

How Does Content Drive Viewership?

Dave Holtz¹, Jeremy Yang², Michael Zhao³

Abstract

Why do some webpages receive massive numbers of page views? To determine how content drives viewership, we construct a unique dataset of all articles published by the New York Times (NYT) in August 2013. Our dataset is built from 2 major components, the NYT's internal web traffic data and article content data parsed from the NYT website. We use the internal web traffic data to accurately track the number of page views of each article as well as construct a set of robust control variables such as the desk and section of each article. To build content features, we use various machine learning and statistical natural language processing techniques on our parsed article content data and construct features such as article perplexity, sentiment, reading difficulty, and indicators that denote the presence of pictures, videos, etc. Additionally, we have access to the NYT's internal website traffic data. We feed all of our constructed features to into a predictive regression model. We find [MAJOR RESULTS HERE].

¹dholtz@mit.edu

²zheny@mit.edu

³mfzhao@mit.edu

Contents

1	Introduction	1
2	Data	2
2.1	NYT Internal Web Traffic Data	2
2.2	Parsed NYT Article Content Data	2
3	Constructed Features	3
3.1	Article Sentiment	4
3.2	Article Perplexity	4
4	Predictive Regression Model	4
	Acknowledgments	5
	References	5

1. Introduction

In today's digital economy, many companies are very interested in attracting users to visit their websites in order to earn ad revenue. While many factors might motivate a user to visit a particular page, certainly one important factor is the content in that webpage. This paper explores the relationship between the content of a webpage and the number of

page views it ultimately ends up receiving by constructing a unique dataset of all articles published by the New York Times (NYT) during August 2013. This dataset is built from two major components: the NYT's internal web traffic data and parsed NYT article content data.

Typically, a study such as ours tends to be very difficult to conduct as either accurate measures of viewership are unavailable¹ or the feature extraction of the content is too challenging (for example Youtube), or or both. Fortunately, our access to the the NYT's internal web traffic data allows us to exactly measure the number of page views an article receives. The web traffic data is rather rich and also includes internal meta-data that we use to build various control features. Moreover, since we are working with mostly textual data, we are able to take advantage of recent advancements in machine learning and statistical NLP to do feature extraction on article text.

¹While oftentimes precise viewership data tends to be not available openly, oftentimes researchers use related observables, such as Facebook likes

A similar study by Berger and Milkman (2012) [1] examines the relationship between content and word-of-mouth virality. They find that the emotional content of a NYT article is predictive of its virality. Using simple measures of an article's sentiment and emotionality, Berger and Milkman show that positive articles are more likely to show up on the New York Times "Most-Emailed" list. They also show that articles that evoke high physiological positive or negative arousal (such as awe or anger) tend to be more viral than articles that evoke deactivating emotions (sadness). We build on this study in two ways: first, we relate an article's content back to the number of page views it receives rather than its virality². Second, we employ more sophisticated machine learning feature extraction techniques to see if they work any better over their simple measures.

2. Data

2.1 NYT Internal Web Traffic Data

Our NYT internal web traffic dataset is a record of all individual user activity on the NYT website covering the period of April 3rd, 2013 to October 31st, 2013. This activity data is stored as individual lines of json and includes who (if available) accessed what page at what time. Overall, it is over 20 terabytes in size and contains over 3 billion page views³. Since the scope of this dataset is so large, we initially restrict this project to a single month, August 2013.

We limit our dataset to consist of pages that only contain articles or blogposts published during the month of August. We parse the data to obtain a list of urls, which need to be stripped of potential garbage. After cleaning up the url data, we are left with 6682 unique pieces of content. We then parse our dataset and aggregate the number of counts each url receives. In order to make the comparison between articles fairer since an article

that's been out longer will have more page views on average, we only count the number of page views received up to 7 days after publication⁴. In total, our data consists of over 250 million page views. As seen in Figure 1 below, the distribution of page views is highly skewed with very heavy tails. After applying a log transformation (as seen in Figure 2), our distribution looks considerably more normal.

