Book info TBD

Vocabulary Entropy in Reddit discussions of US Politics

Nicholas A. Lines*
Ian A. McCulloh*
nicholasalines@gmail.com
imccull4@jhu.edu
Johns Hopkins University
Baltimore, Maryland, USA

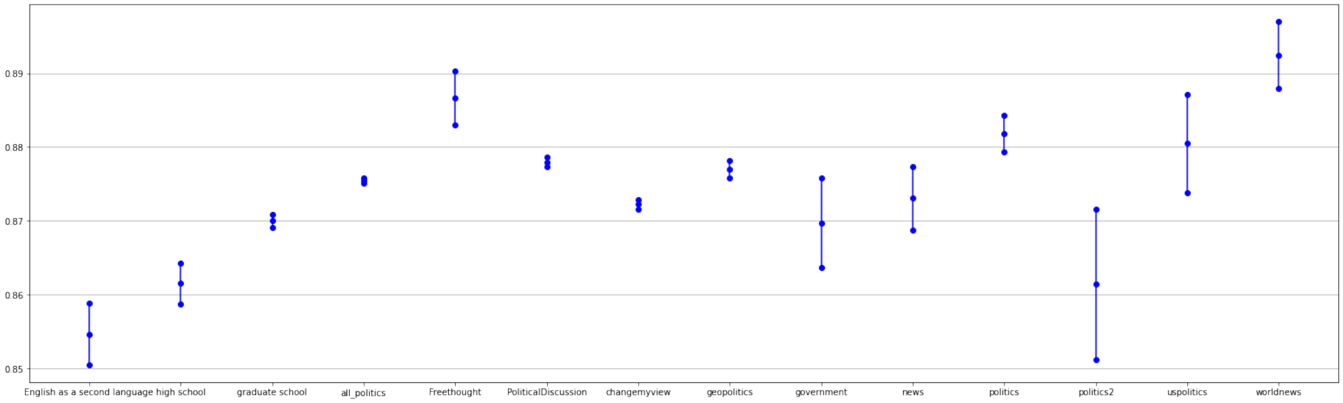

Figure 1: Confidence intervals for the mean vocabulary quotient.

We show here the 95% confidence intervals for the mean population vocabulary quotient of various subreddit groups.

ABSTRACT

Distance reading and text mining applications have suggested the need for more language-specific techniques that can be employed in text analysis. We propose the application of vocabulary entropy to characterize groups of authors, and use this approach to analyze and contrast US political discussions carried out on the social media platform Reddit. We show that normalized sample mean vocabulary entropy of a community corresponds well with education and literacy level, and that vocabulary entropy in political forums varies widely by forum.

CCS CONCEPTS

• **Networks → Social media networks**; • **Information systems → Data mining**; • **Computing methodologies →** *Natural language processing*;

KEYWORDS

social media analysis, text mining

ACM Reference Format:

Nicholas A. Lines and Ian A. McCulloh. 2021. Vocabulary Entropy in Reddit discussions of US Politics. In *Proceedings of Conference TBD.* ACM, New York, NY, USA, 5 pages. https://doi.org/TBD

1 INTRODUCTION

As the amount of text-based media has ballooned in recent decades, it has become clear that the ability to characterize and discuss text corpora at an abstract level is essential. In the literary community this arose as the concept of distance reading, in which literary analysis is pursued without humans actually reading the books [5]. In the statistics and data science communities we find this problem addressed by various text mining and authorship verification techniques. The popularity of topic modeling and other language-agnostic techniques suggests that there is a rising market for universal methods for characterizing author groups and text corpora without relying on language-specific features.

A potentially rich area for research is vocabulary analysis, since all written language relies on a lexicon of vocabulary. Vocabulary size, independent of the exact selection of words in that vocabulary, has long been strongly correlated with mental maturity and other features of interest such as fluency in one or more languages, education level, etc. [6][3] Efforts to measure the true size of individuals vocabularies, however, tend to be onerous: a recent online study with almost 300,000 participants tested vocabulary sizes of Dutch speakers, checking their knowledge of words against a dictionary 53,000 words in length [4]. An alternative to vocabulary size estimation is the much simpler task of vocabulary entropy measurement, since the variety in word use is dependent on the number

*This work was completed as a project for a course taught by Ian A. McCulloh.

Conference TBD, Location, TBD,
© 2021 Association for Computing Machinery.
ACM ISBN ISBN TBD…$Price TBD
https://doi.org/TBD

of words known. We propose that groups of authors can be effectively described in terms of vocabulary entropy. Using a dataset of English Reddit posts and comments, we show that vocabulary entropy is related to education and fluency levels, and then use this tool to explore the landscape of US politics discussion forums on Reddit.

