

# **Lost in the News Sauce?**

**Adventures in classification**

**By: Oluwaseyi Malonia, Chibuzo  
Ugonabo , & George Gee**



# The Data

9500+ links scraped from NPR

- Dating back to 2016

1800+ articles scraped

1610 Articles chosen for modeling.

Perfectly balanced

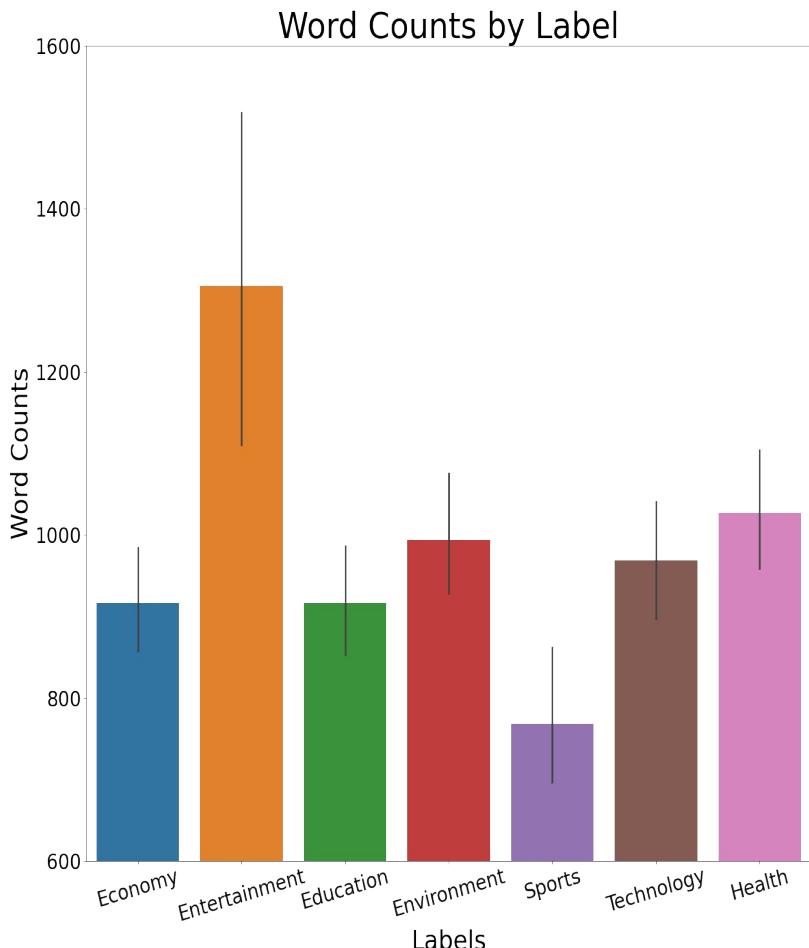
**Train, Validation, Test Sets**

Trained on 1,050 articles

Validated on 350 articles

Tested on 210 articles

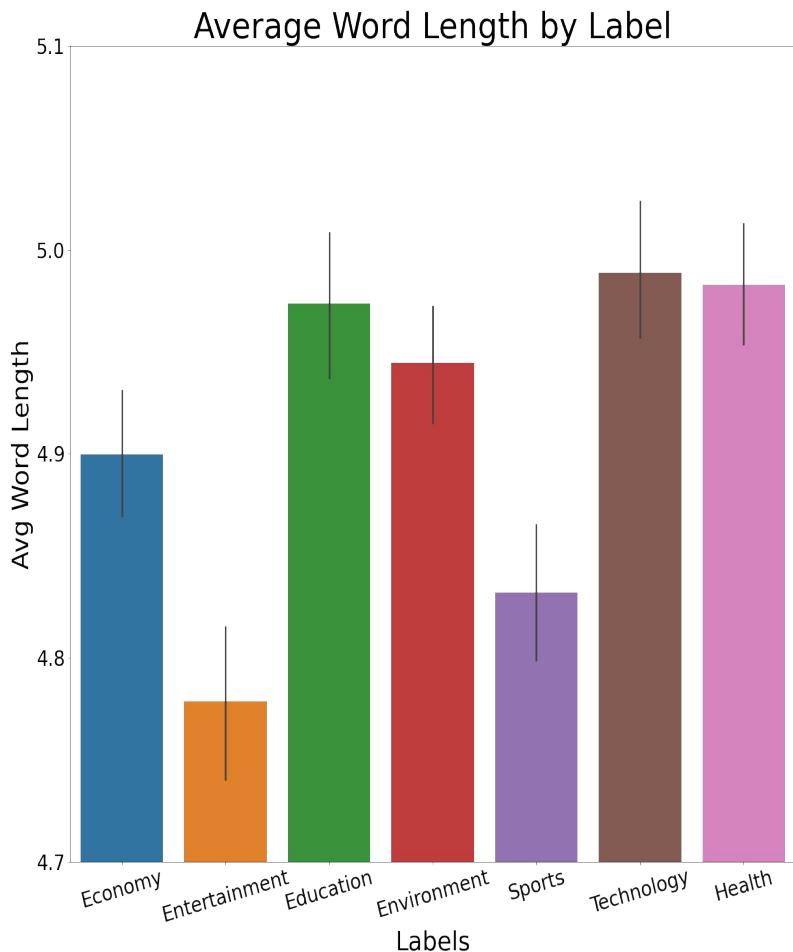
- Topics included:
  - Economy
  - Education
  - Entertainment
  - Environment
  - Health
  - Sports
  - Technology



### Average Word Count per Label.

Sports averaged the lowest amount of word counts.

Entertainment average the highest amount of word counts.



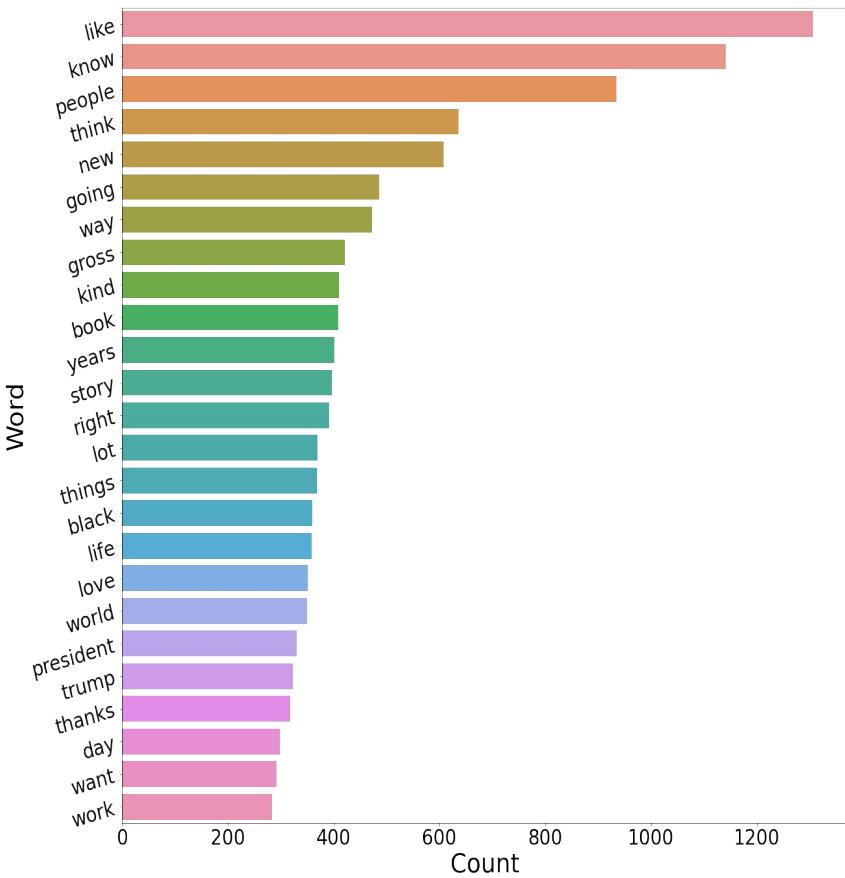
### Average Word Length per Label.

