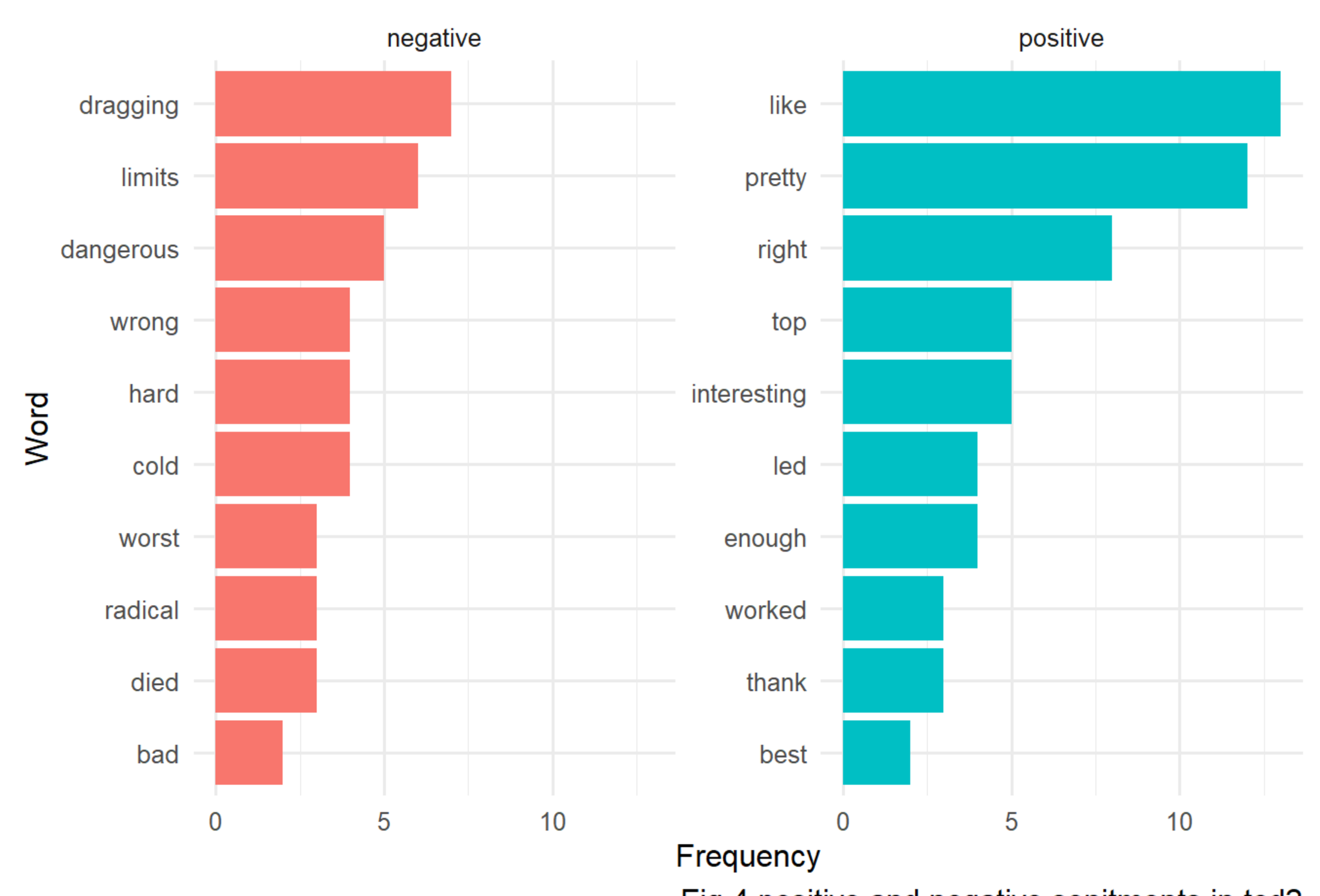

Fig 4 positive and negative senitments in ted2

Figs 3 and 4 show 10 most frequent words with negative and positive sentiments in ted1 and ted2 datasets of speakers. It is evident from the graph that depression, depressed, bad, terrible, difficult, death, disadvantaged, blow and awful are the most frequent negative words in ted1 overall 71 and on the right side well, like, better, work, right, love, great, wonderful, patient and clear are most frequent positive word of ted1 overall 67. While, dragging, limits, dangerous, wrong, hard, cold, worst, radical, died and bad are the negative words in ted2 overall 99 and second graph showing like, pretty, right, top, interesting, led, enough, worked, thank and best are positive words in ted2 overall 73. As obvious frequency of positive and negative sentiments is greater of ted2 than ted1 speaker dataset.