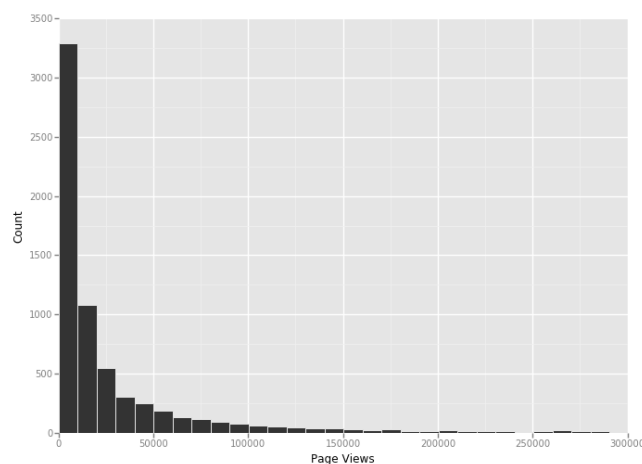


Figure 1. Histogram of Articles by Number of Page Views

Table 1. Page Views Distribution Summary Statistics

Total Page Views	248161455
Min	1
Max	2545288
Mean	37138.8
Median	10298.5
Std. Dev.	88972.9
Skewness	9.52191
Kurtosis	173.061
Observations	6682

In addition to aggregating the counts, when

2.2 Parsed NYT Article Content Data

²Which companies arguably care more about since word-of-mouth virality is usually a means to increase page views

³Not all page views are content views, for example, some events that are also tracked are searches, or user account settings.

⁴Given that page views tend to sharply drop off soon after publication since recency is quite important to the News, the number of page views obtained during the 7 days after an article is published represents the vast majority (usually well over 90%) of total page views an article receives.

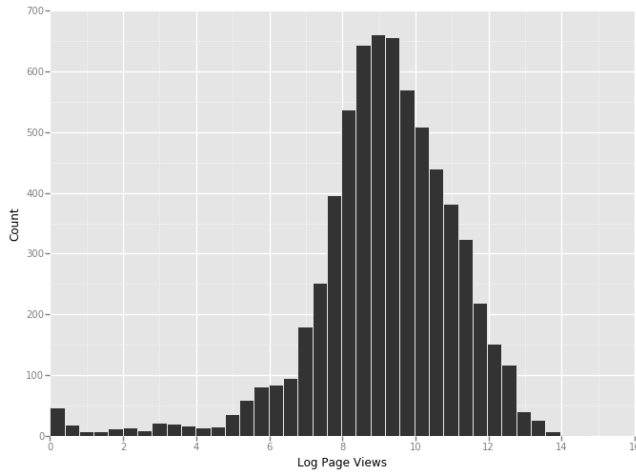


Figure 2. Histogram of Articles by Log of Page Views

Table 2. Log Page Views Distribution Summary Statistics

Min	0
Max	14.74975
Mean	9.122868
Median	9.239754
Std. Dev.	2.028668
Skewness	-1.270368
Kurtosis	3.800911
Observations	6682

3. Constructed Features

In preparing to perform our regression, we construct a number of features, using both the parsed NYT article data and secondary sources. These features include the Flesch reading ease, the (guessed) gender of the author(s), the popularity of the author(s), the sentiment of the article text, the perplexity of the article text, and variables indicating the section the article appeared in and the article's content type. Found below is a full list of these features, as well as the methodology used to construct them and validation of the features, where appropriate.

Flesch Reading Ease The Flesch reading ease is a metric developed by Flesch in 1948 [2]. The score indicates how difficult a piece of English text is to understand. Lower scores correspond to more difficult passages, and

