# MA331- Mid Term Project

## Text analytics of the TED talks by Brian Cox and Terry Moore

2112601-Ryan-Daley

# Introduction

The aim of this report is to present and compare the word frequencies and sentiment analyses of TED Talks given by Brian Cox and Terry Moore. The report proceeds as follows. The introduction provides a brief summary of the two speakers and the nature of the content within their TED Talks. The second section will describe the methods and procedures this report uses, including statistical summary measures and visualizations. The third section will present the results of the report providing explanation of the what the data shows - this will include statistical results and graphs. The report concludes with a brief discussion of the findings, the limitations, considerations and main challenges of this analysis.

Brian Cox is a English physicist, author and TV presenter who specialises in particle physics. He is a professor at the University of Manchester but has also work extensively at CERN in Geneva working on the Large Hadron Collider. This reports analyses two TED Talks given by professor Cox which focuses on his work at the LHC in Geneva. They are titled; “CERN’s supercollider”, which was recorded in 2008 and “What went wrong at the LHC”, recorded in 2009. Professor Cox’s first speech is about the creation and scientific importance of the Large Hadron Collider, in which the aim is to recreate the conditions that were present less than a billionth of a second after the universe began. His second speech, given a year later is an update about the LHC, in which he expands upon a manufacturing default which caused damage to about 50 twenty-tonne magnets. The underlying theme of the second talk is to push forward with scientific inquiry in the face of setbacks and challenges.

Terry Moore is an American entrepreneur and former director of Radius Foundation, a forum designed for exploring and gaining insight from different worldviews. This report analyses two TED Talk’s given by Moore, the first titled, “How to tie your shoes”, released in 2005 and the second, “why is ‘X’ the unknown?”, released in 2012. His first talks starts with the proposition that he believes most people are tying their shoes incorrectly, explaining that there is a ‘strong’ and ‘weak’ form of the knot and that most have been taught the weak form. He demonstrates that by simply going counter clockwise around the bow instead of clockwise, produces the strong form of the knot and makes it less likely that your laces will become undone. His second talk is a history lesson to discover why ‘X’ represents the unknown. Charting its evolution from Common Era Persia and then through to Spain in the 11th century, the words in Arabic became lost in translation. Moore quips at the end that X is the unknown because you can’t say “sh” in Spanish.

# Methods

For the word frequency analysis there are two main tools which will be used, the tidyverse meta-package and tidytext. The transcripts are located in the ‘text’ variable. In order to analyse the data within the variable ‘text’, the data needs to made ‘tidy’. The data must be ‘tidied’ and ‘tokenized’ so that the format is with one observation per row, therefore making it compatible with tidy tools. This is done through the process of tokenization which will identify and break apart the text into individual

tokens. The tidytext 'unnest_tokens()' function will be used to break apart the text into individual tokens. There is a final step after the data has been tokenized and tidied.

The final step before analysing the TED Talks is to identify and remove stop words from the speech. Words like "the", "and", and "to" are expected to occur most often - these are what are referred to as stop words are not interesting for text analysis. The report uses 'get_stopwords' provided by tidytext package to remove these unwanted words from the data set. Now the data set is tidied, the data is filtered so that only the talks by Brian Cox and Terry Moore are selected and a variable is created for each speaker.

A histogram is then created for each speaker to visualize their top 25 most frequent words within their TED Talks. To provide a better visualization of the speakers word frequency, a graph is then plotted with the words of both speakers against each other.

For the sentiment analysis of the two speakers, a sentiment will be assigned to every word of the speakers talks using the 'nrc' lexicon. The sentiment counts for each speaker are then visualized by a bar chart which compares the sentiment against the Log odds ratio.

