TRENT UNIVERSITY

TED Talks: #Trending

Sophie Castel, Kara Fitze, Melissa Van Bussel, and Seyeon Kim
Department of Mathematics and the Applied Modeling and Quantitative Methods (AMOD) Program
Faculty Advisor: Wesley S. Burr

Background

- TED is a non-profit organization devoted to spreading ideas through short (18 minutes or less), powerful talks. TED began as a technology conference in 1984 and has since grown to a global community.
- Web scraping techniques were used to gather data from the TED website about audio and video recordings of TED Talks published online from June 2006 to September 21st, 2017.
- The dataset used in this analysis consists of 2550 observations of 17 attributes, including the published date of the recording, the number of comments, tags, ratings, and the number of views.

Objectives

- Develop a unique composite metric to classify and predict the popularity of TED Talks
- Determine whether the popularity of TED Talks has changed over time through the use of Time Series Analyses
- Determine how, if at all, the themes and tags of a TED Talk will contribute to its popularity
- Use sentiment analysis to determine how the content of TED Talks has changed over time

Methods

- The statistical language R was used to examine the dataset, to perform calculations, and to plot the data. An online word-cloud generator was used to create the word-cloud, while the calculations for size and colour of the words was completed in R.

Results

We created a composite metric to determine the popularity of a TED Talk:

$$P = \log((v + 300r + 300c)e^{-t/11})$$

where v = number of views, r = total number of ratings, c = total number of comments, and t = number of years between the recording's published date and 09/21/17. Our metric was first scaled to range from 0 (exclusive) to 10 (inclusive), and then a logarithmic transformation was applied, making the distribution of P approximately Normal (see Figure 1). We wanted our metric to reflect the level of engagement with the recordings, which is why r and c are weighted more heavily than v. We included a decay term, to account for the fact that more recently published videos have had less time to gain popularity.