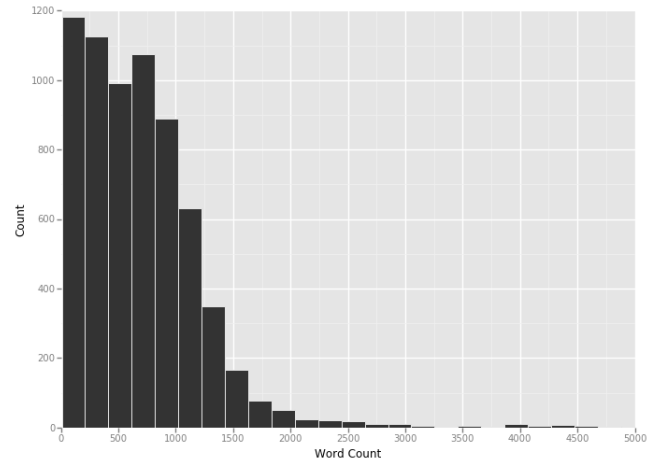


Figure 3. Histogram of Articles by Word Count

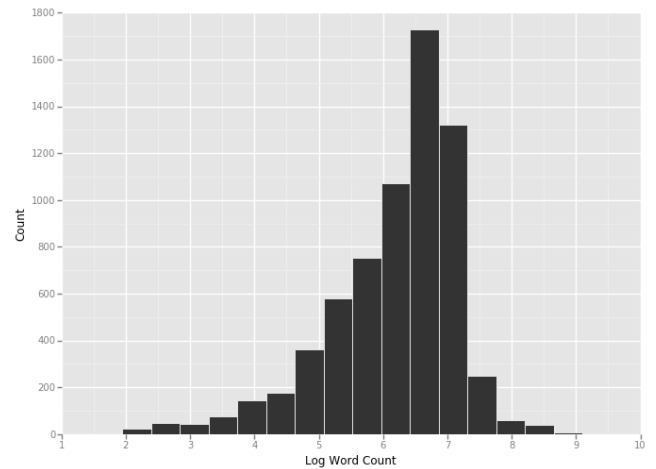


Figure 4. Histogram of Articles by Log Word Count

the highest score attainable is 120.0. The formula for calculating a passage's Flesch reading ease is

$$206.835 - 1.015 \left(\frac{\# \text{ words}}{\# \text{ sentences}} \right) - 84.6 \left(\frac{\# \text{ syllables}}{\# \text{ words}} \right) \quad (1)$$

To calculate the Flesch reading ease, we use the python library "textstat." Despite the fact that the above formula is relatively straightforward, the task of counting the number of syllables in a block of text is non-trivial, so we rely on "textstat" to do so accurately. In cases where the Flesch reading ease was for

some reason null (e.g., a blog post containing only a picture), we assign the Flesch reading ease its median value.

Author Popularity We attempt to include some measure of a particular author’s popularity. It stands to reason that a new article by Paul Krugman or A.O. Scott should garner more readership than a new article by an unknown graduate student enrolled in 6.867 at MIT!

In order to measure something that will serve as a decent proxy for popularity, we programmatically searched for every distinct author in our dataset using Bing, and recorded the number of search results that were returned by the query. In cases where a particular article has more than one distinct author, we calculate an “effective” popularity by simply calculating the average number of search results over all article authors.

Author Gender We also attempt to construct a feature that indicates the most likely gender of the article author(s). In cases where the gender of the author is unclear (e.g., Robin) or there are likely multiple authors with different genders (e.g., The New York Times Staff), we record a third gender value, “ambiguous / unknown.”

We construct our gender data by cross-referencing the first names of all of the authors in our dataset against U.S. Social Security Administration baby name data from 1935 to 1997. If over 90% of the babies with a given name have been male, we assume a given author is male. If over 90% of the babies with a given name have been female, we assume a given author is female. Otherwise, we record “ambiguous / unknown.”

Material Type, Section, Desk, and Article Type

We also include dummy variables including the material type (e.g., ‘News’ or ‘Obituary’), desk (e.g., ‘Weekend’ or ‘Real Estate’), article type (‘Blog post’ or ‘Article’), and section (e.g., ‘Movies’ or ‘World’).

We also include features that attempt to capture the article sentiment and the article text perplexity. Since the construction of these features was more involved and involved validation of our algorithms, we discuss these two features in separate subsections.