2 BACKGROUND

2.1 Related work

To the best of our knowledge, this is the first discussion published on the subject of author group characterization using vocabulary entropy. However, vocabulary has long been a subject of interest for similar studies. Authorship verification in academic and court evidence has often relied on vocabulary analysis, including measures of vocabulary richness [1]. Some of Claude Shannon's original work in entropy was motivated by a study of variety in English character use [8]. In 2000, the authors of [2] discussed several measures of vocabulary richness, in the context of author identification, including normalized Shannon entropy. In 2018 an identical approach for measuring a single author's vocabulary richness was independently proposed in [7], and called the *vocabulary quotient*. It is important to note that both publications described this normalized entropy measure as independent of the length of the text in words, and this is not the case: the vocabulary quotient still decreases with sample size. However, we will use the vocabulary quotient throughout this work to remain consistent with the most recent publications related to the subject.

2.2 Vocabulary quotients

Let n be the length in words of a text sample in words composed using V distinct words $X_1, X_2, ..., X_V$. These words's frequencies are observed and used to write the vocabulary's empirical probability distribution for the random variable for word choice, X. The Shannon entropy of the document is defined to be

$$H(X) = -\sum_{i=1}^{V} P(X_i) \log_{10} P(X_i).$$

However, vocabulary use in writing follows Heap's law, which states that the number of unique words observed, V, in a document of length n words is given by

$$V(n) = \alpha n^{\beta},$$

where α and β are parameters dependent on the author(s) and possibly the language and context of the document. This means that document length will greatly affect the entropy measure. As in [2], [7] we use the vocabulary quotient $\frac{H(X)}{-\log_{10}(1/n)}$, the entropy divided by the maximum entropy of a document of the same size, to dampen this effect, and also restrict our review to universal fixed sizes of n to ensure true consistency in entropy measures.

3 METHODOLOGY

3.1 Data collection

Reddit is a social media microblogging platform that consists of many individual communities called subreddits, whose posts are focused on a particular subject. Users may post original content, links, etc, and also provide comments on posts. Using the Python Reddit API Wrapper (PRAW), we collected samples of the most recent posts and comments from groups of subreddits associated with non-native English speakers, high school students, graduate school students, and US politics. Since some of the subreddits surveyed are sparsely used, some of the collected posts date as far back as 2012, though most occur in spring of 2021. Each post or comment was tokenized using the NLTK "casual tokenize" strategy (designed for processing English Twitter data), after having its punctuation and any handles (user names) stripped and the casing leveled. Any post shorter than 150 tokens was discarded, and the remaining posts were truncated to the first 150 tokens. The resulting sample sizes are shown in Table 1.

Table 1: Subreddit sample sizes

Subreddit group	Sample size
English as a second language	261
High school	496
Graduate School	2370
All political subreddits below	11937
Freethought	98
PoliticalDiscussion	4576
changemyview	5345
geopolitics	1246
government	62
news	81
politics	314
politics2	54
uspolitics	66
worldnews	95

3.2 Vocabulary quotient computation

The vocabulary quotient was then computed for each sample text, and the means recorded for each subreddit group. We found a 95% confidence interval for each subreddit group's true mean vocabulary quotient using the sample variances and the Normality of the sample means. The resulting distributions are shown in Figure 2. We also processed the comments and posts individually, and found that the mean vocabulary quotient of comments was usually higher than that of the original posts. We repeated the experiment with sample sizes of $n = 100$ but did not observe any significant difference in the results.

4 RESULTS

4.1 Analysis

The confidence intervals for population means of the subreddit groups, given our samples, are shown in Figure 1. This confirms our expectation that the casual conversation of graduate students exhibits higher vocabulary entropy than that of high school students or English learners in similar forums (note that we do not use any professional writing, such as essays or dissertations, but instead use only Reddit-style conversation). The distribution of high-school speech is extremely broad, suggesting that students