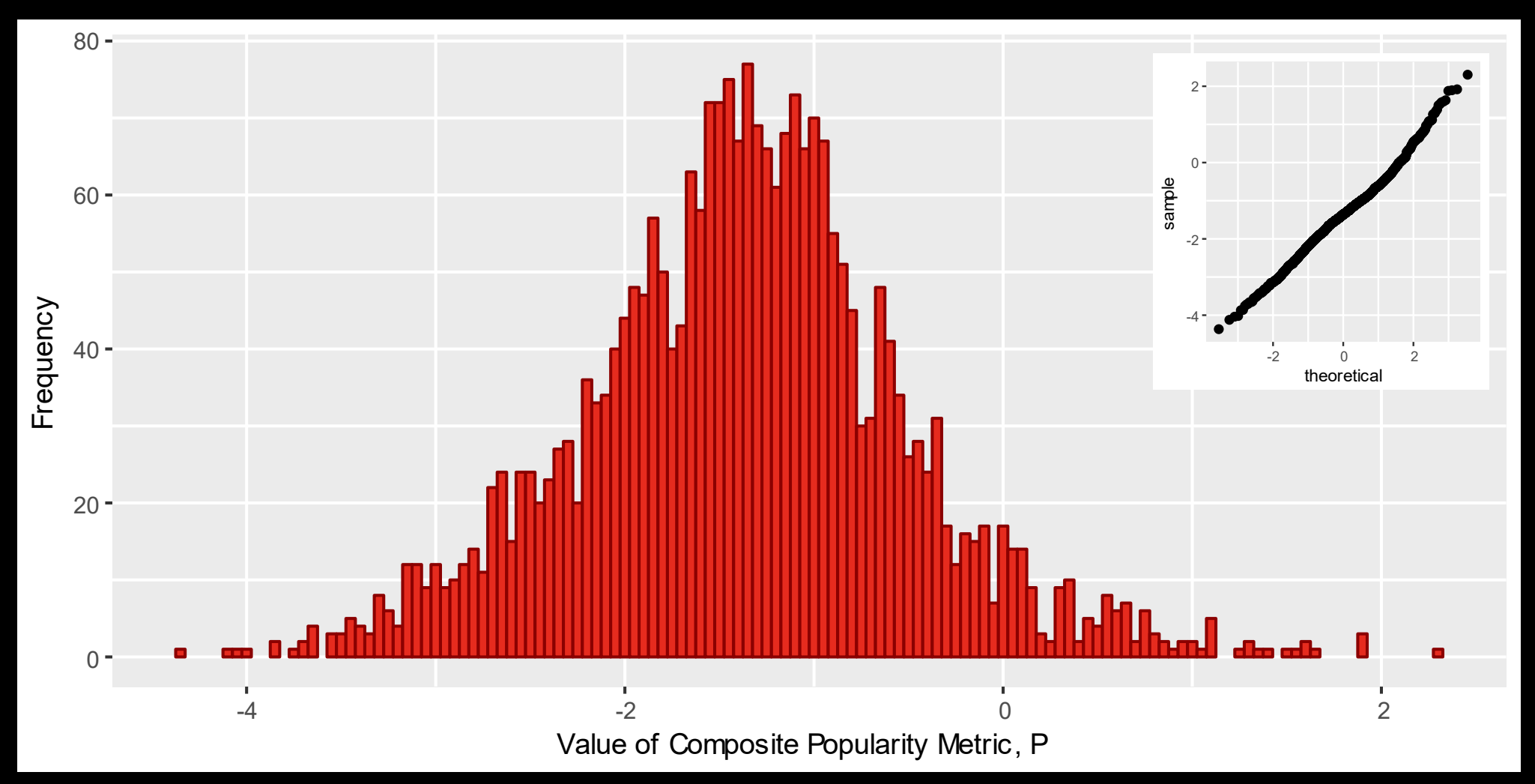


Figure 1: Distribution of our Composite Popularity Metric, *P*, is approximately normal.