Results

In ted_talks dataset we have chosen two speakers who are called as ted1 and ted2 in this coursework report and as mentioned above ted2 is greater than ted1 data. By presenting and comparing word frequency and sentiment analysis, it was noticed in both ted1 and ted2 dataset that they have almost similar sentiments. In sentiment analysis, all three lexicons which are used presented the results depending on both ted1 and ted2 dataset. Bing lexicon divided the positive and negative sentiments after that afinn lexicon stated the score between positive and negative sentiments and nrc provided several sentiments as shown in Fig 7 and Fig 8. It is engrossing to see the sentiment scores that ted2 has higher positive and negative words compared to ted1 dataset.

Summary of Ted1

```
##      word                 n
##  Length:20          Min.   :11.00
##  Class :character   1st Qu.:11.75
##  Mode  :character   Median :13.00
##                     Mean   :15.45
##                     3rd Qu.:16.00
##                     Max.   :28.00
```

Summary of Ted2

```
##      word                 n
##  Length:20          Min.   :11.00
##  Class :character   1st Qu.:11.75
##  Mode  :character   Median :13.00
##                     Mean   :15.45
##                     3rd Qu.:16.00
##                     Max.   :28.00
```

Sentiments of Ted1

```
##
## negative positive
##       71       67
```

Sentiments of Ted2

```
##
## negative positive
##       99       73
```

Fig 5 Sentiment scores for Sherwind Nuland

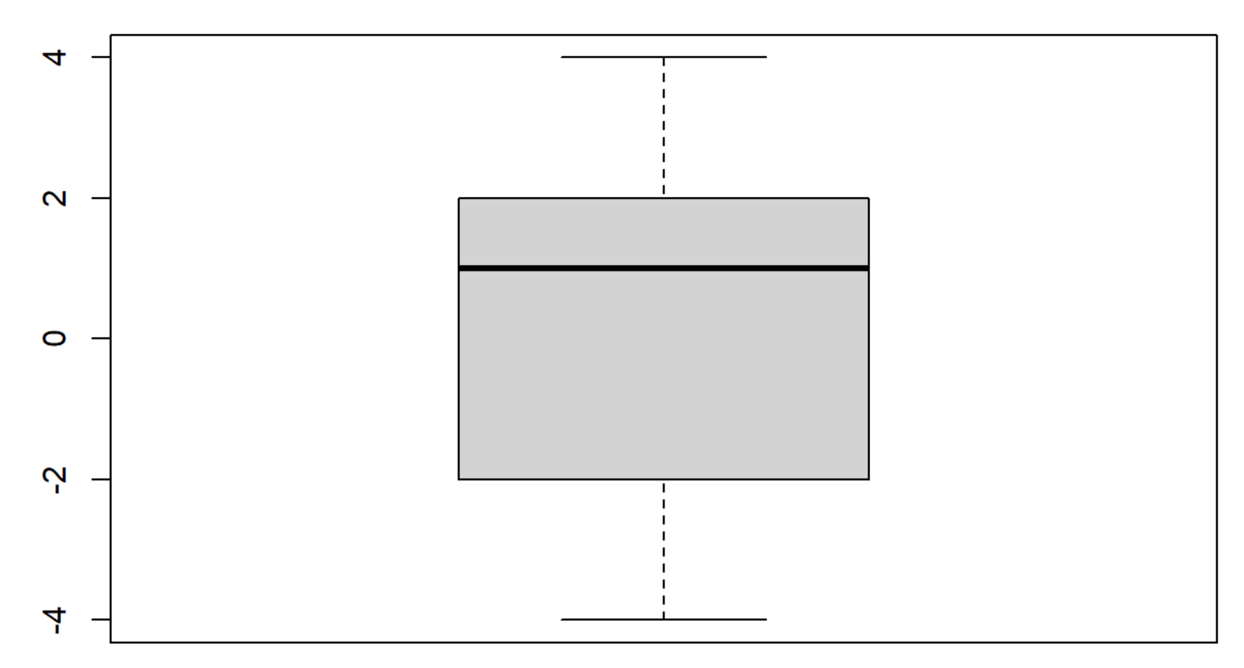

Fig 6 Sentiment scores for Ben Saunders

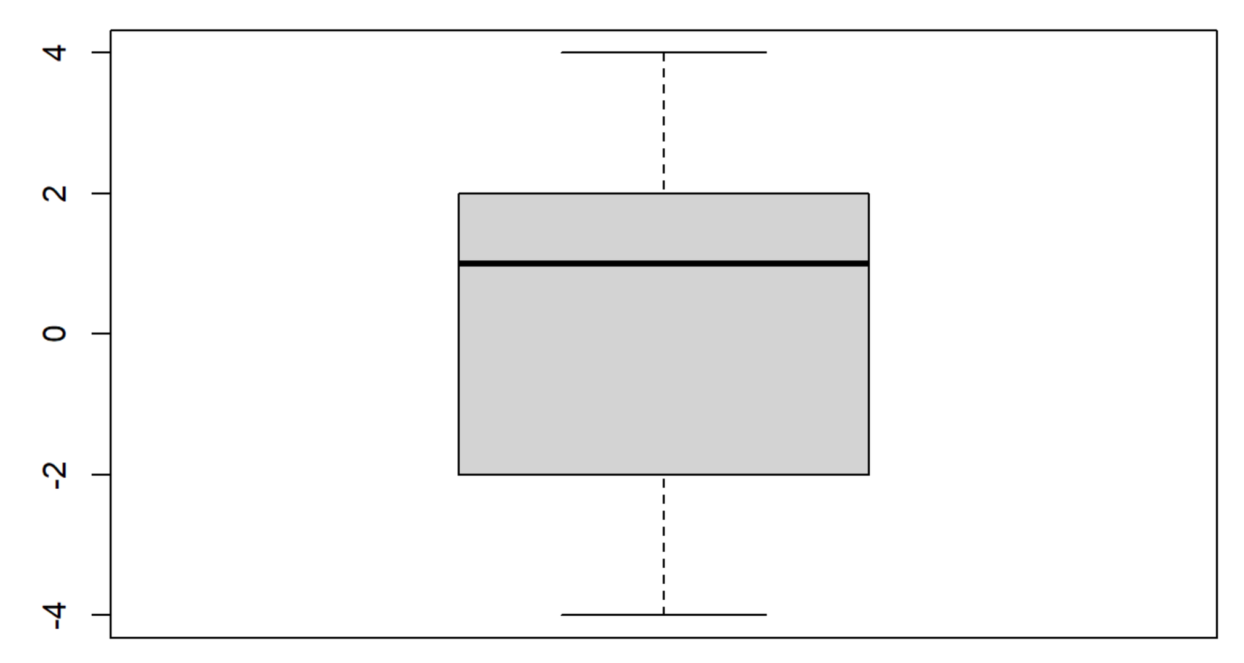

Fig 5 and Fig 6 shows the dispersal of negative and positive sentiments of both ted1 and ted2 speakers' dataset. Sentiment scores for ted1 data are given by -4 to 4 and the median score is observed to be between 0 and 2. Secondly, for ted2 data is given by -4 and 4 and median score is between 0 and 2. Sentiment score given by afinn lexicon is comparatively same in both ted1 and ted2 dataset.

