MA331-Report: 2211606

TED Talks by Speaker Alan Eustace and Speaker Tshering Tobgay

Ghosh Hansda, Sourav

## Introduction

This report is an exploratory analysis of two TED talks, one each by Alan Eustace and Tshering Tobgay. Alan Eustace’s speech took place on October 21, 2014. He is a former Google executive. Tshering Tobgay delivered the speech on April 1, 2016. He is a Bhutanese politician who served as the Prime Minister of Bhutan from 2013 to 2018. The objective of this report is to identify common themes in their speeches and conduct a sentiment analysis of the words uttered.

## Methods

The data used in this analysis is from the ted\_talks dataset available in the dsEssex package in R. The dataset contains information about various TED Talks, including the speaker, the title of the talk, and the transcript of the speech.The data preparation process involves loading the required packages, dsEssex, tidyverse, tidytext, ggplot2, and ggrepel. So the dataset is filtered to include speeches by the two speakers and then tokenize the text using the “unnest\_tokens()” function from the “tidytext” package. The analysis removes stop words, as well as some words like “laughter” and “applause,” to create a word count table for each speaker. Visualizing the top 25 words used by each speaker is done by using the “slice\_max()”, “mutate()”, and “geom\_col()” functions from the “ggplot2” package. To compare the top words used by both speakers, the report uses “bind\_rows()”, “group\_by()”, “filter()”, “pivot\_wider()”, and “geom\_abline()” functions from “ggplot2” and labels the words using “geom\_text\_repel()” function from “ggrepel”. Performing sentiment analysis uses the NRC lexicon. It uses the “get\_sentiments(”nrc”)” function to obtain the NRC lexicon and joins the lexicon with the tokenized text using “inner\_join()”. The analysis counts the number of words associated with each sentiment for each speaker using “count()”, and pivots the data using “pivot\_wider()” to create a table with columns for each speaker and rows for each sentiment. It calculates the odds ratio and log odds ratio of each sentiment for each speaker using “mutate()”. The odds ratio is calculated as the ratio of the number of words associated with a sentiment to the number of words not associated with the sentiment. The log odds ratio is calculated as the natural logarithm of the odds ratio.The most common positive and negative sentiment lexicons are extracted from a dataset of talks. First, the unnest\_tokens() function is used to separate the text into individual words. The anti\_join() function is then used to remove stop words, followed by the inner\_join() function to match the remaining words with the National Research Council (NRC) sentiment lexicon, which classifies words as positive, negative, or neutral. The resulting words are then filtered to select only those with positive or negative sentiment, and their frequency is counted using count(). Two plots are created to show the most common positive and negative words. The slice\_max() function is used to select the top 10 words, and the fct\_reorder() function is used to order the words by frequency. The plots are created using ggplot(), with geom\_col() to create a bar plot, and coord\_flip() to flip the axes. The x- and y-labels are set using xlab() and ylab(), and the plot titles are set using ggtitle(). Finally, the plots are displayed side by side using grid.arrange().

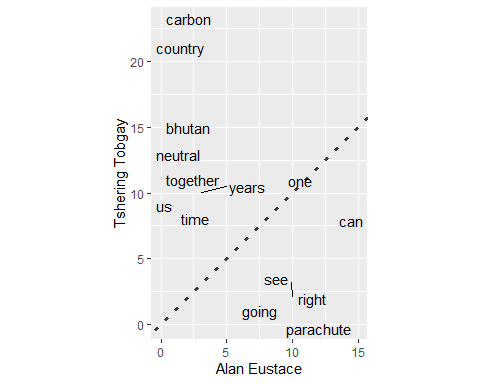
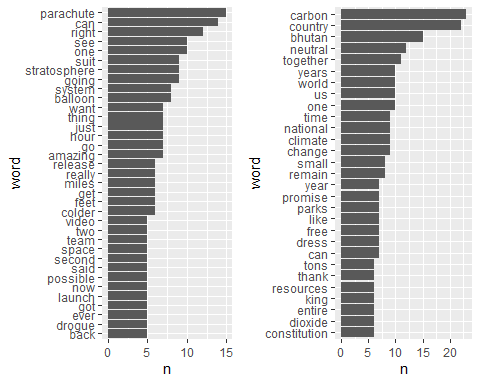
## Results

The tables of Alan Eustace and Tshering Tobgay show the words they used most frequently in their respective talks.Here Alan has used maximum the word ‘parachute’ 15 times whereas Tshering has used ‘carbon’ 23 times.