3.1 Article Sentiment

3.2 Article Perplexity

4. Predictive Regression Model

Feature	Weight
intercept	9.047
log_wcount	0.729
desk_Foreign	0.657
section_World	-0.553
desk_Travel	-0.472
typeOfMaterial_Schedule	-0.291
section_Movies	-0.277
section_Opinion	0.261
typeOfMaterial_Review	-0.255
desk_National	0.253
typeOfMaterial_Letter	-0.243
section_Sports	-0.215
section_Education	-0.199
section_Books	0.192
section_Corrections	-0.179
desk_Weekend	-0.175
section_Health	0.172
type_BlogPost	-0.170
desk_BookReview	-0.169
typeOfMaterial_Op-Ed	0.167

Acknowledgments

References

- [1] Jonah Berger and Katherine L Milkman. What makes online content viral? *Journal of marketing research*, 49(2):192–205, 2012.
- [2] Rudolph Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948.

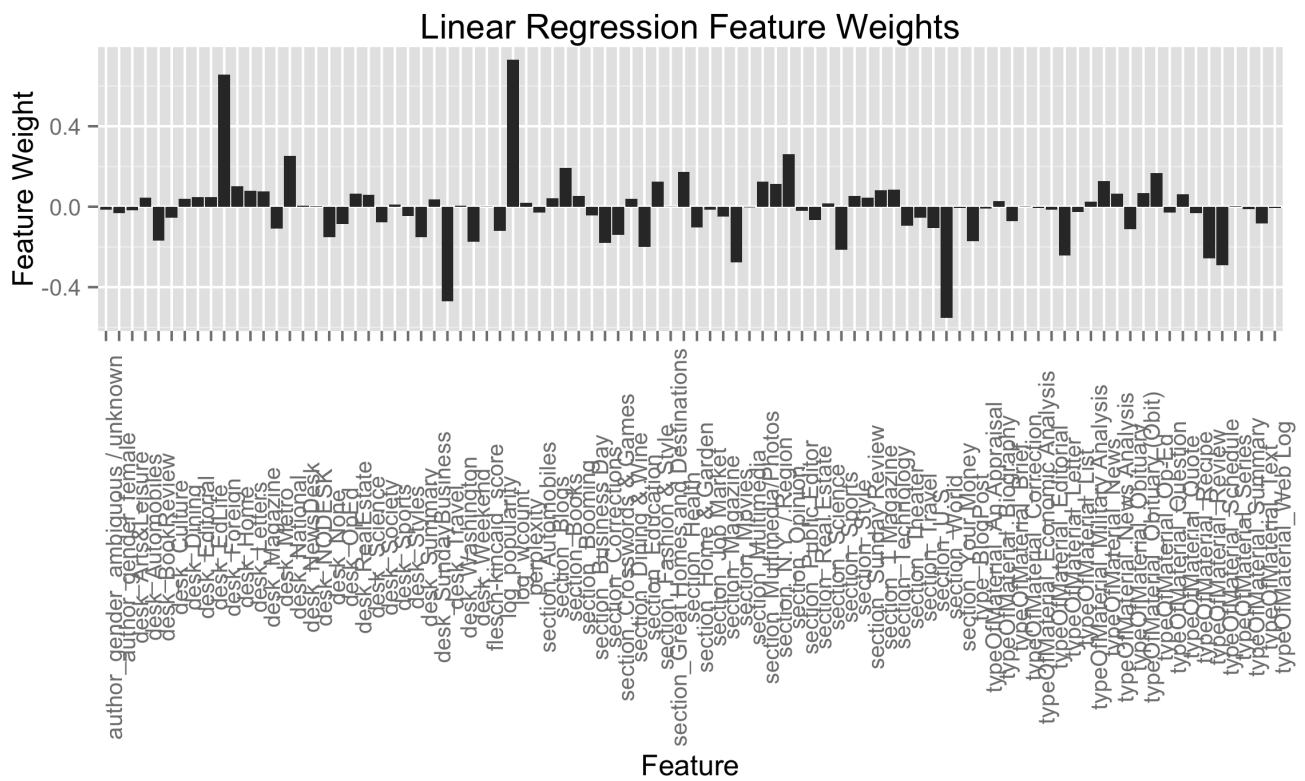


Figure 5. Linear Regression Feature Weights (excluding intercept term)