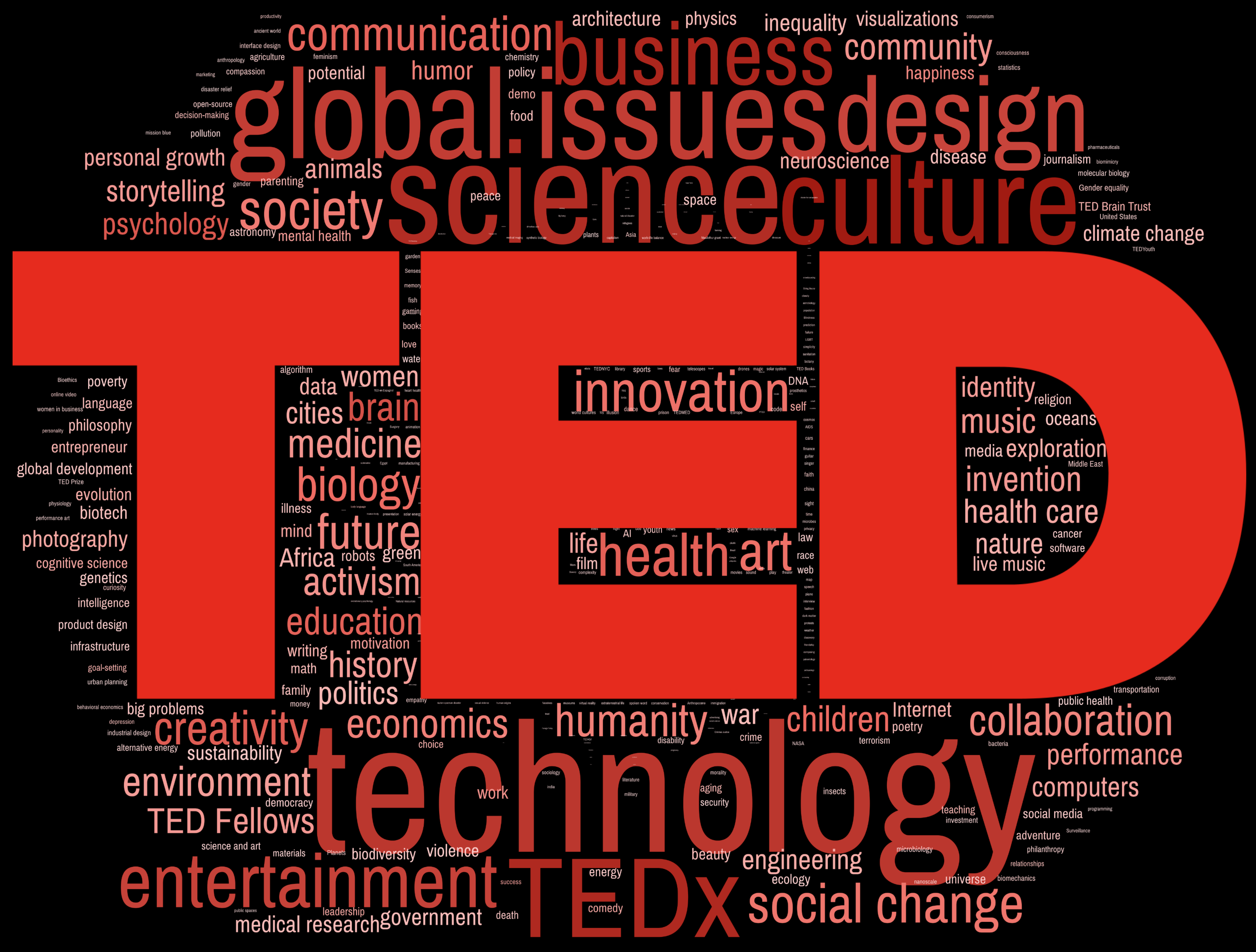
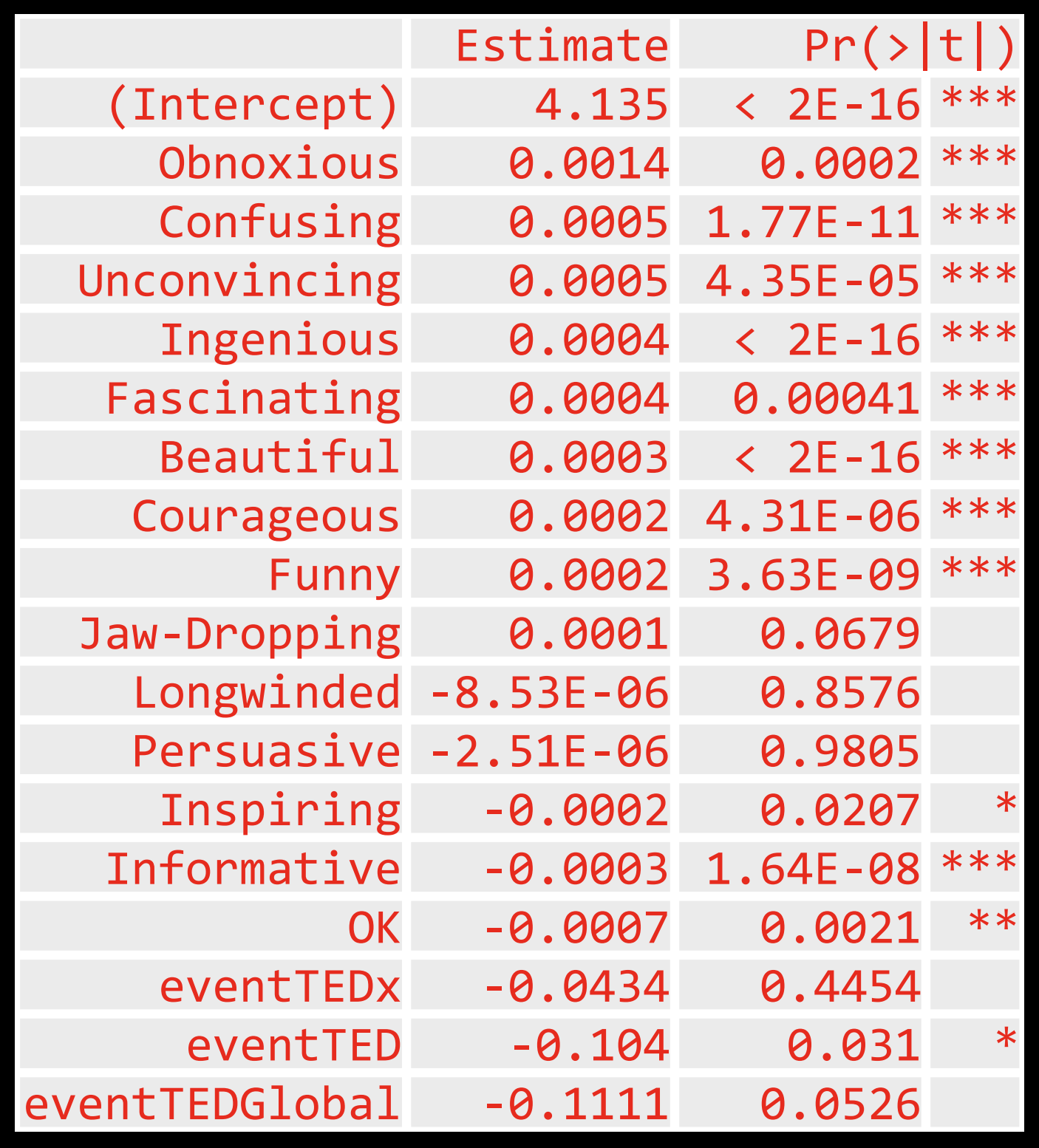