Entertainment uses smaller words.

Technology uses larger words.

Likely a direct correlation with the target audiences education level.

## Word Count Distribution For Entertainment Articles



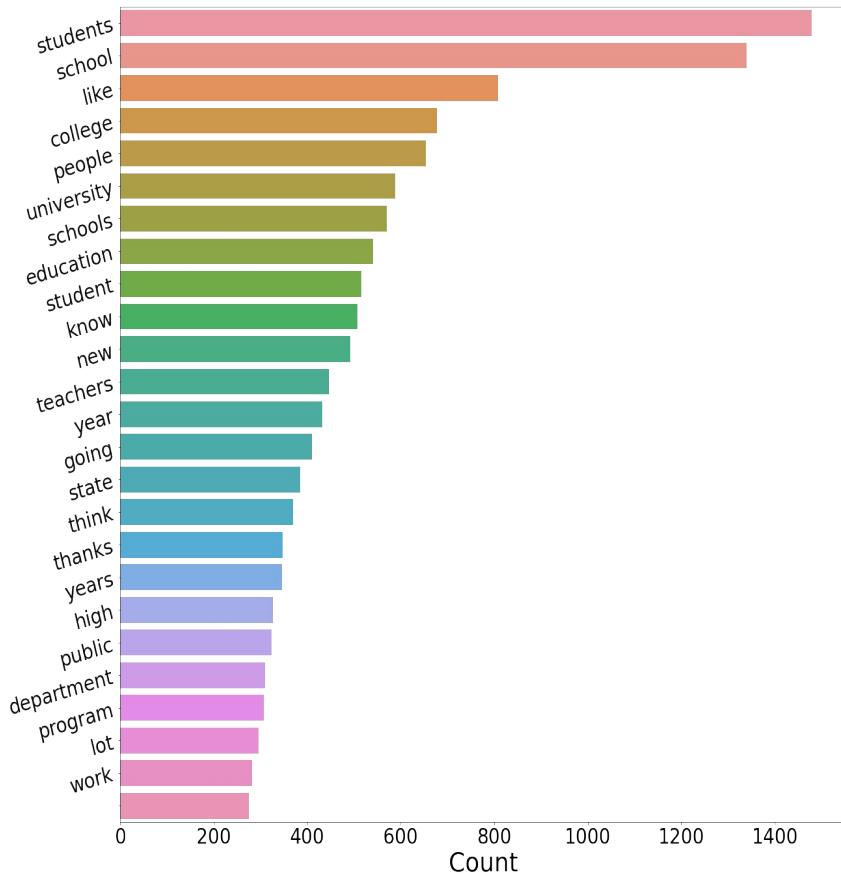
## Entertainment Word Frequencies

Entertainment contains buzzwords such as 'book', 'story', 'like', and 'love'.

Contains oddities like President Trump...

Say what you will but the oval office had never been so "entertaining".

## Word Count Distribution For Education Articles

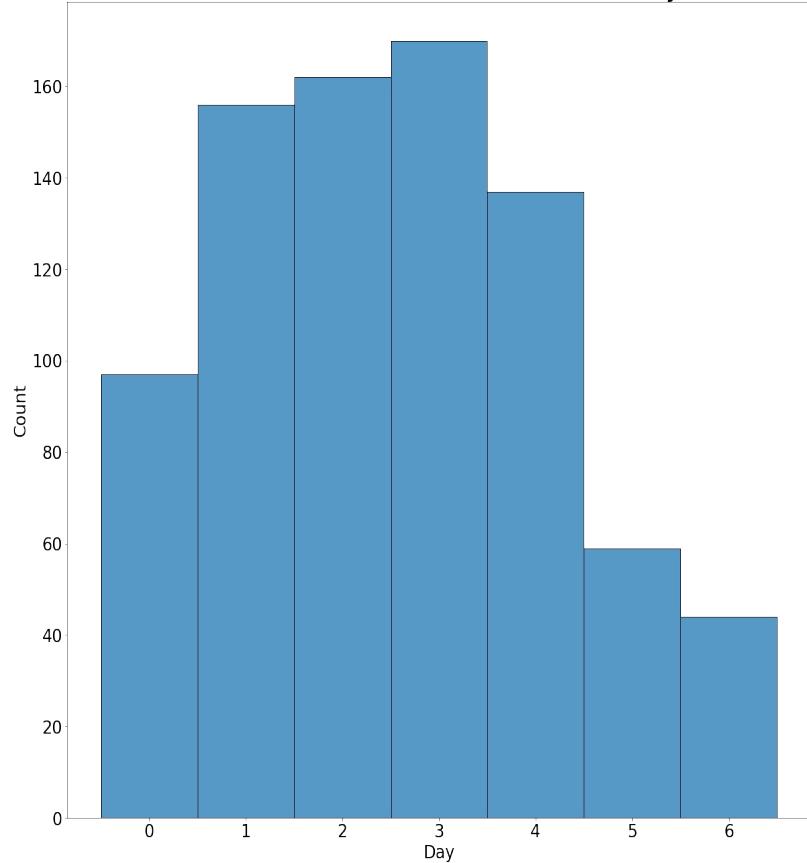


## Education Word Frequencies

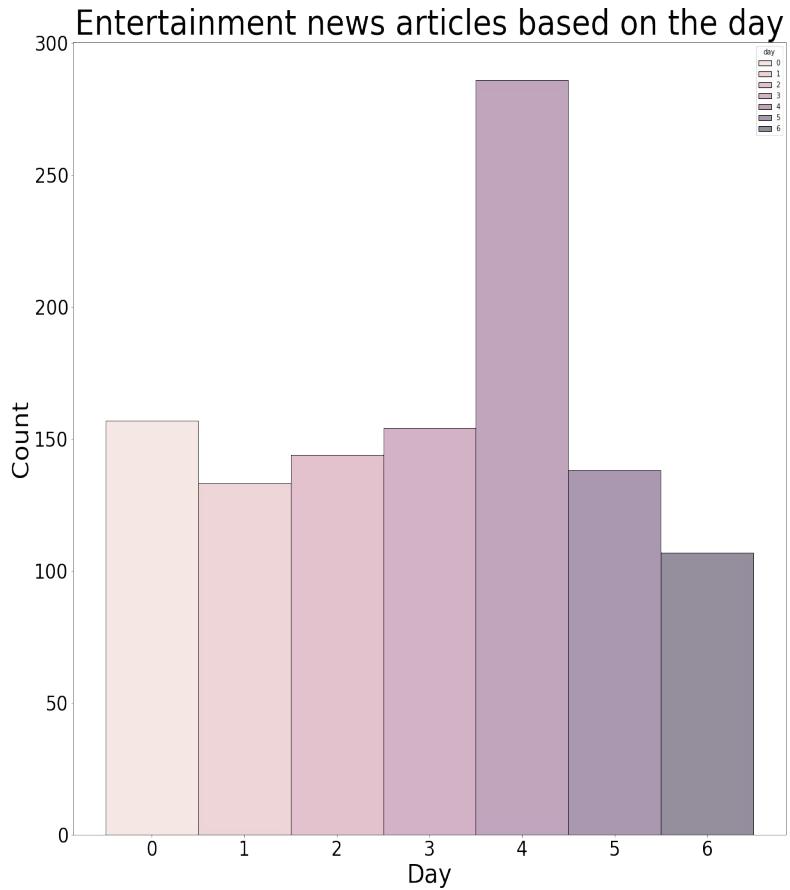
Education contained mostly expected buzzwords such as:

- Student
- University
- School
- Teachers

Tech news articles based on the day



News articles for Technology tend to be released at the start of the week, peaking at Thursday before tapering off for the weekend.



News articles for Entertainment tend to be calm at the start of the week. Peaking on Friday before tapering into a calm weekend.

---

# Data Processing

# Text Processing

Stripped of all HTML meta-data.

