# Are Good Headlines Simpler: A Study from the Upworthy Research Archive

Mingjiamei Zhang

*Department of Physics, The University of Chicago*

(Dated: November 12, 2023)

# I. BACKGROUND AND MOTIVATION

## A. The Upworthy Research Archive

Upworthy Research Archive is a dataset collected by Upworthy.com which covers the results of A/B tests from early 2013 into April 2015 [1]. In each test, Upworthy randomly assigned different readers to see different packages which contains a headline and an image for the same story. The content management system recorded the number of participants that were shown a given package (*impressions*) and the number that clicked on the package (*clicks*). After a period of time, an editor would review a dashboard that reported the results and either conduct an additional experiment (potentially with new packages) or choose which package to finalize for a given article. Editors sometimes finalized a package other than the best performing one. From that decision point, Upworthy would only display the final chosen package.

The `packages` dataset records the result of a series of tests. Each test includes various number of packages with the same `test_id` and for each package, the headline text `headline` and the `image_id` are recorded as well as the associated number of `impressions` and `clicks`. An example test with 4 packages is shown in Fig. 1.

## B. Hypothesis: Simplicity affects the click rate

Media creators always hope to attract more readers, but how to achieve it is usually unclear. With the Upworthy Research Archive dataset, we hope to dig into the pattern of reader's response and help the media creator to craft more attractive headlines for their news articles.

What could be characteristics of attractive headlines? Of course there are many possibilities. **One intuition is that, the readers usually come from a diverse audience and the headlines that attract most shouldn't be too complicated to understand. The simplicity (or, complexity) of the text will affect the response of the user.**

In the following report, we will dig into the dataset to find out what the relationship between the simplicity and the popularity.

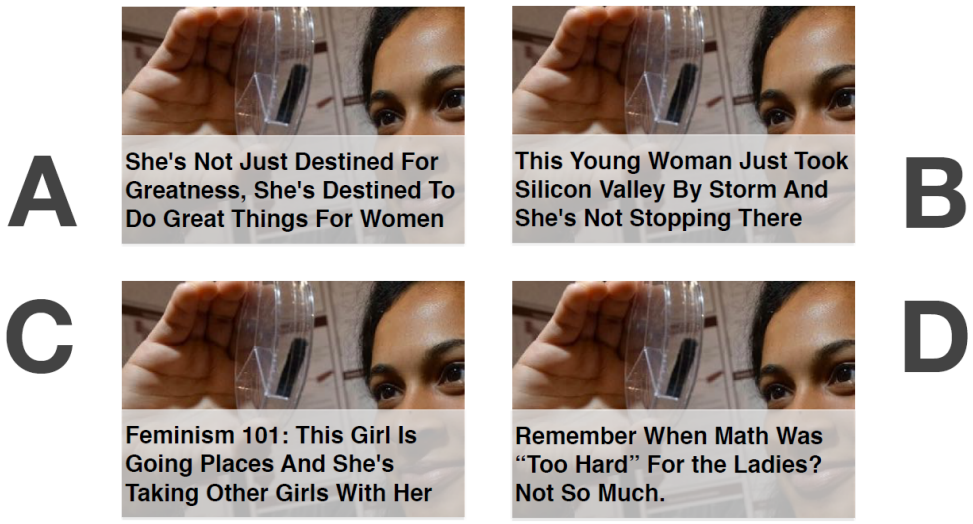


FIG. 1. An example of a test of 4 packages (A, B, C, D). All packages have the same `test_id` and are tested within the same week, but with potentially different images or headlines.

## II. DATASET STATISTICS

To get a feel of the dataset, we first perform exploratory data analysis on the `packages` dataset. Since we are mostly interested in studying the effect of headlines, we dropped the entries that do not have valid headline information. For all the remaining packages, as shown in Fig. 2, most tests contain 2 to 8 packages with an average number of impression of about 3600 per package.