Figure 2: A 'word cloud' representation of all tags defining the complete set of TED Talks. The colour of each word is proportional to the total number of views associated with each tag, where the contribution to the total view count is based on a weighted sum proportional to the tag index, *j*, where *j* = 1,2, ...,32. Word size is proportional to the tags frequency across all *j*. The tags are placed randomly in order to adequately fit the shape of the TED logo.



	Estimate	Pr(>|t|)	
(Intercept)	4.135	< 2E-16	***
Obnoxious	0.0014	0.0002	***
Confusing	0.0005	1.77E-11	***
Unconvincing	0.0005	4.35E-05	***
Ingenious	0.0004	< 2E-16	***
Fascinating	0.0004	0.00041	***
Beautiful	0.0003	< 2E-16	***
Courageous	0.0002	4.31E-06	***
Funny	0.0002	3.63E-09	***
Jaw-Dropping	0.0001	0.0679	
Longwinded	-8.53E-06	0.8576	
Persuasive	-2.51E-06	0.9805	
Inspiring	-0.0002	0.0207	*
Informative	-0.0003	1.64E-08	***
OK	-0.0007	0.0021	**
eventTEDx	-0.0434	0.4454	
eventTED	-0.104	0.031	*
eventTEDGlobal	-0.1111	0.0526	

Table 1: A summary of the estimates generated in the full model equation.

Our interest lies in identifying the rating attributes that are positively correlated with popularity. Constructing such a model could help content creators set the tone of their presentation so as to maximize public engagement and impact. We suspected that the event type would also explain a significant amount of the variation in popularity.

In order to reduce the dimensionality of the *event* variable, a categorical variable was created with four levels: TEDGlobal, TEDx, TED Annual, and "Other". A linear model regressing the fourteen rating counts for each descriptor and the new event variable onto our Composite Popularity Metric was then fit. The optimal model specification was realized using a stepwise algorithm based on Akaike's Information Criterion (AIC). This optimal model used all of the significant predictors in Table 1.

We attempted a classical STL decomposition of the average monthly Composite Popularity Metric and found no evidence of periodic structure. Fitting an ARIMA model to the data gave an ARIMA(2, 1, 0) as the model which minimized the AIC. This allowed for a short forecast, indicating that the popularity of TED Talks will continue to increase in the near future. In addition, since the time series structure appears minimal, we fit a linear trend to the data which indicated the same result.

