Megan Eichelberger
Willamette University Capstone Summer 23
Jameson Watts

Introduction

Understanding the presentation of language and the way people understand and connect with the content is imperative to the success of large language models, text prediction, and general readership. Looking at various measures including readability, sentiment, and variability the author seeks to gain greater understanding of the differences – should they exist – across news articles on broad topics of sport, politics, finance, and art.

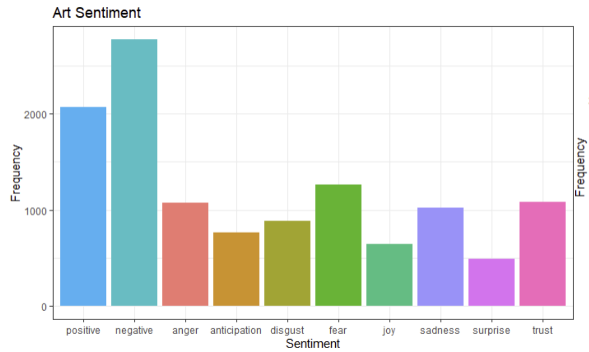
Data

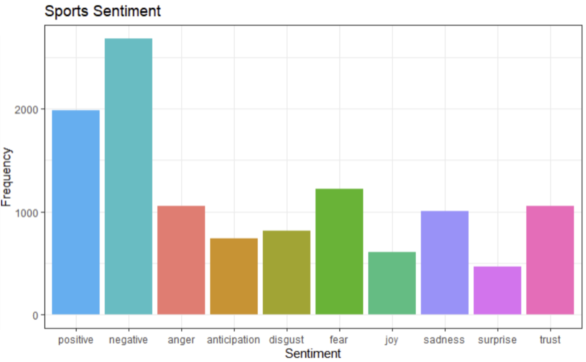
The articles covered in this analysis are from The Guardian, a British newspaper with UK, US, Australian, and International editions. The Guardian provides an API with robust accessibility and *r* integration, allowing the author to access a material body of text. Further, an analysis conducted by online readability toolkit provider Readable showed that of the nine news outlets investigated shows the Guardian with a readability score in the medium-low range (averaging a 6th-7th grade level), giving this author the opinion that the Guardian would be a fair representation of an accessible set of texts (2019).

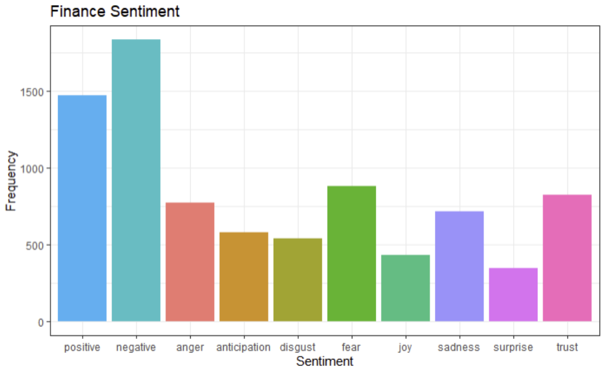
The Guardian's API grants access to articles published between 1999 and the present day. For the purpose of this analysis, the author chose the time frame between January 1 2023 through June 30 2023. This provides a material body of text while also narrowing the scope enough to allow for the analysis to encompass various topics. The author imported the body text then created connections with sentiment libraries that allow for analysis of text by both binary positive or negative sentiment and more complex sentiment categorization. The body text was tokenized to allow for analysis at a more granular level. The text was also coerced into a corpus, necessary for topic modeling.

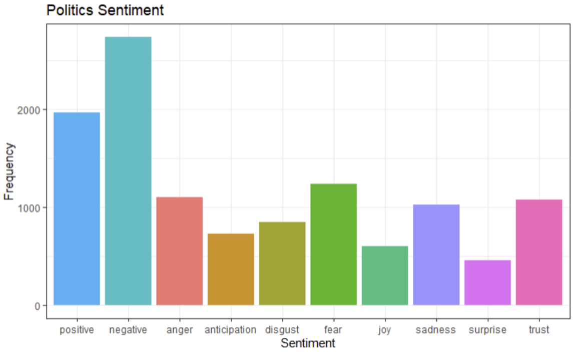
Sentiment and Word Count

Sentiment analysis performed on the topics showed that the types of sentiment across all topics is fairly commensurate. The author hypothesizes this is due to the text being from the same publication – many news outlets have specific standards for reporters. In fact, the only major departure is the anticipation sentiment in financial reporting.