We denote a test as a set of packages, for example, a test $t$ of $n$ packages is $t = \{p_1, p_2, ..., p_n\}$. To see if the packages in the same test are being assigned to similar number of readers, we define the *impression percentage difference* for a given test $t$ as

$$\texttt{max\_diff}_t = \frac{\max_{p\in t}(\texttt{impression}_p) - \min_{p\in t}(\texttt{impression}_p)}{\text{mean}_{p\in t}(\texttt{impression}_p)} \tag{1}$$

which is the ratio of the maximum difference in impressions between two packages within the same test divided by the average impression for the test. As shown in Fig. 3, for 99% of the tests, the impression percentage difference is below 25%, which means that most of the tests have relatively even impression distribution among packages.

We define *click rate* for a given package $p$ as

$$\texttt{click\_rate}_p = \frac{\texttt{clicks}_p}{\texttt{impressions}_p} \tag{2}$$

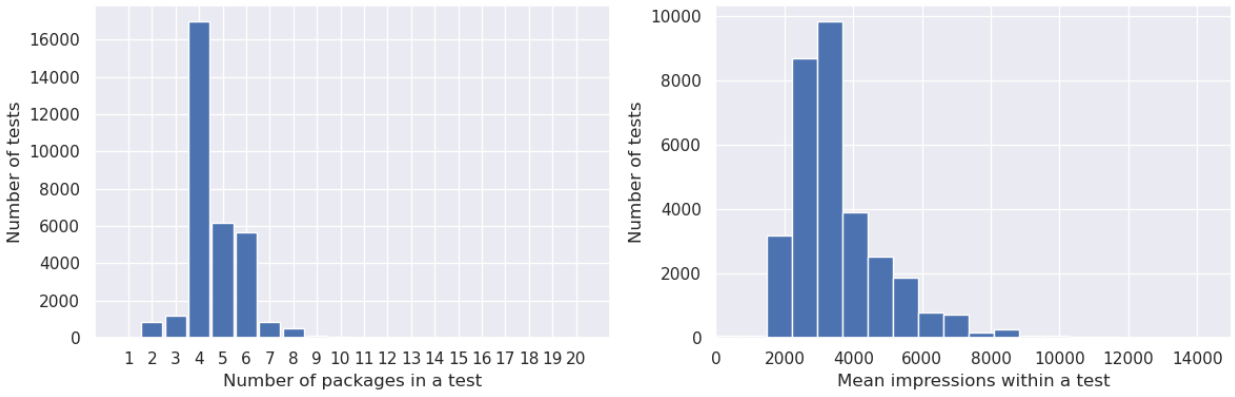


FIG. 2. Statistics of the tests. The left plot shows the number of tests with different number of containing packages. The right plot shows the number of tests with different average number of impression per package.