Stripped of embedded hyperlinks and emails.

Stripped of special characters, digits, punctuation, and stop words.

All text converted to lowercase.

Stemming and Lemmatization

## Stemming vs Lemmatization



# Modeling

# TF-IDF

Term Frequency - Inverse Document Frequency

Numerical statistic to reflect how important a word is to a document in a collection of documents.

How many times a term appears in a given document vs how many documents that term is present in.

Utilized the lemmatization linguistic morphology.

```
params{min_df= , max_df= , ngram_range= max_features= ,}
```

# Random Forest

Ensemble Learning for Classification

Random Forest is an ensemble of single decision trees to create a forest of predictors.

Robust against the overfitting issues of a single decision tree.

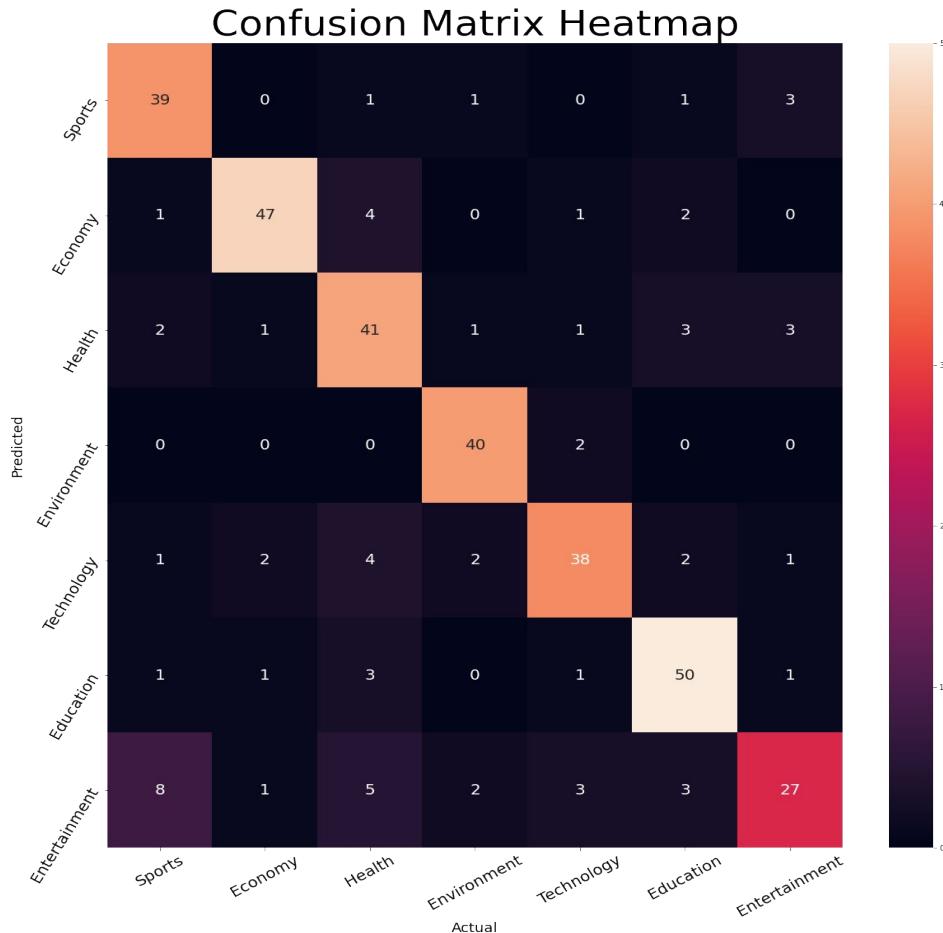
```
params{max_depth= , max_features= , n_estimators= ,  
min_samples_leaf= , min_samples_split= }
```

# Confusion Matrix

High misclassification between technology and economy.

A number of economic articles discuss F.A.A.N.G. companies.

Which are all organizations entrenched in technology.



# Performance

Training and Validation:

Trained on 1050 articles

Validated on 350 articles

Accuracy of 81%

F1 Score of 80%

Testing:

Tested on 210 articles

Accuracy of 80%

F1 Score of 80%

# Next Steps

Deploy model with continued supervision.

Hire interns to confirm the labels are correct.

Scrape more articles and continuously train the model.



# Thank You!!

Questions???

Comments!!!



# Appendix

# Stemming and Lemmatization

Stemming is the process of converting a word into its stem, base, or root form. Can often be a form that doesn't exist in the dictionary.

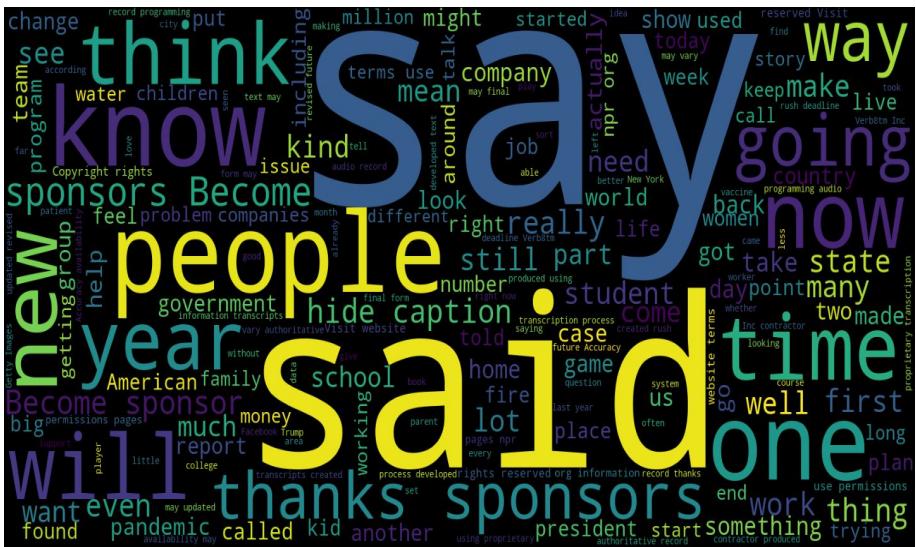
Lemmatization is the process of grouping together the inflected forms of a word and converting them to a common lemma or dictionary form.

# GenSim Summarization

In an effort to decrease the potential abundance of superfluous sentences that'll noisey up the model we implemented a summarization algorithm.

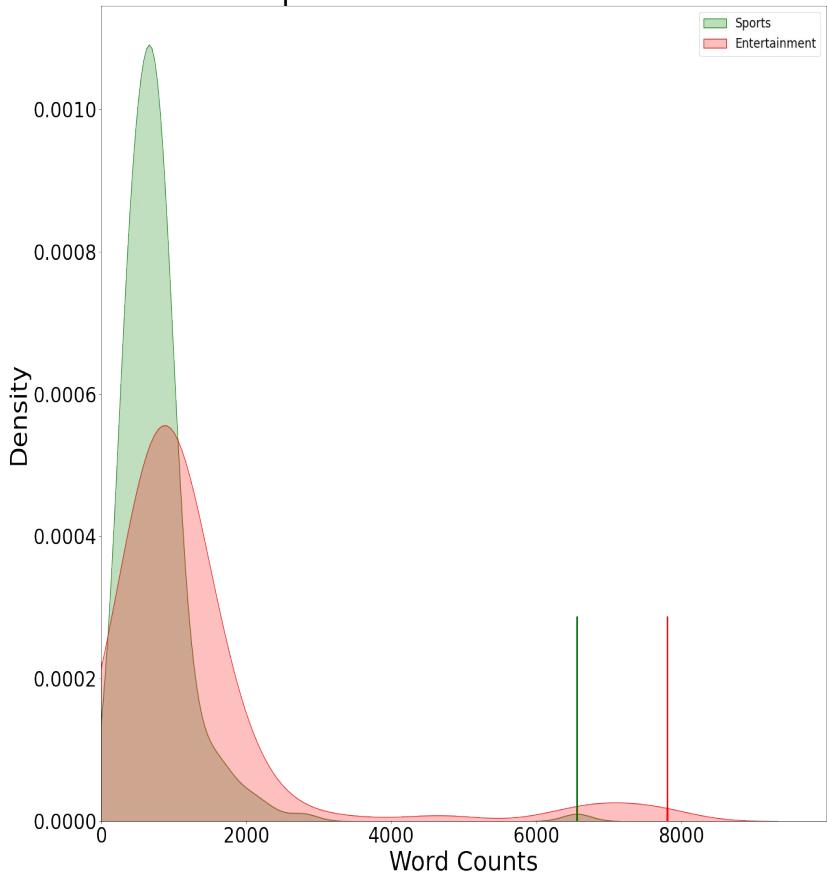
Summarization is based on ranking the sentences via the TextRank algorithm. TextRank is a graphical based ranking model to find relevant sentences and keywords.