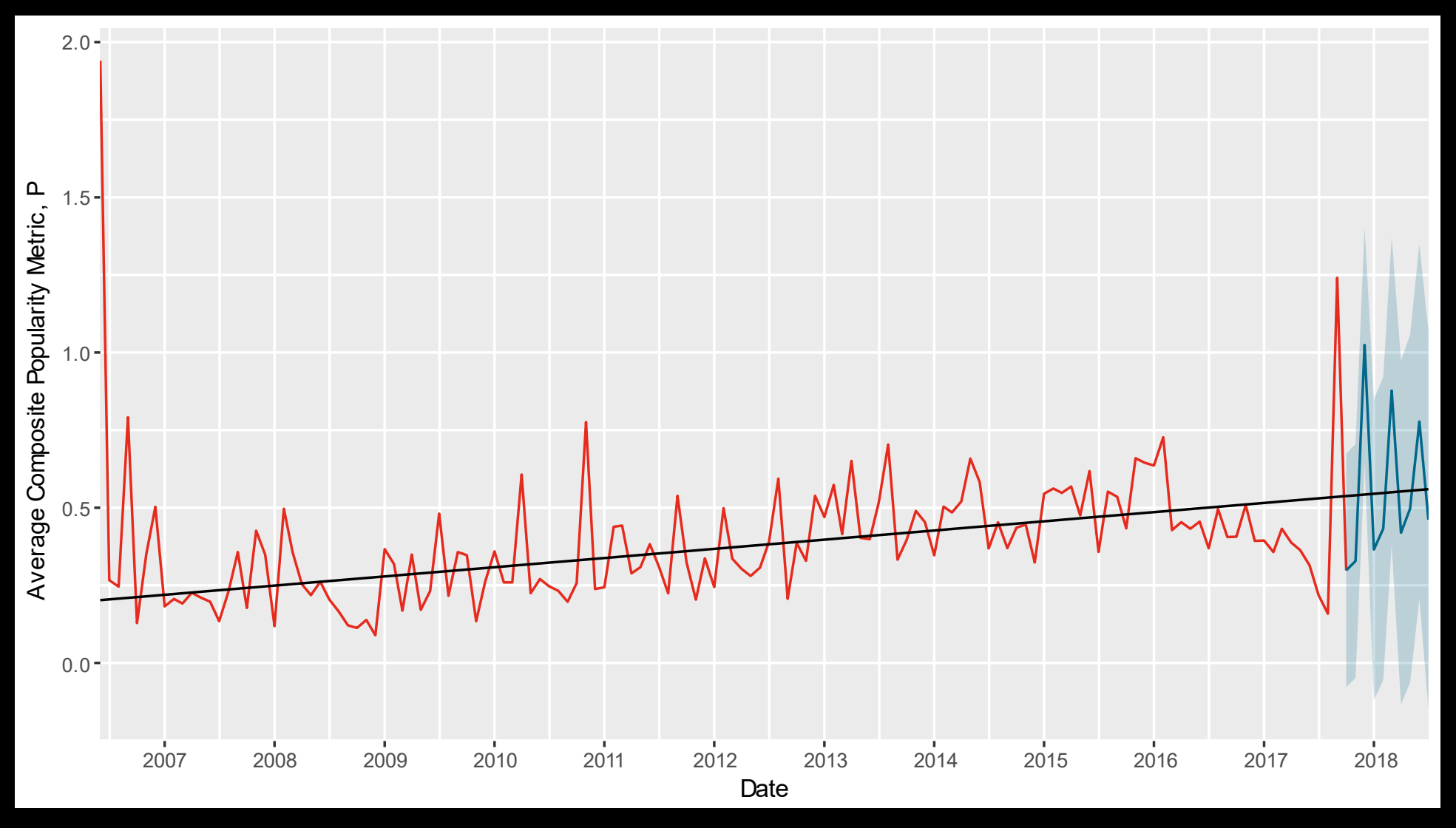


Figure 3: Observed monthly average Composite Popularity Metric, *P*, with forecast from ARIMA(2, 1, 0).

Sentiment Analysis

- Sentiment analysis is a form of text mining that identifies opinions and emotions expressed in a piece of writing or other media. There are many different ways of doing this, but the `tidytext` package in R contains several sentiment lexicons.
- The AFINN lexicon (created by Finn Arup Nielsen) assigns a value from -5 to +5 to each word, where positive values correspond to positive sentiment and negative values correspond to negative sentiment.
- We performed sentiment analysis on the tags used in the TED Talks videos using the AFINN lexicon.