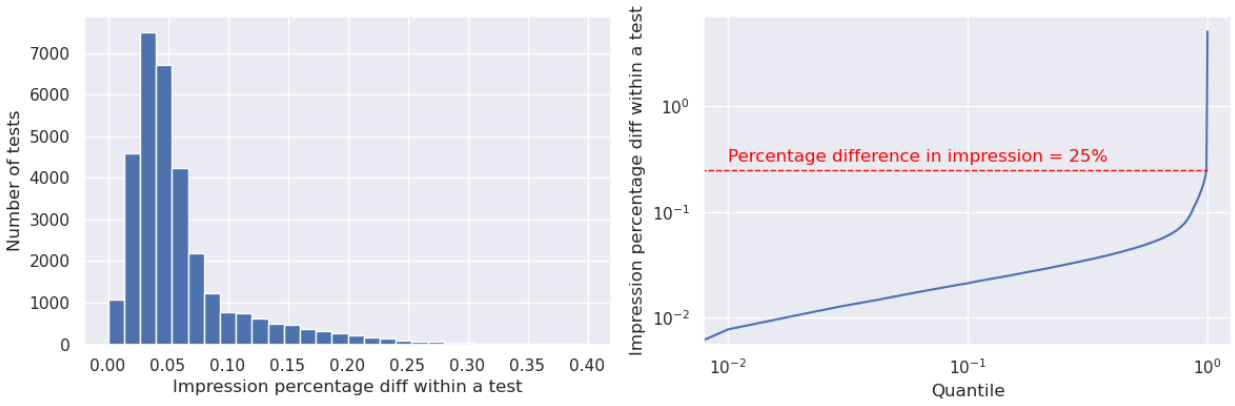


FIG. 3. Distribution of impression percentage difference. The left plot shows the distribution of impression percentage difference. The right plot is the quantile distribution.

The distribution of click rates among packages is shown in Fig. 4, which indicates that most of the packages have a click rate of less than 7%. The average click rate among all packages in the dataset is

$$\texttt{mean\_click\_rate} = \frac{\sum_{p \in \texttt{packages}} \texttt{clicks}_p}{\sum_{p \in \texttt{packages}} \texttt{impressions}_p} \tag{3}$$

which is around 1.52%.

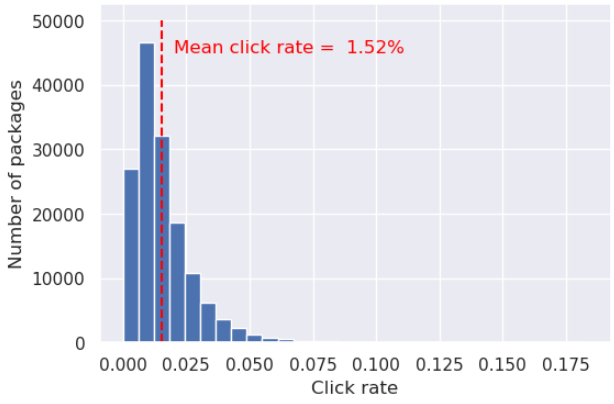


FIG. 4. Distribution of click rate among all the packages with headlines.

## III. MODEL OF CLICK RATE

### A. Simple model of reader's response

To see if the difference in click rate between two packages in a test is significant, we developed a simple model of reader's response that takes into account of the number of samples (impressions) collected and the number of clicks received among the two packages.

The response of each reader is binary and stochastic, so we can model it as a Bernoulli random variable $X$ such that $X = 1$ denotes that the user decides to click and $X = 0$ is not to click. To model the behavior of a group of readers exposed to the same package, we can make the following assumptions:

- Each reader make choice of whether to click independently
- The probability of clicking is the same for all readers

The first assumption is reasonable because the readers are typically independent of each other. The second assumption is not very realistic since each reader should have his/her own preferences over topics/content, but it is a good approximation if the number of readers assigned for the package is large.

For a package with $\texttt{impression}_k = N$, the number of received clicks $Y = X_1 + X_2 + ... + X_N$ out of $N$ readers is a binomial random variable with parameters $N$ and $p$, where $p$ is the probability of clicking on the package. In practice, since the click rate is usually small ($p \ll 1$), we can use a Poisson distribution to approximate the binomial distribution, therefore the variance of $Y$ is $Np$ which is the same as the expected value.

### B. Pairwise t-statistic of headline click rates

From the observed clicks $Y$, we can estimate the click rate of the package $p$ as $\hat{p} = Y/N$ and the variance of the estimate is $\hat{p}/N$. Given two packages 1 and 2 in the same test, the t-statistic of the difference in their click rates is then

$$\texttt{t\_stat}_{1,2} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_1/N_1 + \hat{p}_2/N_2}} \tag{4}$$

where $N_1$ and $N_2$ are the impressions of the two packages respectively, and $\hat{p}_1$ and $\hat{p}_2$ are the estimated click rate of the two packages. The t-statistic quantifies the statistical significance of the difference in click rates. With high absolute value of `t-stat`, we can say with strong confidence that the click rates are different.