Params{ ratio=.2 & word\_count=375 }



# Sports vs Entertainment

Sports and Entertainment

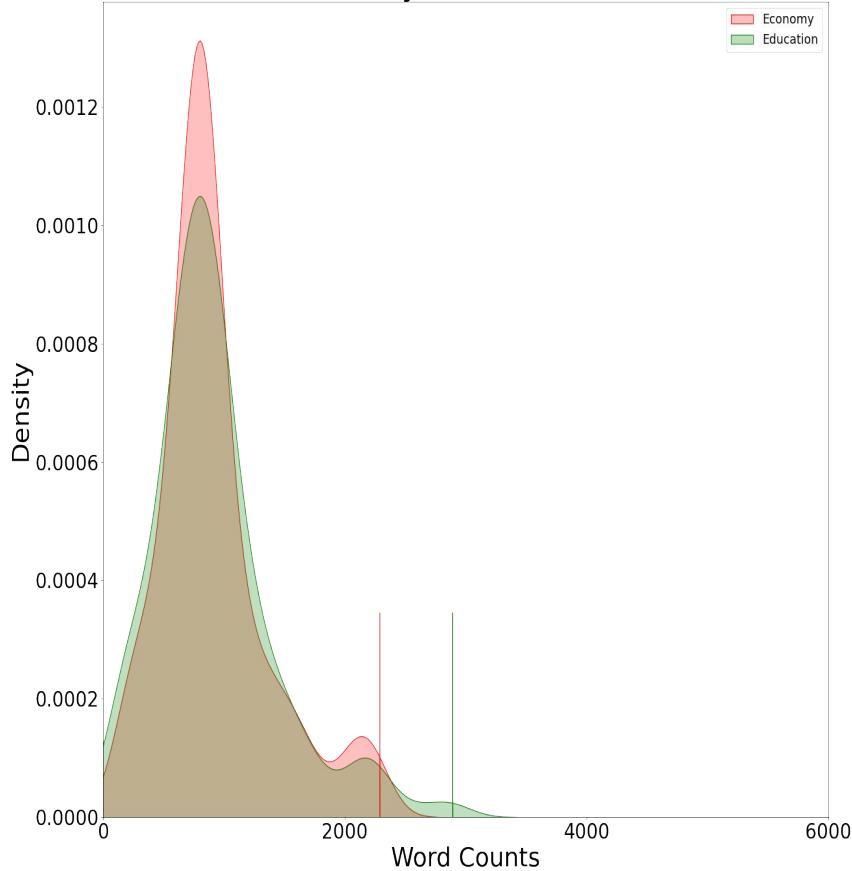


Sports had the lowest average word count while entertainment had the highest average word count per article.

Most articles for both categories were below 2000 but extreme outliers existed for both.