# A tibble: 476 × 3  
 speaker word n  
 <chr> <chr> <int>  
 1 Alan Eustace parachute 15  
 2 Alan Eustace can 14  
 3 Alan Eustace right 12  
 4 Alan Eustace one 10  
 5 Alan Eustace see 10  
 6 Alan Eustace going 9  
 7 Alan Eustace stratosphere 9  
 8 Alan Eustace suit 9  
 9 Alan Eustace balloon 8  
10 Alan Eustace system 8  
# … with 466 more rows

# A tibble: 610 × 3  
 speaker word n  
 <chr> <chr> <int>  
 1 Tshering Tobgay carbon 23  
 2 Tshering Tobgay country 22  
 3 Tshering Tobgay bhutan 15  
 4 Tshering Tobgay neutral 12  
 5 Tshering Tobgay together 11  
 6 Tshering Tobgay one 10  
 7 Tshering Tobgay us 10  
 8 Tshering Tobgay world 10  
 9 Tshering Tobgay years 10  
10 Tshering Tobgay change 9  
# … with 600 more rows

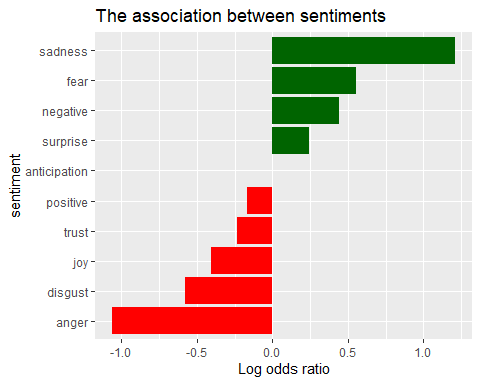
The first two bar graphs depict the top 25 words used by both the speakers.Alan has mostly been using words like ‘parachute’,‘can’ & ‘right’ whereas Tshering has mostly been using words like ‘carbon’,‘country’ & ‘bhutan’.The third figure shows a scatter plot where each pair of words is represented by a point on the plot, with the x-axis representing the frequency of the word in Alan Eustace’s speech, and the y-axis representing the frequency of the word in Tshering Tobgay’s speech. The line of best fit shows the trend in the data.



The table below shows which emotions were expressed more frequently by each speaker during the talks.The values in the table show the number of words associated with each emotional tone for each speaker.Here Tshering is showing more trust than Alan & having a more positive & joyous tone.

# A tibble: 10 × 3  
 sentiment `Alan Eustace` `Tshering Tobgay`  
 <chr> <int> <int>  
 1 anger 3 16  
 2 anticipation 36 68  
 3 disgust 3 10  
 4 fear 29 33  
 5 joy 24 65  
 6 negative 27 34  
 7 positive 71 151  
 8 sadness 17 10  
 9 surprise 20 30  
10 trust 42 96

The bar chart below displays the log odds ratio for each sentiment, with sentiments ordered based on their association with the speakers. Sentiments that are more strongly associated with Alan Eustace appear on the left side of the chart, while those more strongly associated with Tshering Tobgay appear on the right side. The color of the bars indicates whether the log odds ratio is positive (dark green) or negative (red). The x-axis shows the log odds ratio, which is a statistical measure that reflects the strength and direction of the association between the sentiment and the speakers. A lower log odds ratio indicates a stronger association with Alan Eustace, while a higher log odds ratio indicates a stronger association with Tshering Tobgay. Overall, the visualization helps to identify which sentiments are more strongly associated with each speaker and to compare the differences in sentiment expression between them.



In the following chart, we have displayed the most common positive words and most common negative words used in both the talks.The words like ‘laughter’ & ‘applause’ have more postive impact & on the other hand words like ‘government’, ‘impossible’ & ‘fight’ are having a more negative impact.