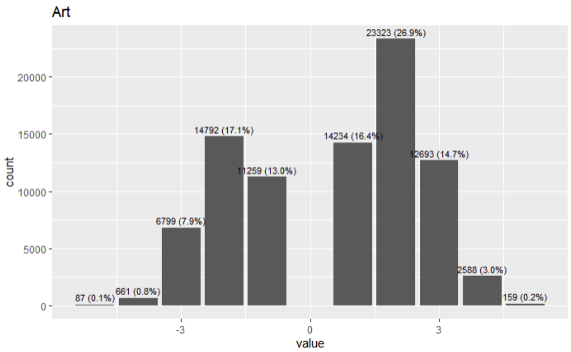


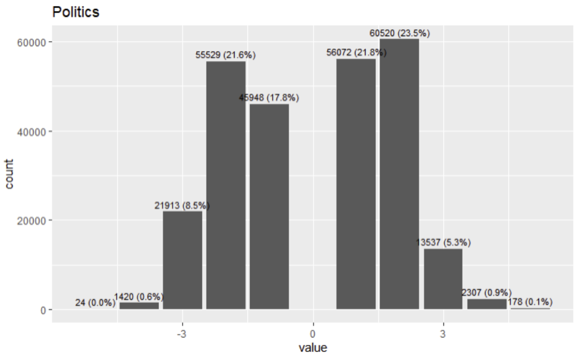


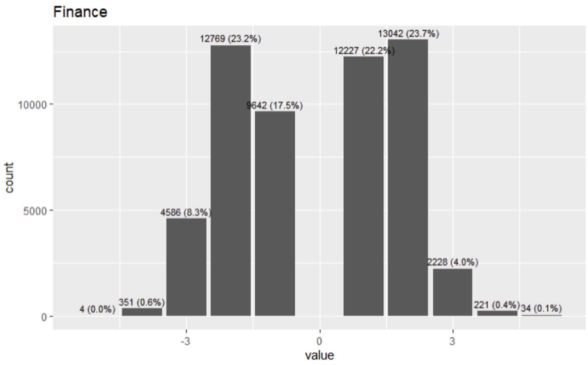


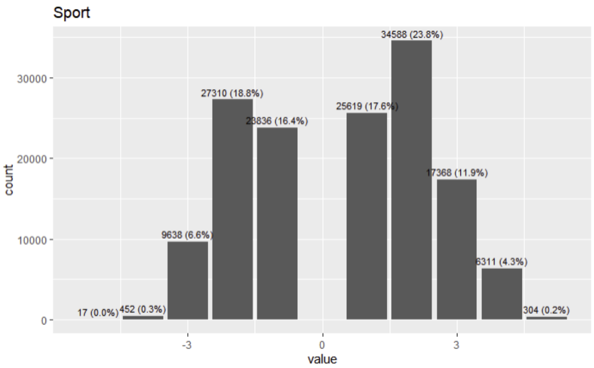


While these measures of sentiment are generally aligned, the intensity of the sentiment is not necessarily so. The below graphs show the measure of sentiment on a scale of -5 to 5, -5 being the most negative and 5 being the most positive. Words scoring 5 include outstanding and superb while -5 includes words the author will not print in this paper due to their extreme vulgarity. Interestingly enough, articles about art tend to have the greatest concentration of these -5 words, while perhaps less surprisingly sports have a greater concentration of superlative 5s.