## Economy vs Education

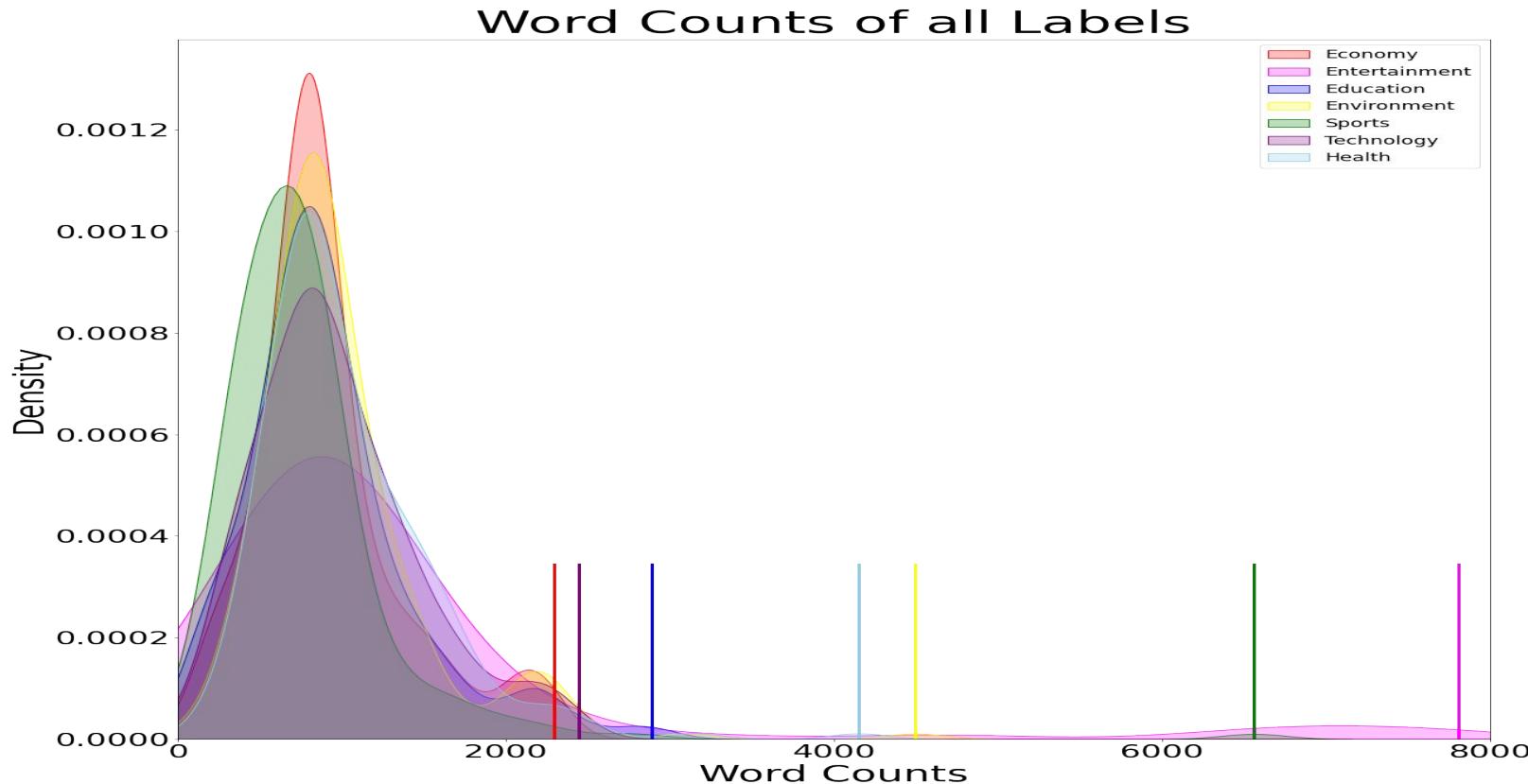
Economy and Education



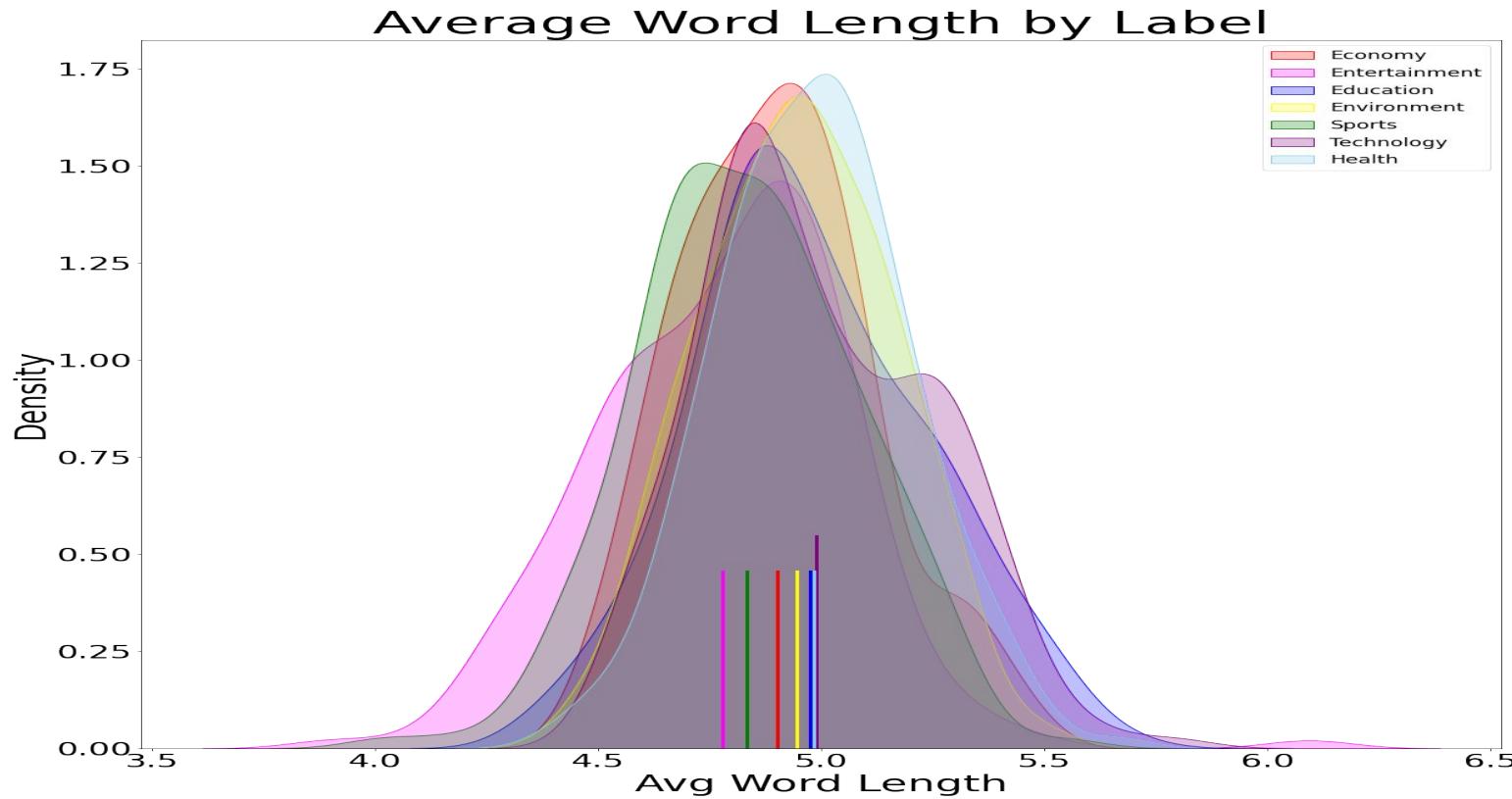
Very similar in terms of average word count per article as such the distributions are also quite similar.

Education outliers led to the expansion of the distribution.

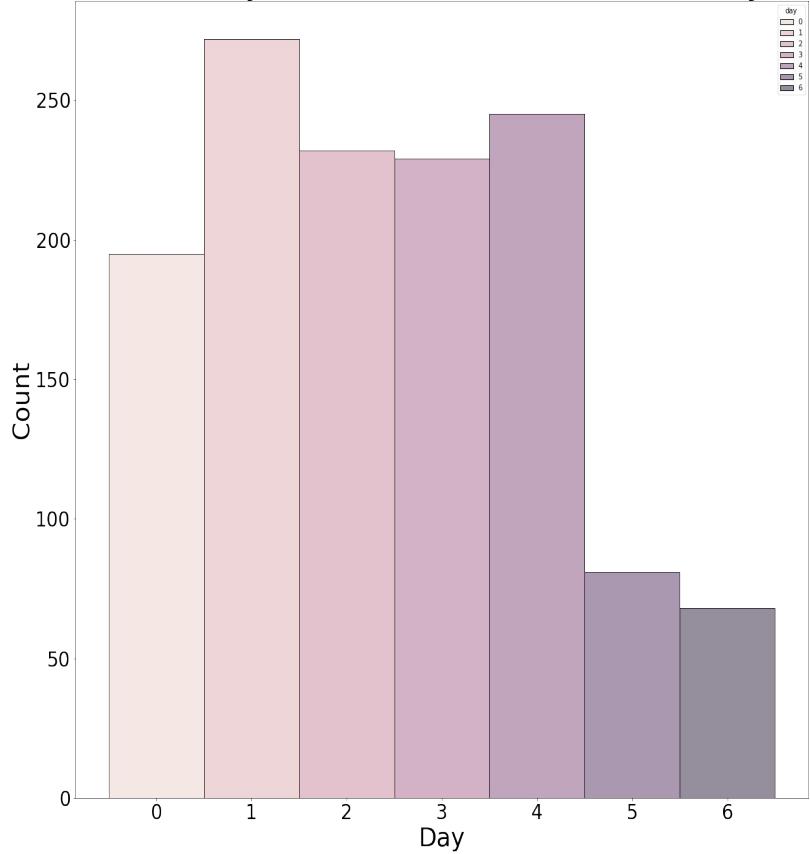
# Average Word Counts (All Labels)



# Average Word Lengths (All Labels)

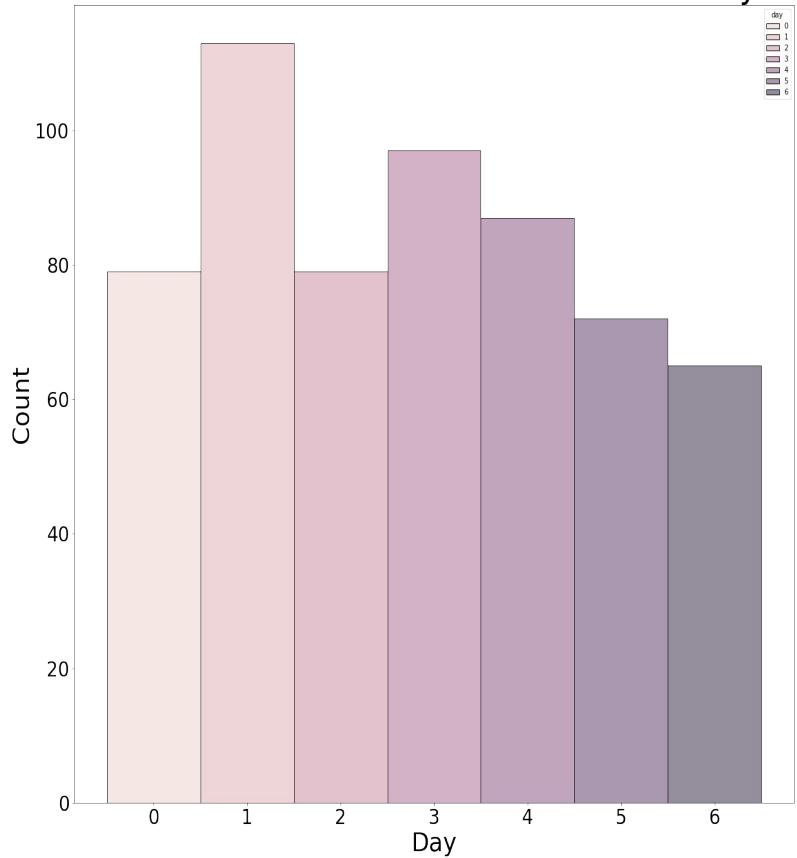


Economy news articles based on the day



News articles for Economy tend to be at the start of the week. Peaking at Tuesday before tapering into a calm weekend.