Because the number of packages within a test could be more than two, we define the *pairwise t-statistic* as the t-statistic of the difference in click rate between two packages within the test. This way, a test of 4 packages with the same image but different headlines can generate $\frac{4\times3}{2} = 6$ pairs of headlines and corresponding pairwise t-statistic.

### C. Good headlines and bad headlines

We define a *good headline* is the headline in a pair of packages that results in higher click rate, and a *bad headline* is the one with lower click rates. To get a deeper understanding of what are the characteristics of good headlines and bad headlines, we compute the pairwise t-statistic of all pair of packages that are different *only* by their headlines. This way the difference in their click rate will only be due to the change in the headlines.

The total number of distinct pairs of headlines found is 140,621 across the `package` dataset with 63,209 unique headlines. Within each pair, the headline with higher click rate is identified as a good headline while the other is a bad headline. Since an actual test may contain more than two packages and a headline could be better than more than one headline in the same test, the same headline could appear many times as "good headline" and is also possible to appear as "bad headline" when compared to a stronger headline. However, within each pair, the "good headline" is always the one with higher click rate than the "bad headline". An example of 6 pairs of headlines with relatively high t-statistic is shown in Fig. 5.

| test_id | Good headline | Bad headline | t |
|---|---|---|---|
| 53a9a4f6f98fb2ebad0000b2 | he stopped her at a train station and told her she was beautiful. that's when the nightmare began. | being a 'slave to love' sounds romantic. but this survivor's poem reminds us what that really means. | 9.91 |
| 52323aaacb03337988000ec8 | here's a debate between an atheist and a christian. you'll be surprised at how it plays out. | a debate about religion that doesn't end in fisticuffs. seriously. | -9.32 |
| 51c8cece234062bbdd003366 | "someone gave a guy a megaphone, and you'll be awed by what he did with it" | spoken word schooled: the harsh truth of politics that you didn't learn in class | -8.90 |
| 528126ede296c2e8eb009687 | this man on the verge of breaking down says more in 3 minutes than most say in a lifetime | this man lives right in the path of typhoon haiyan and was pleading for help a year before it hit | 8.62 |
| 5149bcc7011bb50002004a7a | watch this bride dump her very important groom at the altar | "with this ring, mcdonald's, i thee wed" | -9.20086 |
| 52392441411723a48000b624 | “watch this family sweetly, and lovingly totally destroy one of the worst stereotypes of christians” | a message from a christian family that i can totally get behind | -8.84872 |

FIG. 5. Example of 6 pairs of headlines that result in different click rates. The t-statistic is evaluated with the order that the pair appear in the dataset, which could result in negative value if the package with good headline appears later.

We can also look at the distribution of the pairwise t-statistic among all the 140,621 identified pairs of headlines. As shown in Fig. 6, the distribution is broader than a standard normal distribution with a standard deviation of $\sigma = 1.967$, indicating that there are factors other than statistical fluctuation that contribute to the difference in click rates such that there are more pairs with high t-statistic than expected by assuming a normal distribution.

## IV. ANALYSIS

Our question is: Are the *good* headlines typically *simpler* than the bad headlines? To answer this question, our approach is to identify and collect good headlines and bad headlines, and then analyze the richness and variety of the vocabulary.

### *1. Collect good headlines and bad headlines with a set threshold*

With the pairwise t-statistic, we can collect a list of good headlines and a list of bad headlines and analyze the richness and variety of the vocabulary used in the two lists.