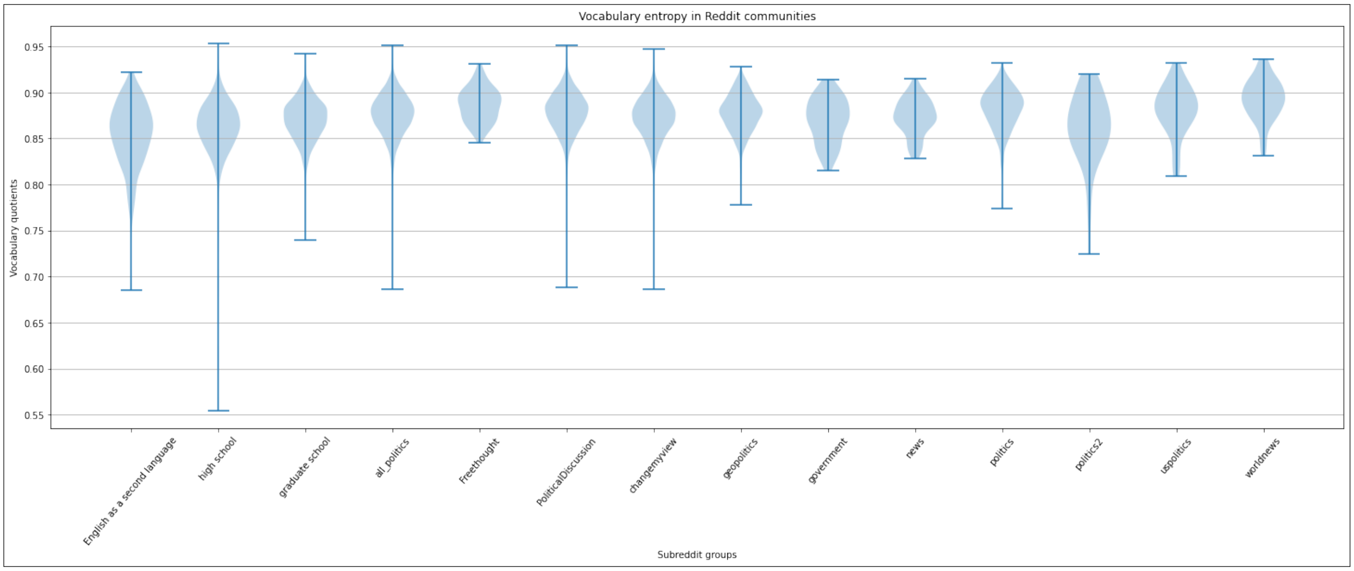

Figure 2: Distributions of vocabulary quotients in various subreddit groups.
These violin plots show the distribution of vocabulary quotients in each of the subreddit groups examined in this study.

with high vocabulary entropy show that potential before college education.

To our surprise the political discussion subreddits, on average, show higher vocabulary entropy than any of the above student groups, suggesting more linguistic sophistication in the discussion content. Individual subreddits vary, of course, with some closer to high-school level writing and others significantly higher than graduate school level writing. There are several facts that contribute to the unusually high scores. The highest performing subreddit, worldnews, is mainly devoted to linking and quoting news articles, rather than posting original content, so it is not surprising that the vocabulary entropy is similar to that of professional writing. In contrast, politics2 is a subreddit frequented by users posting from a variety of world regions and contains much more original, unprofessional content, which explains its extremely low vocabulary entropy.

4.2 Future Work

These first steps into vocabulary analysis of social media groups suggest many further areas to investigate. While we did not explicitly check for duplication in the data collection process, we did review the results to show that there were not any frequently occurring quotient values, so it appears this is a minor concern. However, many subreddits have official and unofficial bots that create posts and comments, and these may well be artificially boosting the group's mean quotient. We might also control for quotations and group-authored content, since this is known to exhibit higher vocabulary entropy. Improvements could also be made in the text cleaning approach such as a dictionary lookup, since we found some hyperlinks and emojis slipped through the casual tokenizer. Despite this limitation, the highest scoring posts and comments in our data were found to genuinely exhibit diverse vocabulary and more professional writing.

A severe limitation of all micro-blogging data is the minimal length of posts hosted by the platforms. Because the length filtering could not occur within PRAW's scraping routines, over 87% of the collected posts and comments collected had to be discarded. This sampling bias toward unusually long posts means that our analysis is not actually representative of the whole Reddit community. An important question to investigate is just how low the sample size n can be dropped before the vocabulary quotient is uninformative.

Since Reddit is principally home for informal writing, it would be informative to contrast these results with samples from more formal writing such as student essays, LinkedIn posts, or school blogs. How much an individual or group vary their vocabulary entropy when switching from informal to formal settings is a question of interest. And finally, we also wish to investigate the variance of vocabulary quotients across multiple languages.

5 CONCLUSION

We have shown that the vocabulary quotient of a group of authors can indicate their education and literacy level, and that the content of recent US political discussion threads on Reddit exhibits a high degree of linguistic sophistication in terms of vocabulary entropy. Individual subreddits vary significantly in their vocabulary entropy, and these variations are correlated with features specific to those subreddits such as frequent quotations from professional journalism. All code related to this paper, including that used for data collection, is provided at https://github.com/linesn/reddit_analysis.

ACKNOWLEDGMENTS

This work was done under the instruction of Dr. McCulloh during a course on Social Media Analysis at JHU.

REFERENCES

[1] C. Chaski. 2001. Empirical evaluations of language-based author identification techniques. *International Journal of Speech Language and The Law* 8 (2001), 1–65.

[2] R. Dale, H. Moisl, and H. Somers. 2000. *Handbook of Natural Language Processing*. Taylor & Francis. https://books.google.com/books?id=VoOLvxyX0BUC

[3] Charlotte Fox and James E Birren. 1949. Some factors affecting vocabulary size in later maturity: Age, education, and length of institutionalization. *Journal of gerontology* 4, 1 (1949), 19–26.

[4] Emmanuel Keuleers, Michaël Stevens, Paweł Mandera, and Marc Brysbaert. 2015. Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *The Quarterly Journal of Experimental Psychology* 68, 8 (2015), 1665–1692.

[5] Franco Moretti. 2000. Conjectures on world literature. *New left review* 1 (2000), 54.

[6] ISP Nation. 1993. Vocabulary size, growth, and use. *The bilingual lexicon* 6 (1993).

[7] Nikhil Kumar Rajput, Bhavya Ahuja, and Manoj Kumar Riyal. 2018. A novel approach towards deriving vocabulary quotient. *Digital Scholarship in the Humanities* 33, 4 (2018), 894–901.

[8] Claude E Shannon. 1951. Prediction and entropy of printed English. *Bell system technical journal* 30, 1 (1951), 50–64.