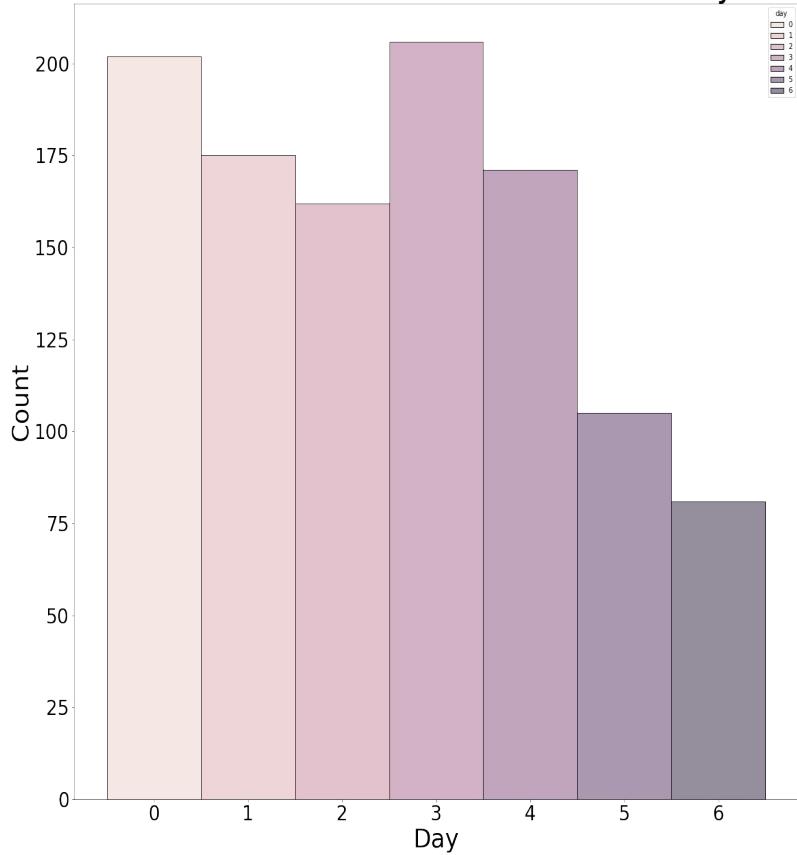
Education news articles based on the day



News articles for Education tend to be at the start of the week. Peaking at Tuesday before tapering into a calm weekend.

---

Health news articles based on the day

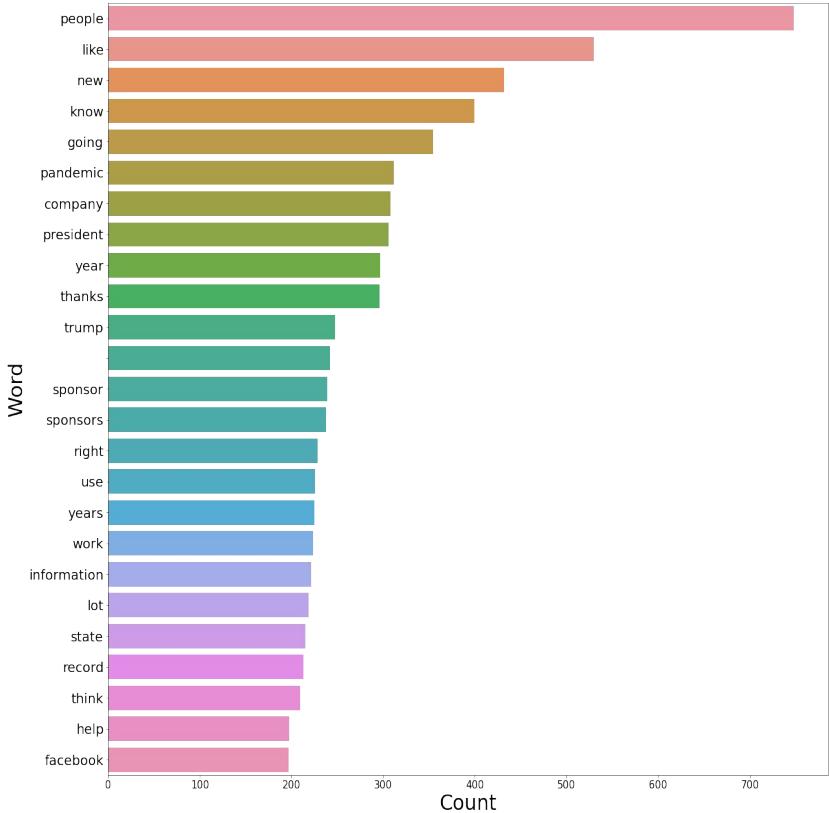


News articles for Health tend to be at the start of the week. Peaking at Thursday before tapering into a calm weekend.