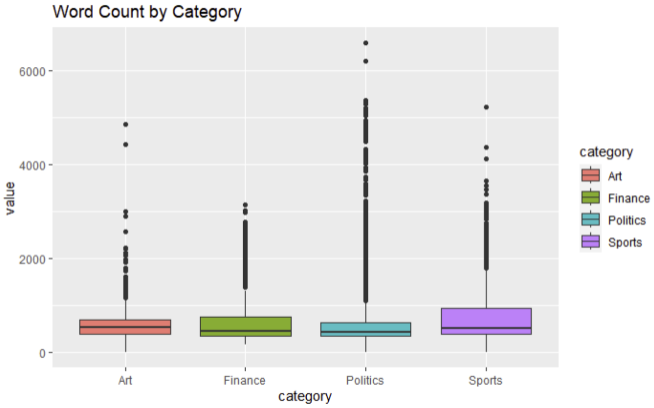






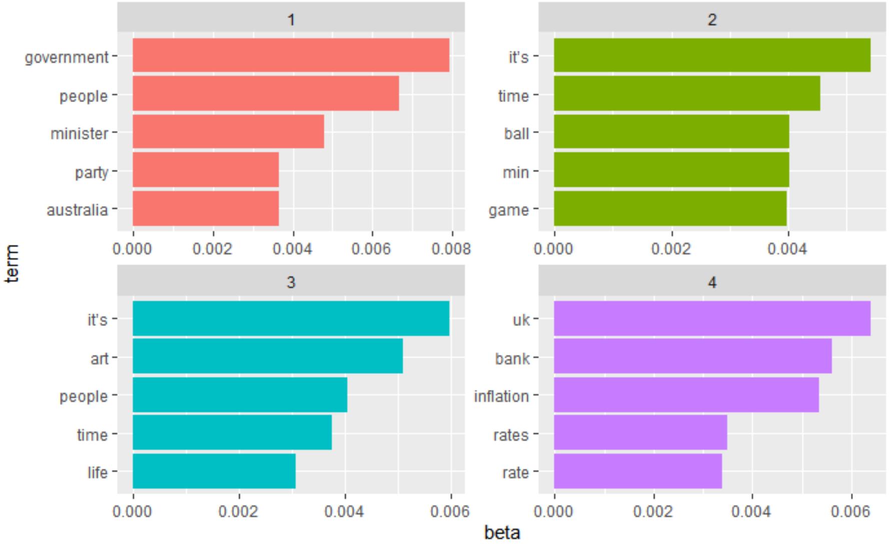


Looking at the word count of each article, we see that while each category – apart from Finance – has some significant outliers, the length of article does generally align.

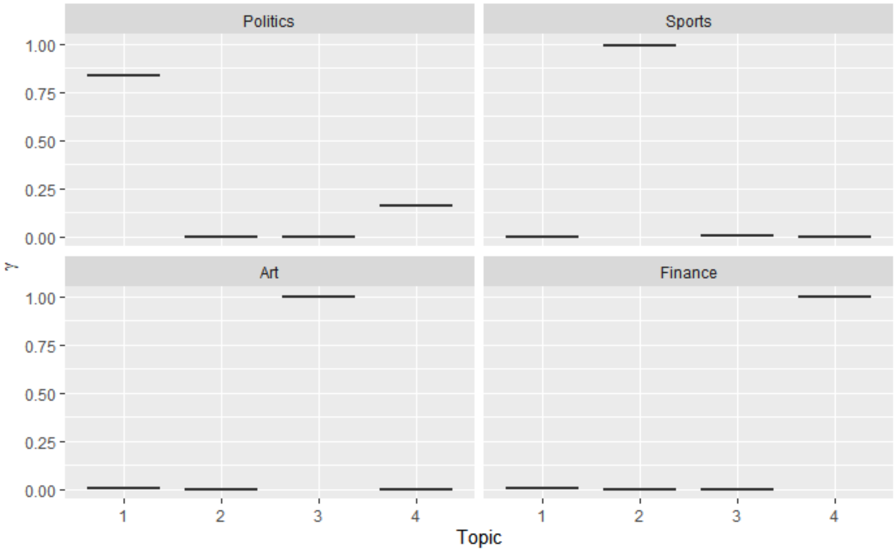


LDA and Predictability

With there being so much alignment in sentiment we instead look to other factors of predictability. The author uses topic modeling techniques to investigate predictors, most specifically LDA – Latent Dirichlet allocation. An LDA model allows us to look at corpora which may have overlap and make predictions as to which specific corpus – henceforth referred to as topic – it belongs. We run the LDA with four topics in the model, and the model returns the likelihood that the word belongs to a particular topic. We then visualize the top 5 for each topic.