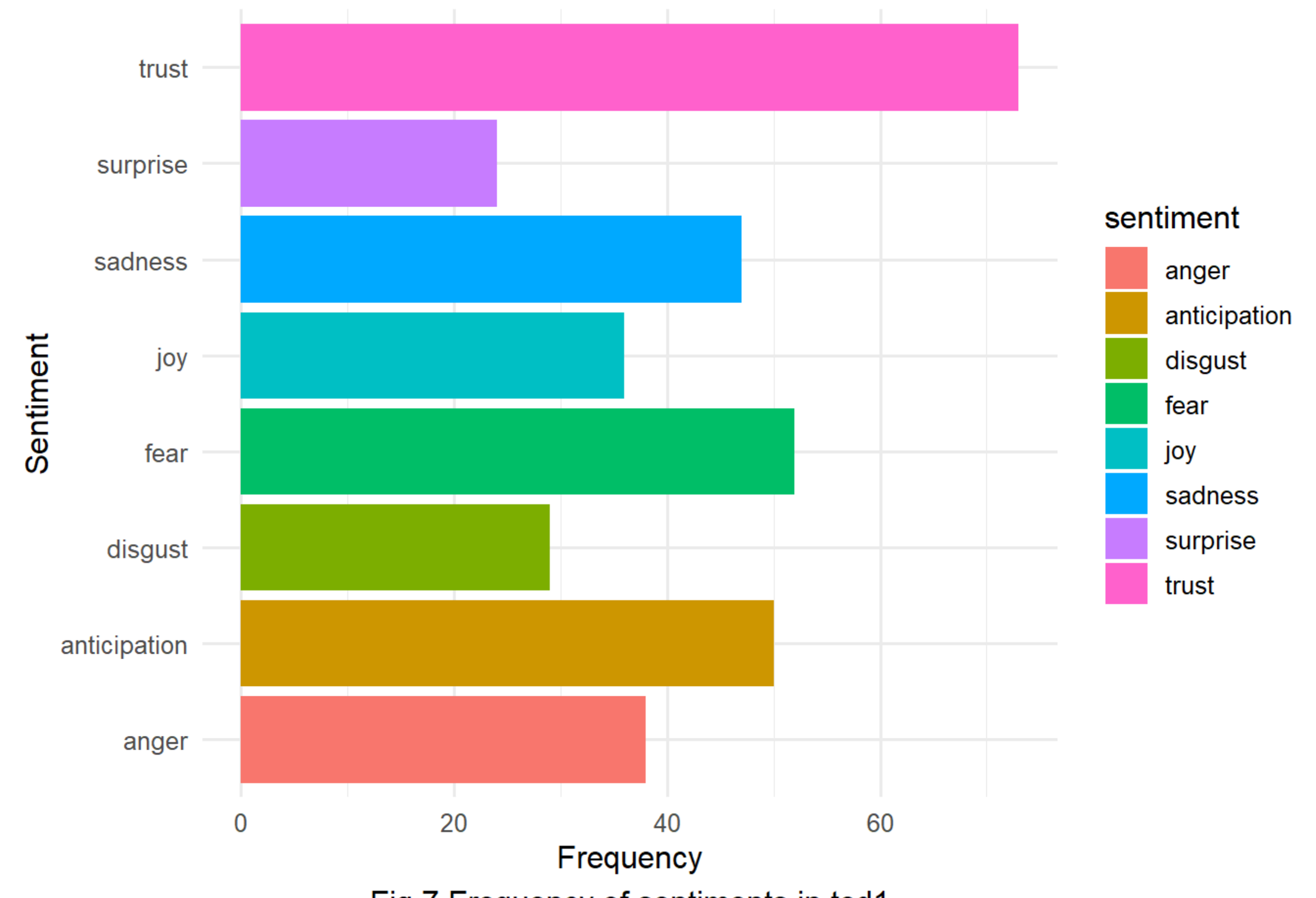

Fig 7 Frequency of sentiments in ted1

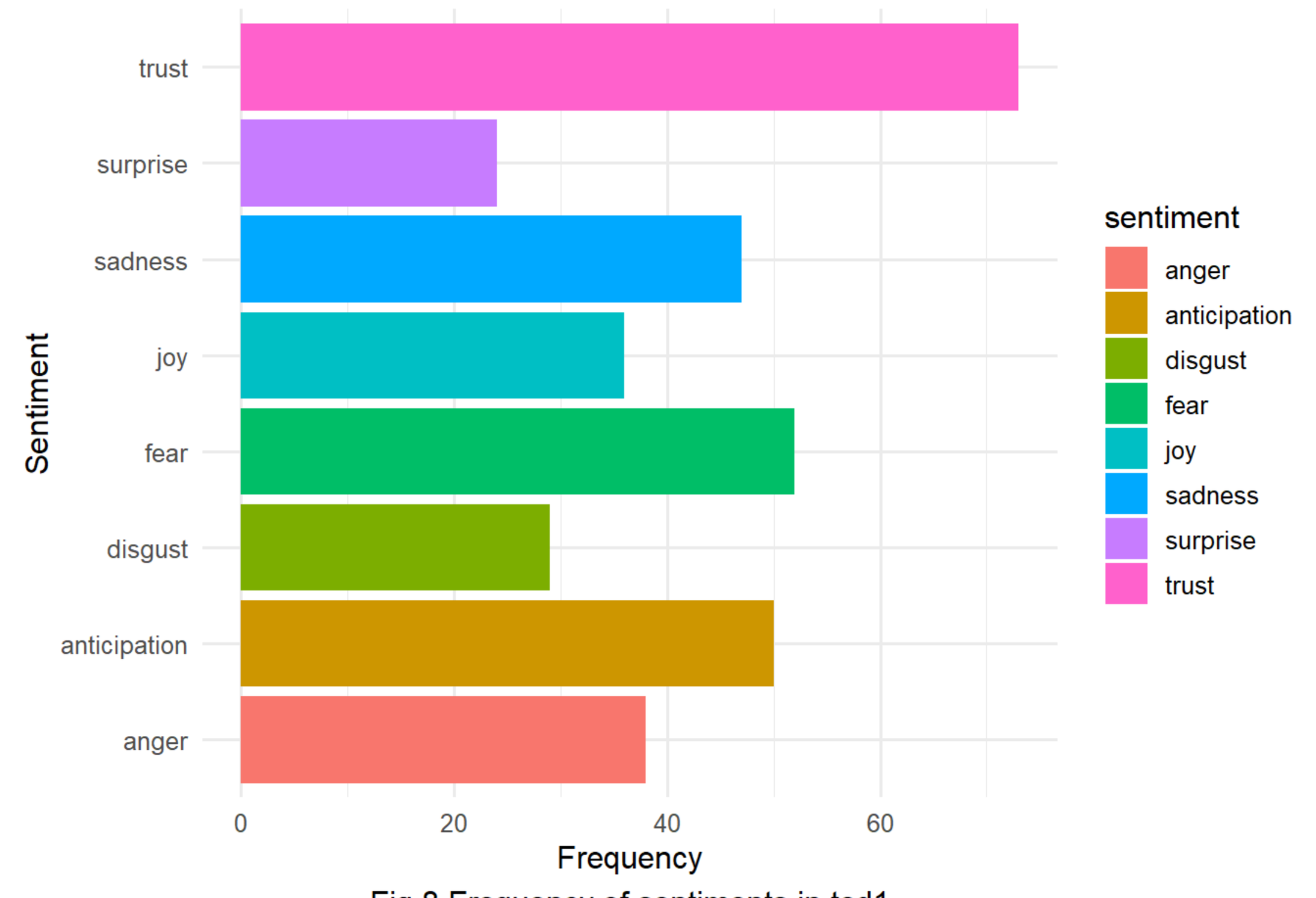

Fig 8 Frequency of sentiments in ted1

Fig 7 and Fig 8 shows the frequency of sentiment given by the nrc lexicon in both ted1 and ted2 datasets. Trust appears to be the dominant sentiment in ted1 data and also predominant in ted2 data as shown in Fig 8.

```
##
##        anger anticipation      disgust         fear          joy      sadness
##           38           50           29           52           36           47
##     surprise        trust
##           24           73
```

```
##
##        anger anticipation      disgust         fear          joy      sadness
##           33           69           25           45           52           39
##     surprise        trust
##           33           75
```

Discussion:

Sentiment analysis provides us different ways to understand the point of view expressed in these talks of speakers. We surveyed how to approach sentiment analysis using tidy data principles and many other packages used in this coursework. Initially we stated, the ted1 data taken from dataset ted_talks have lesser words than ted2. It was found that both the texts have almost same sentiments and words. Ted2 has higher negative emotion in talks and the positive sentiments than ted1. Due to limited words in the dictionary our sentiment analysis is also restricted because all words are not included. Sentiment analysis provides us different ways to understand the point of view expressed in these talks of speakers.

Bibliography

1 The tidy text format | Text Mining with R (tidytextmining.com)

2 Sentiment analysis with tidy data | Text Mining with R (tidytextmining.com)