---

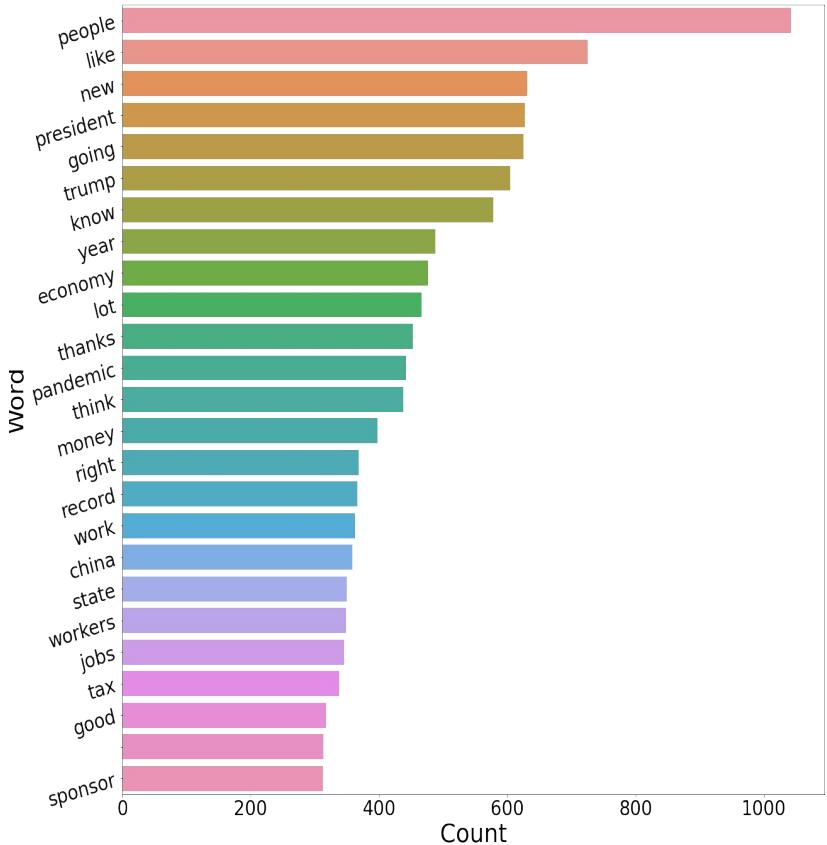
# Business

Word Count Distribution For Business Articles

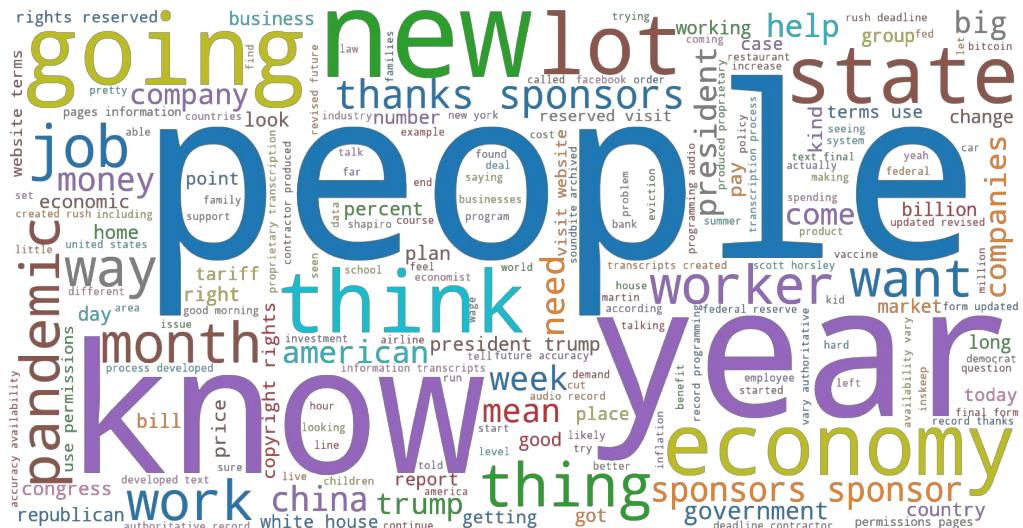




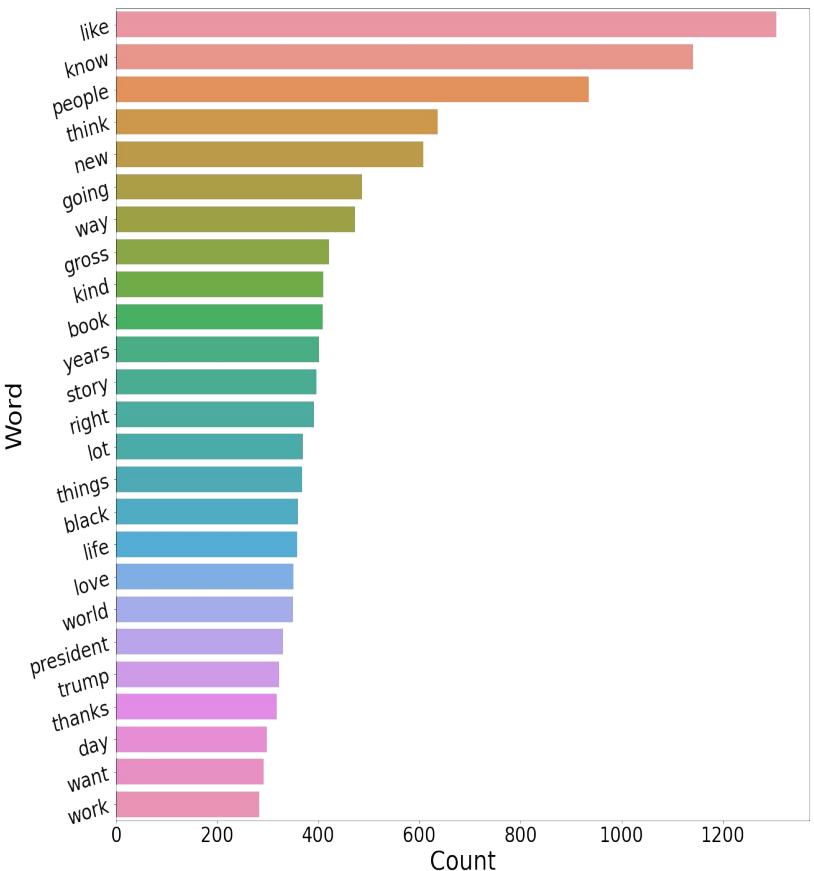
## Word Count Distribution For Economy Articles



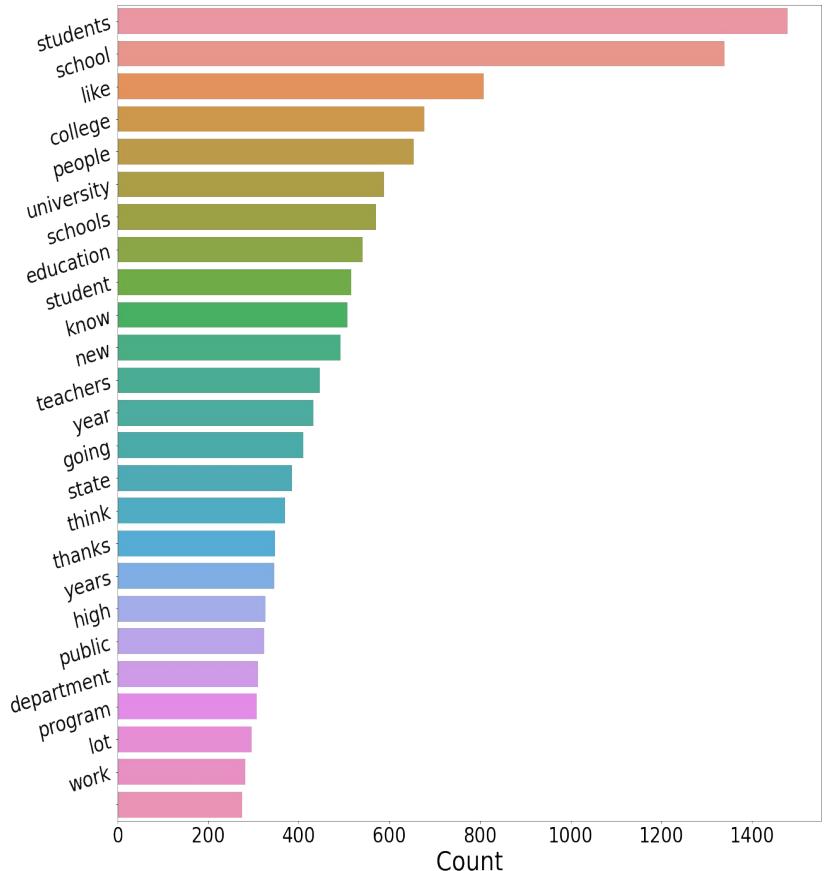
# Economy



Word Count Distribution For Entertainment Articles



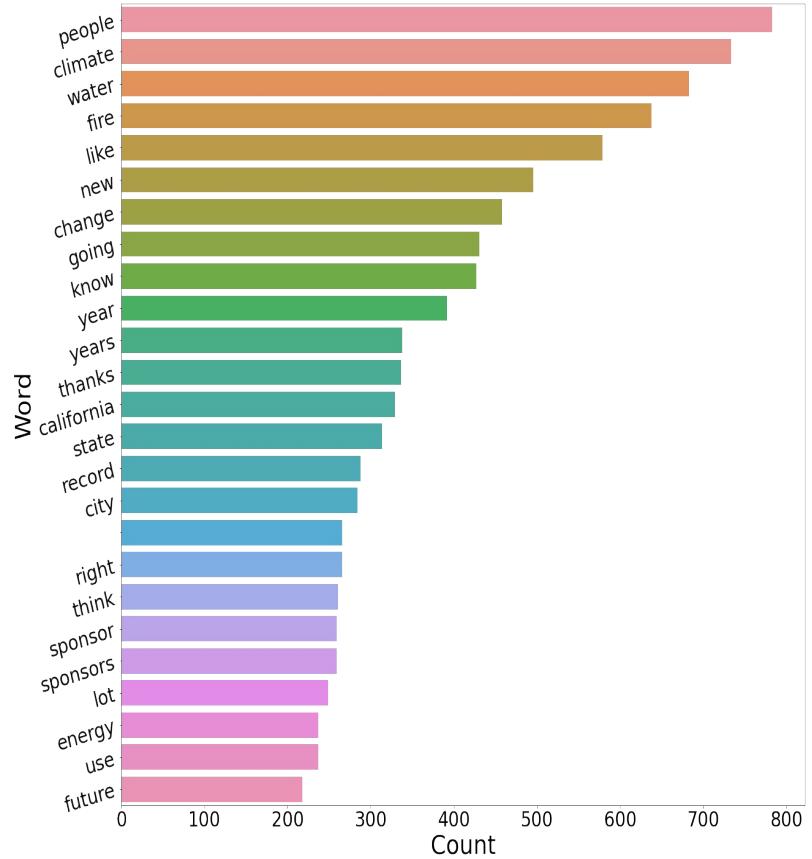
Word Count Distribution For Education Articles