# Results

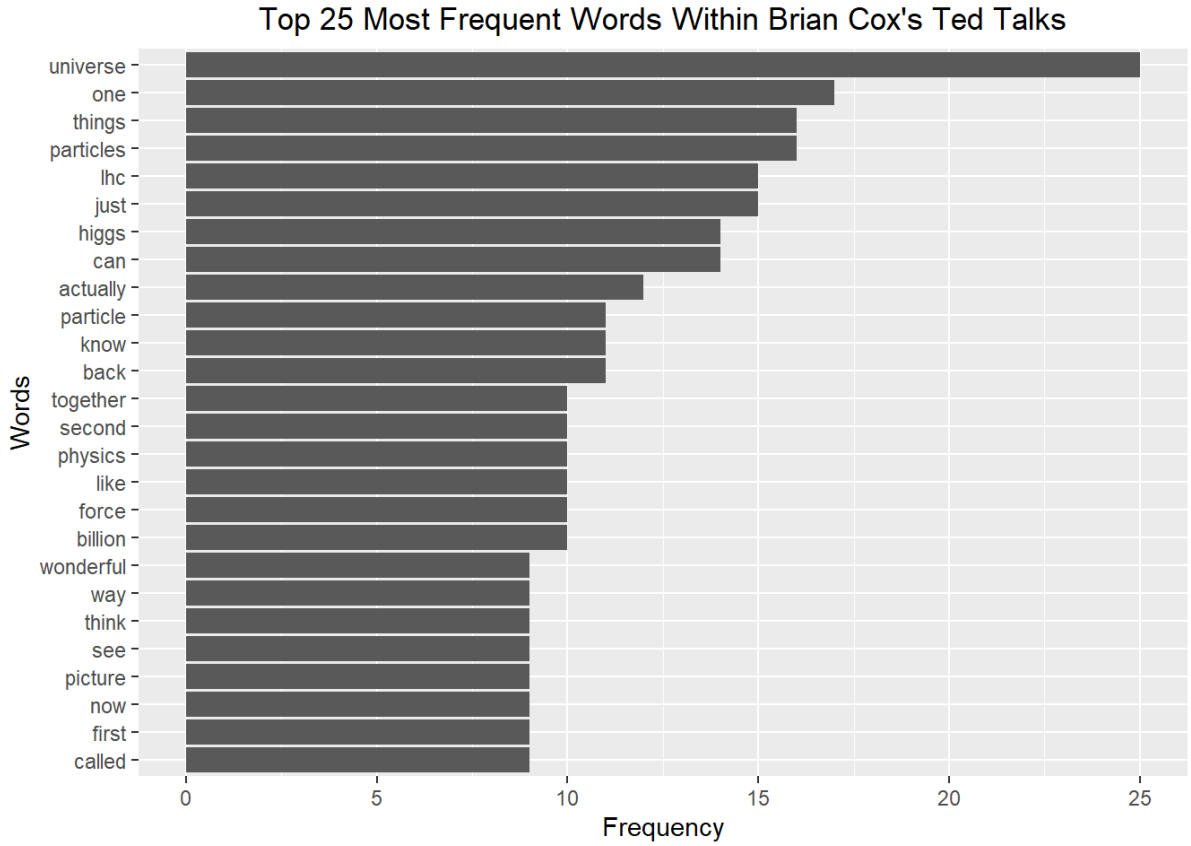


*Figure 1: Brian Cox's 25 most frequent words*

The graph above shows the top 25 most frequent words within Brian Cox's Ted Talks. From this we can see that the word 'universe' was by a substantial margin the most frequent word within his talks used 25 times. This is in keeping with both talks regarding the LHC and about its implications for discovering the building blocks of the universe.