To do that, we define a threshold parameter `threshold` $\geq 0$, such that only when the

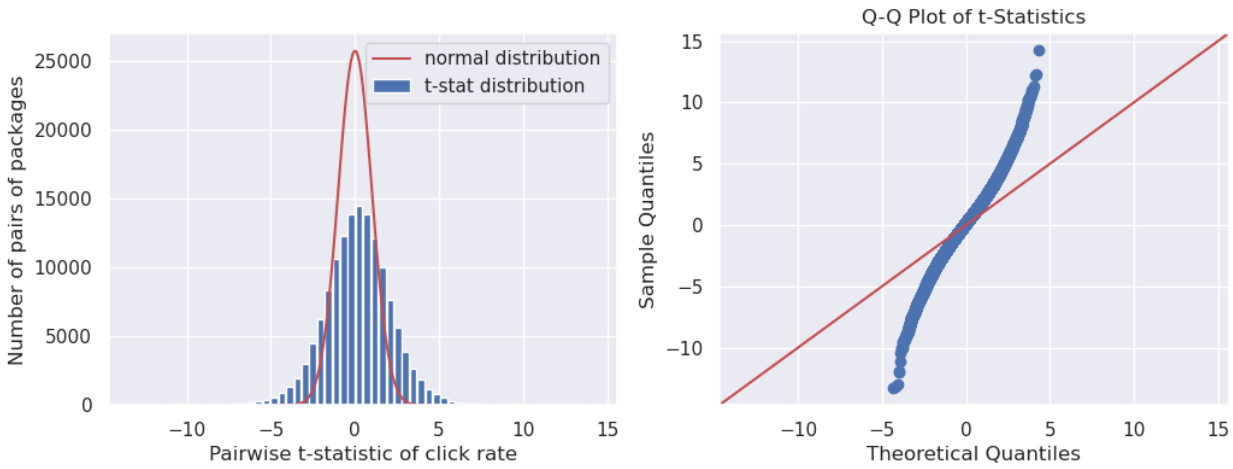


FIG. 6. The distribution of pairwise t-statistic of the click rates across 140,621 pairs of packages that are different only by their headlines. Left plot shows the histogram of the distribution. Right plot shows the Q-Q plot of the distribution versus a standard normal distribution.

absolute value of the pairwise t-statistic is greater than the set threshold, we consider the good headline *better* than the bad headline and put those two to the corresponding list. The choice of threshold is a trade-off between the data size and the cleanness. A higher threshold will result in a smaller data size for analysis because most of the headline pairs are filtered out, but can ensure that the good headlines are better than bad headlines with higher confidence level. A lower threshold will preserve more headlines in the collected lists, but since it basically puts a headline to good headline list even if the higher click rate it receives is not statistically significant, it could potentially mix good and bad in the analysis. As shown in Fig. 7, the number of headlines pairs decreases as the threshold increases. When threshold is set to 5, the filtered pairs contain only about 2% of the 140,621 pairs found in the original dataset.

If we consider each pair of headlines individually when constructing the lists, the same headline may appear multiple times in one list or even appear in both two lists, depending on the threshold value and its pairwise t-statistic with other headlines in the same test. Although such repetition could be meaningful, we also consider constructing a *unique* set of good headlines and bad headlines from the threshold-filtered list for analysis.

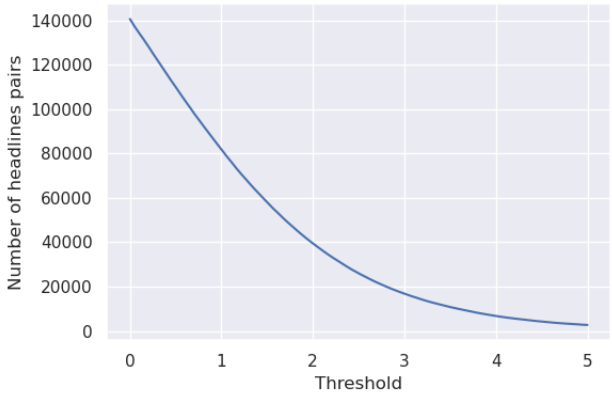


FIG. 7. The number of remaining headline pairs after applying a threshold filter on the absolute value of pairwise t-statistic.