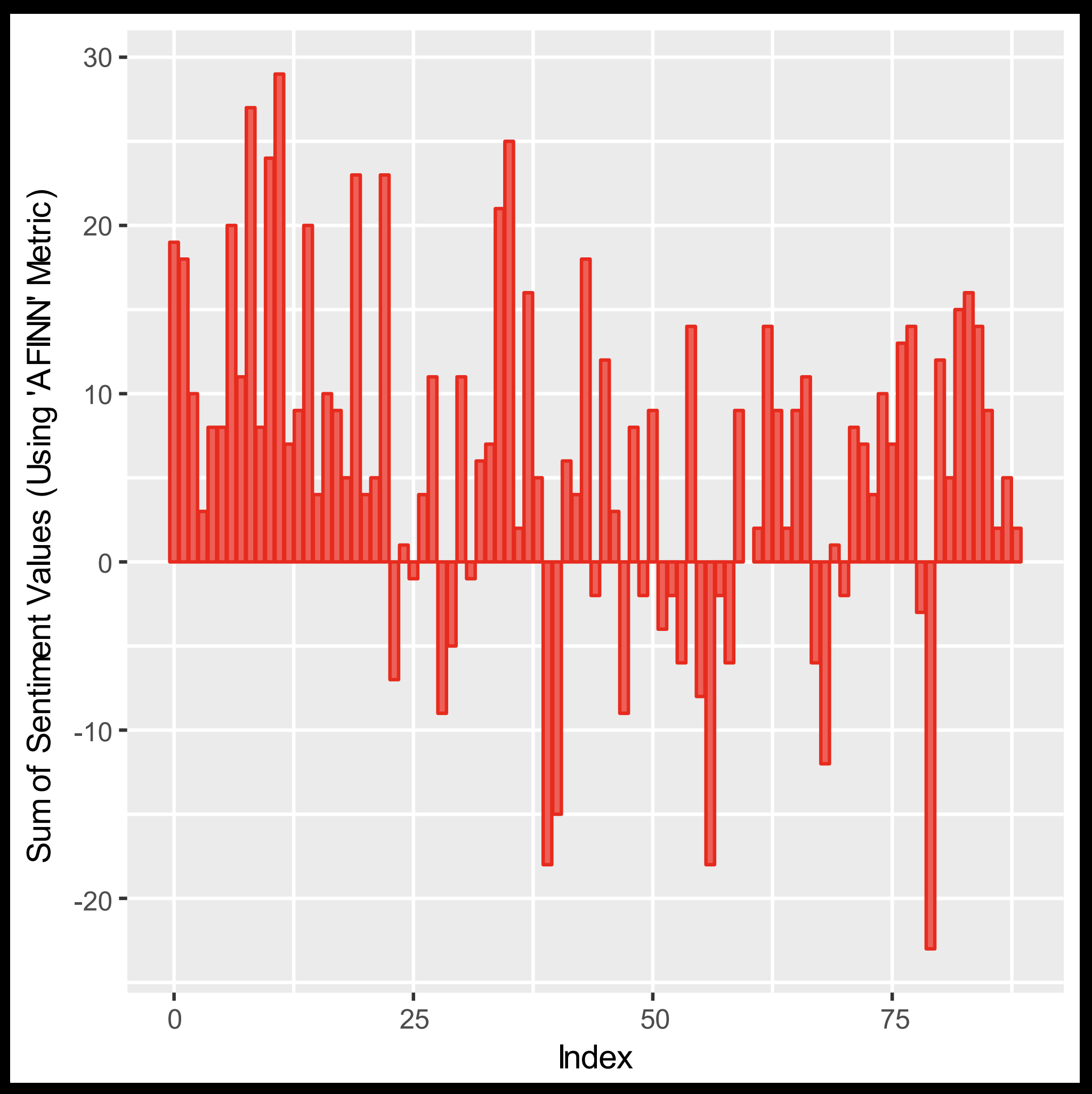


Figure 4: Composite sentiment analysis over time of TED Talks. Each tag in the dataset was given one line in a data frame, in chronological order. This data frame was then split into groups of 250 tags for a total of 89 groups, and a sentiment metric was then calculated for each of the groups. This metric consisted of the sum of the AFINN sentiment values for each of the tags in the group. These metrics were then plotted in chronological order, to demonstrate how the sentiment of TED Talks has changed over the years.

Conclusions

- Using our Composite Popularity Metric, we determined that the popularity of TED Talks peaked in 2006, dipped, and then slowly began to increase over time.
- Attributes like "Obnoxious", "Confusing", and "Unconvincing" are the highest contributors to a talk's popularity. Initially, this fact seems contradictory; however, one must consider that our definition of popularity reflects audience engagement, which is not restricted to be positive. In addition, our findings contradict our initial belief that event type is highly significant.
- TED Talks have primarily been focused on technology, science, and global issues.
- An ARIMA(2, 1, 0) forecast indicated that the popularity is expected to continue to increase in future years.
- The sentiments of TED Talks have decreased over the years, demonstrating that TED Talks have become more controversial over time.

References

TED Talks dataset obtained from https://ssc.ca/en/case-study/ case-study-2-what-predicts-popularity-ted-talks.
Word cloud created using worldclouds.com.