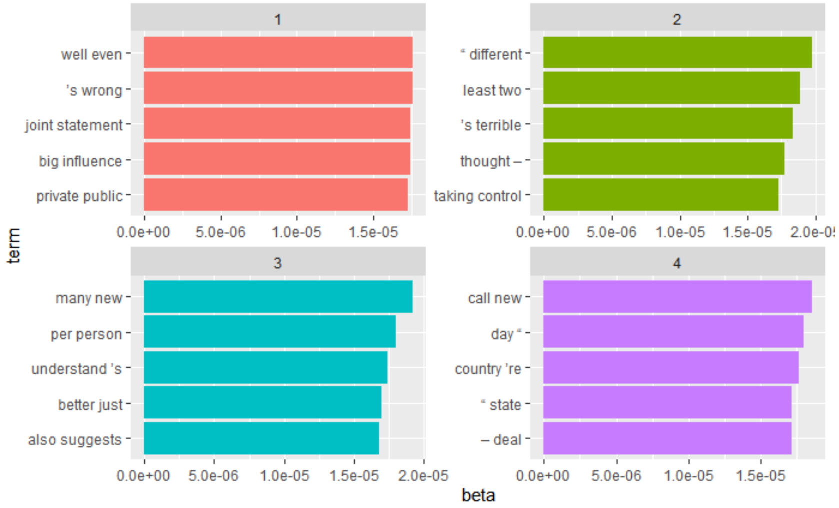


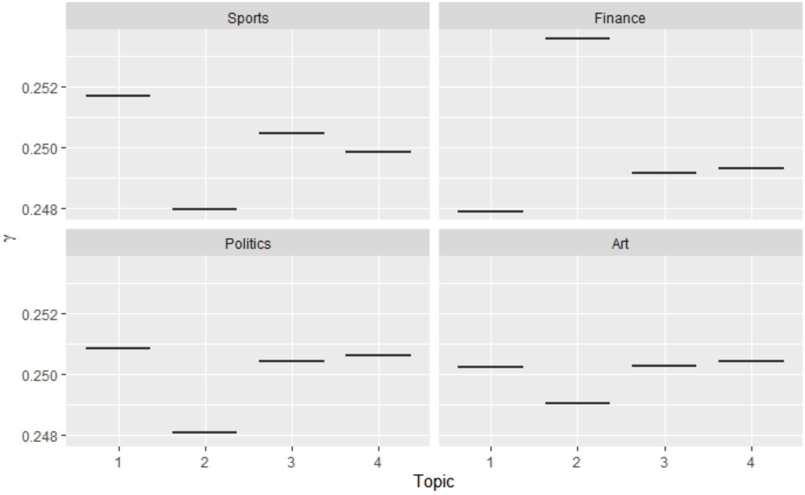
Knowing beforehand that our topics are art, finance, politics, and sports we may be able to surmise on our on which topic is which – there are some keywords that an English speaker may be able to use to make a quick prediction. However, we can work backwards and make sure. Calculating the LDA using the gamma, we can check our work.



The model shows that sports, art, and finance can be predicted perfectly based on the text provided. Politics is a little more questionable, but predictive power is still very high.

Using modeling techniques we are able to identify predictive words in a corpora of text. It is worth noting that when we move from single words to bigrams the predictive power declines significantly. Take special note of the scales on the below graphs in contrast with the preceding graphs. The most common bigrams simply are not specific enough to predict with any real confidence the topics to which they belong.





Conclusion and Further Research

Applying these techniques to language corpora has potential to bolster our understanding of language and predictability. The limitations to this project as it stands center mostly around availability. The author would in the future like to apply the concepts within to greater bodies of texts and see if the findings bear out across other types of media. Gaining access to additional modern bodies of texts is difficult, as most are not readily available until they leave the public domain 70 years after the death of their author. With a significant decrease in API connections with major news outlets and little to no availability of current books there are significant roadblocks to further investigation. The author looks forward to more robust analysis in the future. These findings are an exciting start to broaden the understanding of predictability and sentiment in text.