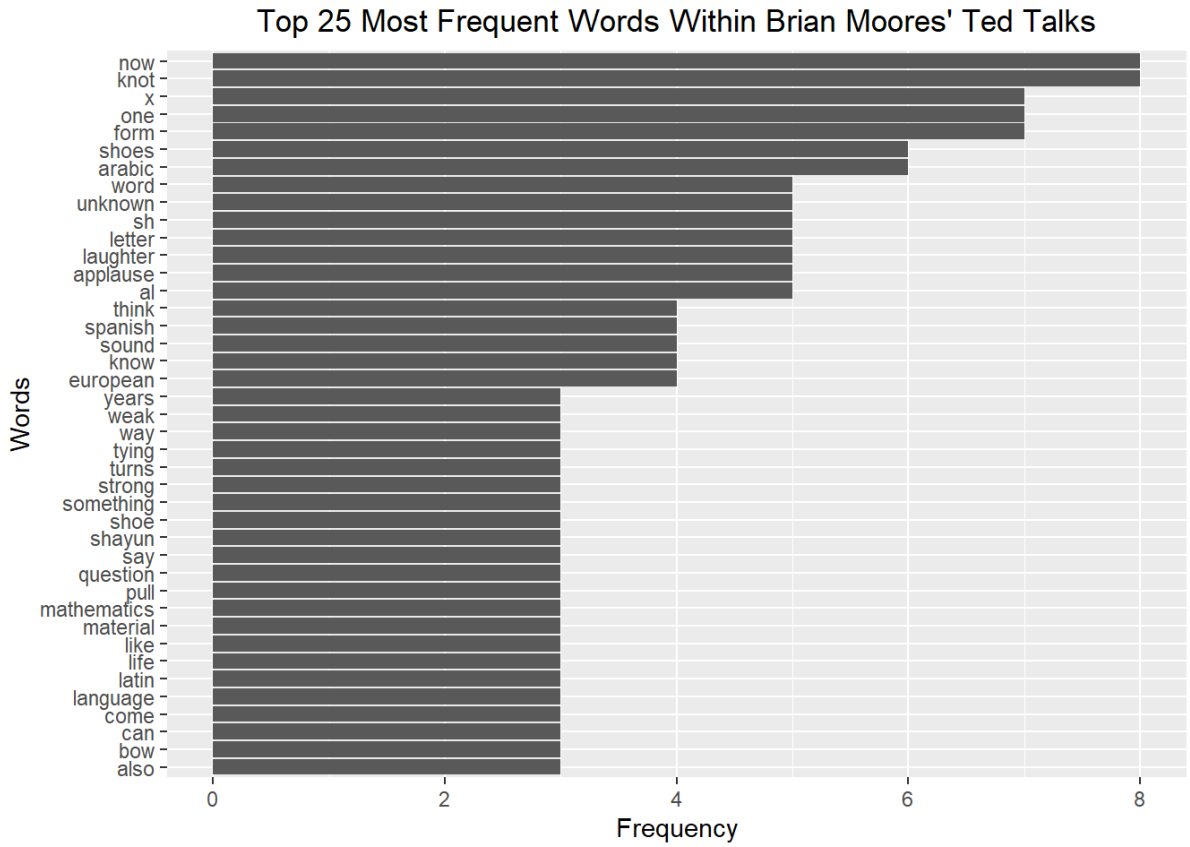


*Figure 2: Terry Moore's 25 most frequent words*

The graph above shows the top 25 most frequent words within Terry Moore's TED Talks. 'Now' and 'knot' are both tied as the most frequent having been repeated a total of 8 times each. Again this is understandable as one of Moore's talks was on the correct way to tie a knot for your shoelaces. Third most frequent letter was 'x' again in keeping with the subject matter of his second talk.

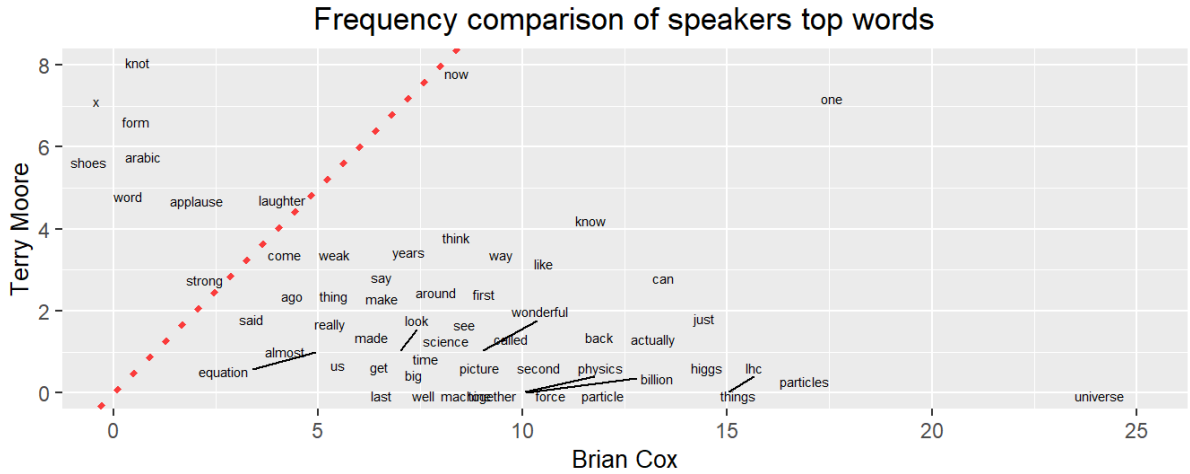


*Figure 3: Frequency comparison of speakers top words*

This graph represents the highest frequency words among both speakers. Words are included in the graph if they had a frequency of 5 or more. From this graph it is clear to see that Brian Cox has a significantly larger amount of words with a frequency of 5 or more. This can be explained in part by the length of Brian Cox's TED Talks in comparison to Terry Moore. The sum of words within Brian Cox's

talks is 1543, whereas in comparison Moore's sum of words is 447. With Cox having more than triple the amount of words than Moore, the skew towards Brian Cox is to be expected.