### 2. *Lexical diversity: Text-Token Ratio*

To quantitatively study the difference in the language used by the good and bad headlines, we can look into the *Text-Token Ratio* (TTR) [2]. TTR is a linguistic metric used to quantify the lexical diversity of a text or corpus which is calculated as the ratio of the total number of unique words (*types*) to the total number of words (*tokens*) in the given text, as defined below

$$\texttt{TTR} = \frac{\text{Number of types}}{\text{Number of tokens}} \tag{5}$$

TTR usually goes down as the number of tokens increases, which is because as more types are already included in the vocabulary, there is smaller chance to see a new type when including more tokens. The corpus with less diversified vocabulary will typically have a lower TTR with the same token length, which means the number of unique words for a given number of words is lower. A highly technical corpus is expected to have high TTR because lots of its words have low frequency. As a reference, for English, it is usually sufficient to cover $> 90\%$ of the content with 2000 to 3000 types [3].

We analyzed the TTR of the 4 headlines lists: `good`, `bad`, `good_unique` and `bad_unique`. The lists are constructed with varying threshold parameter to filter the pairs and the unique headlines sets are constructed by removing repeated entries. The lower-cased headlines are also pre-processed to remove punctuation and non-alphabetical words. Fig. 8 shows the TTR vs. number of tokens curves of the 4 groups and all of them show decreasing trend as token number increases. After the number of tokens reaches around 10k, the TTR starts to

diverge between the 4 groups. Since the TTR at low token number can fluctuate a lot, we look at the TTR at 20k tokens and find that the good (unique) headlines has a lower TTR than bad (unique) headlines with all threshold settings. When the threshold is higher, the difference in TTR becomes larger and the TTR start to diverge at smaller token number, potentially due to the more distinct "good" and "bad" headlines included in the analysis.

To make sure the difference we see is statistically significant, we repeat the TTR measurement at 20k token number for 10 times by drawing random 20k tokens from the 4 groups. As shown in Fig. 9, the TTR in the good headlines is consistently lower than in the bad headlines with or without the repeated headlines removed. We can compute the t-statistic of the difference in TTR as

$$t_{\mathrm{TTR}} = \frac{\mathrm{mean}(\mathrm{TTR}_{bad}) - \mathrm{mean}(\mathrm{TTR}_{good})}{\sqrt{\mathrm{var}(\mathrm{TTR}_{good}) + \mathrm{var}(\mathrm{TTR}_{bad})}} \tag{6}$$

where the mean and var are taken among the 10 random draws. Between `good` and `bad` headline lists, for all the threshold settings, the t-statistic is greater than 4.2 which indicates statistical significance of the difference in TTR. For the `good_unique` and `bad_unique` lists, the t-statistic is 1.05 when threshold is set at 0, which is not significant ($p > 0.05$). However, with higher threshold ($\geq 2$), the t-statitic is larger than 2.3 which indicates the significance in the difference in TTR.

When threshold is set to 2, for every 20k tokens (words), the `bad` headlines include on average 3709 types, while the `good` headlines include 3347 types, which is about 10% lower. If repeated headlines removed from the list, the `bad_unique` headlines include 3745 types, while the `good_unique` headlines include 3510 types, which is still about 6% lower. Since the difference is statistically significant, we can conclude that the better headlines are using less diversified words.

*3. The effect of less frequent words in a headline*

We can also study the effect of less frequent words in a headline. From the previous analysis, we see that the good headlines typically use a smaller vocabulary, then a natural expectation is that, if a headline contains very infrequent words, it is less likely to be a good headline.