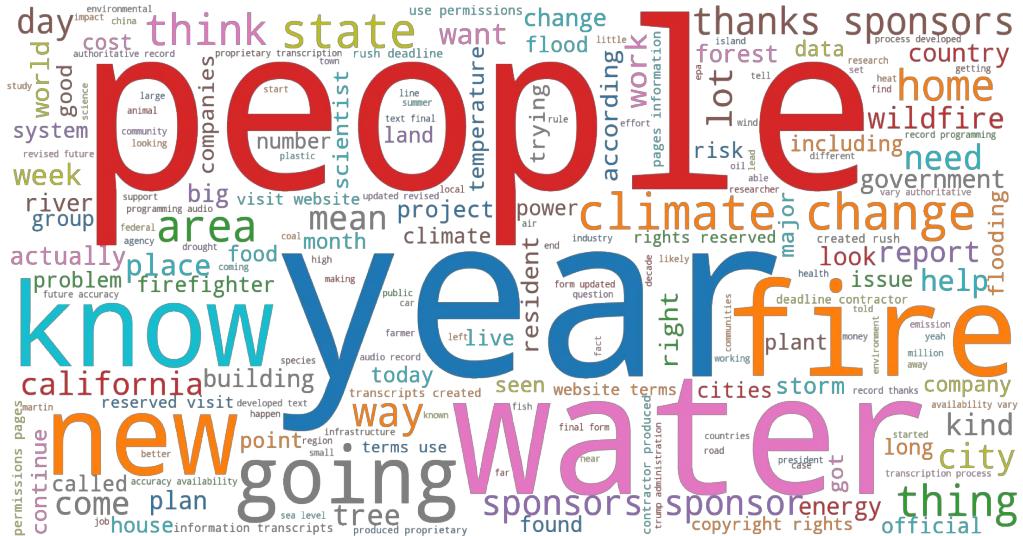
# Education

A large, colorful word cloud centered around the word "Education". Other prominent words include "School", "Student", "Teacher", "Year", "New", "Program", "Way", "Know", "Think", "Work", "Help", "Place", "Kid", "Campus", "Case", "Need", "Got", "Sponsor", "Student", "Public", "University", "High", "School", "Education", "Parent", "Families", "Talk", "Support", "Visit", "Website", "Found", "Instead", and many smaller words like "Record", "Thanks", "Borrower", "Life", "Find", "Live", "Worker", "Set", "Likely", "Long", "End", "Friend", "Page", "Rule", "Strike", "Pandemic", "Community", "Public", "Company", "Point", "Study", "Data", "Open", "Graduate", "Future", "Law", "Team", "Revised", "Good", "Rush", "Deadline", "Feel", "Look", "Big", "Call", "Thing", "Week", "Right", "Big", "Developed", "Text", "Book", "Education", "Proprietary", "Accuracy", "Available", "Future", "Accuracy", "Families", "Talk", "Support", "Visit", "Website", "Found", "Instead", and "Instituted".

### Word Count Distribution For Environment Articles

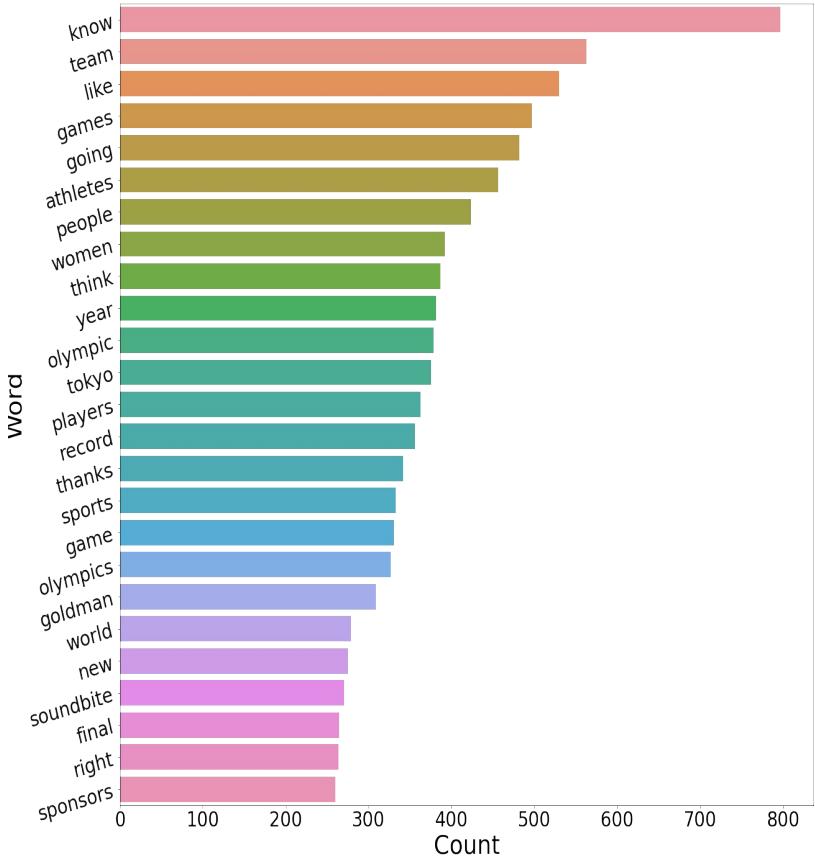


# Environment





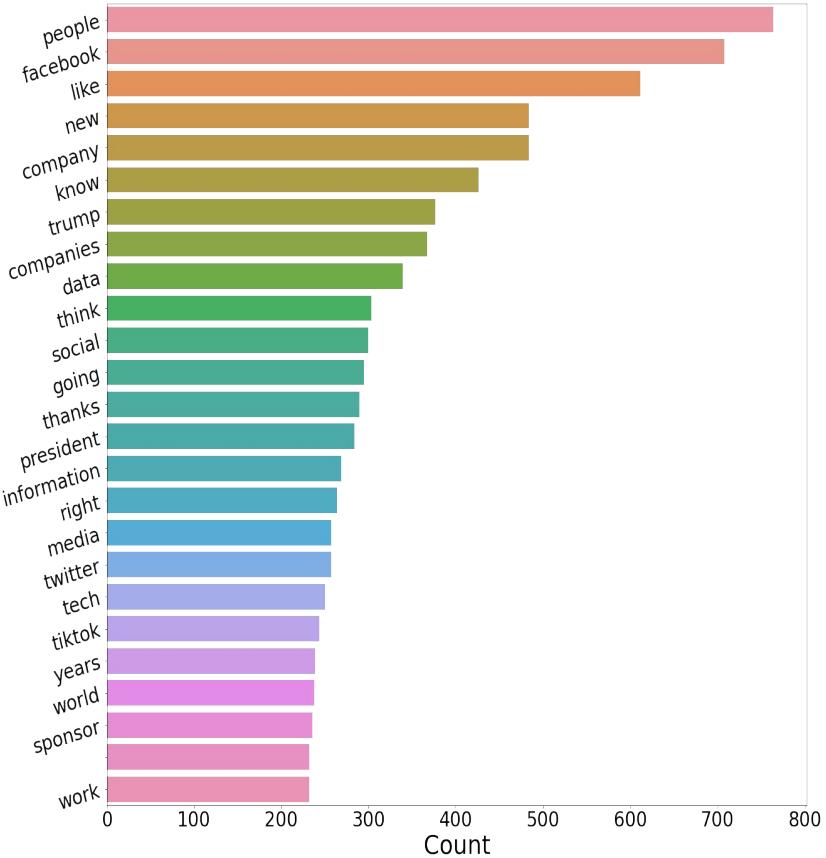
## Word Count Distribution For Sports Articles



# Sports



## Word Count Distribution For Technology Articles



# Tech