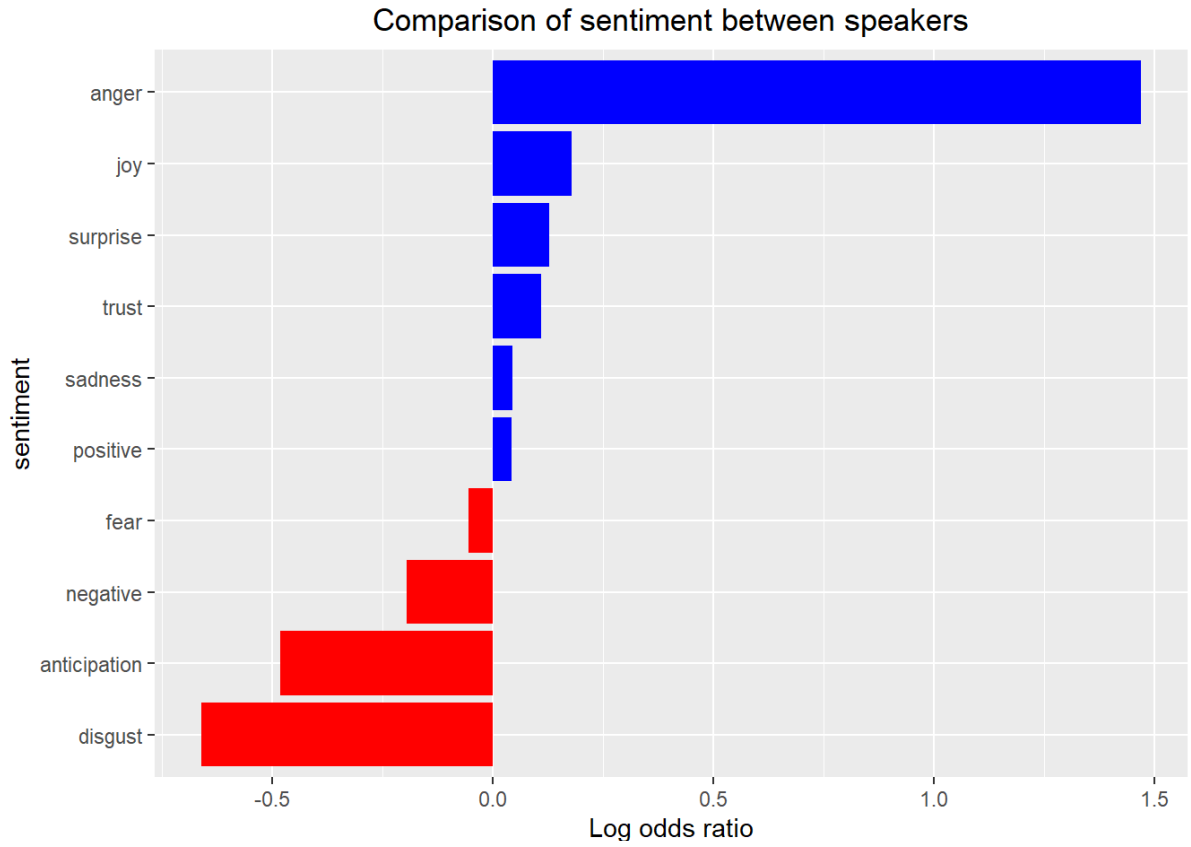


*Figure 4: Sentiment comparison between both speakers*

The graph above shows the adjusted log odds ratio of the comparison of sentiments between Brian Cox and Terry Moore. The blue bars represent sentiments that had a higher prevalence in Brian Cox's Talks based on the words used and the red are the sentiments that had a higher prevalence with Terry Moore's talk based on the words used. From this we can see that Brian Cox's most prevalent sentiment was Anger and Terry Moore's was disgust. Interestingly, both speakers most prevalent sentiment were negative.

# Discussion

The aim of this report was to present and compare the word frequencies and sentiment analyses of the TED Talks given by Brian Cox and Terry Moore. This was done through the process of tokenization, stop word removal, followed by analysis and visualization. From this, the report has identified Brian Cox's most frequent word was 'universe' and Terry Moore's was tied between 'now' and 'knot'. The sentiment analysis has shown that Brian Cox's most prevalent sentiment was Anger and Terry Moore's was disgust, both speakers most prevalent sentiment were negative. The main challenges and limitations of this report was within the sentiment analysis. In sentiment analysis, a word is assigned to a sentiment. However this approach can miss context-dependent sentiment. Although the tone of Brian Cox's speeches were positive and enthusiastic, one might get an impression that he was angry through his talks. Such an example can limit the validity and usefulness of the sentiment analysis. Overall however, this project has demonstrated that text analytics and sentiment analysis can be useful tools in gleaning insight from data.