To study the effect of less frequent words in a headline on the click rate, we first build a

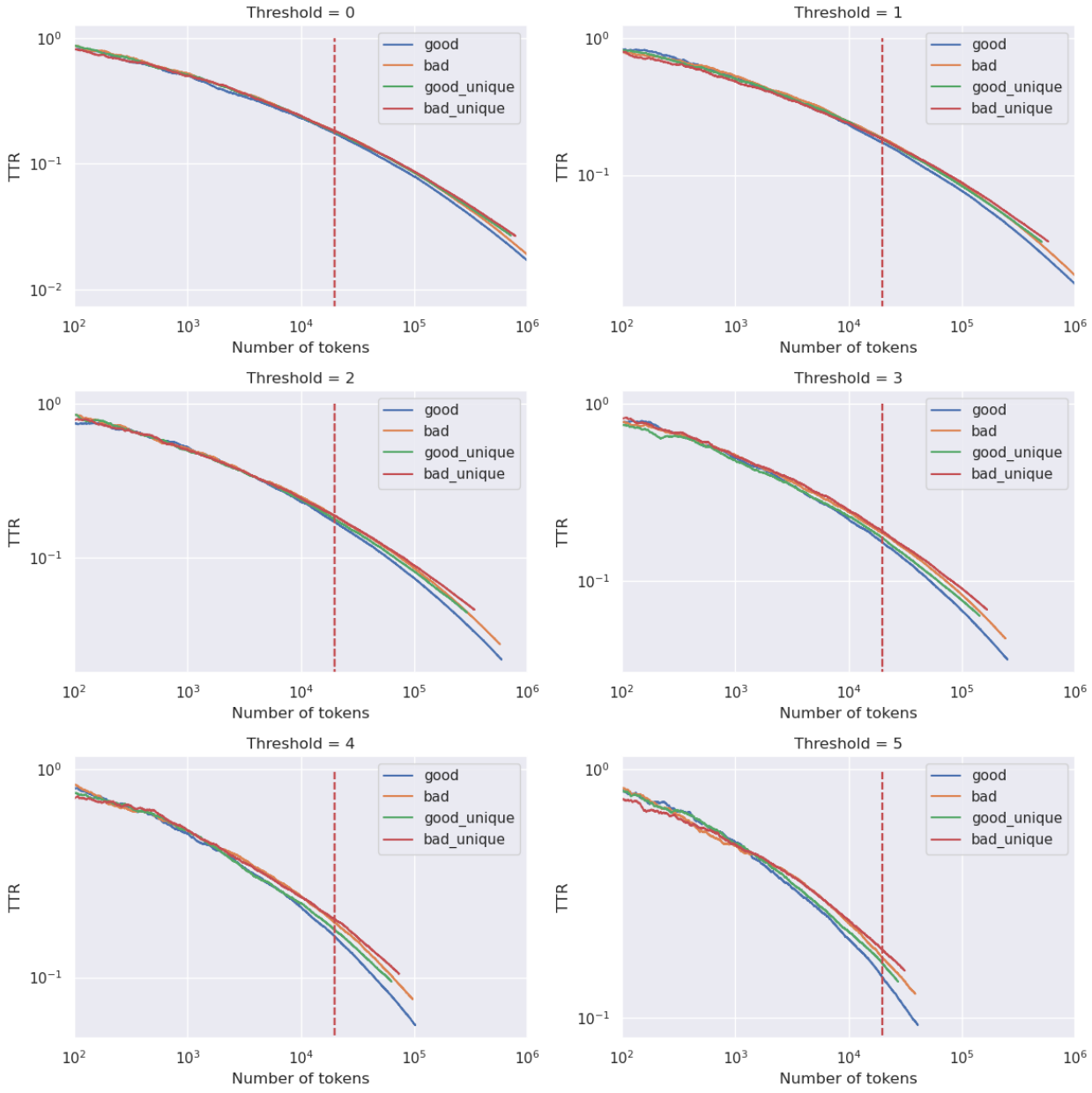


FIG. 8. The Type-Token Ratio as a function of number of tokens between different groups of texts. Different plots are with different threshold setting for constructing good/bad headline lists. The "good_unique" and "bad_unique" are the unique headlines from the good/bad headline lists. As the threshold increases, the number of tokens included in the headline lists decreases, so the maximum number of tokens is different between the plots. As a reference, the red dash line marked the position of 20,000 tokens.

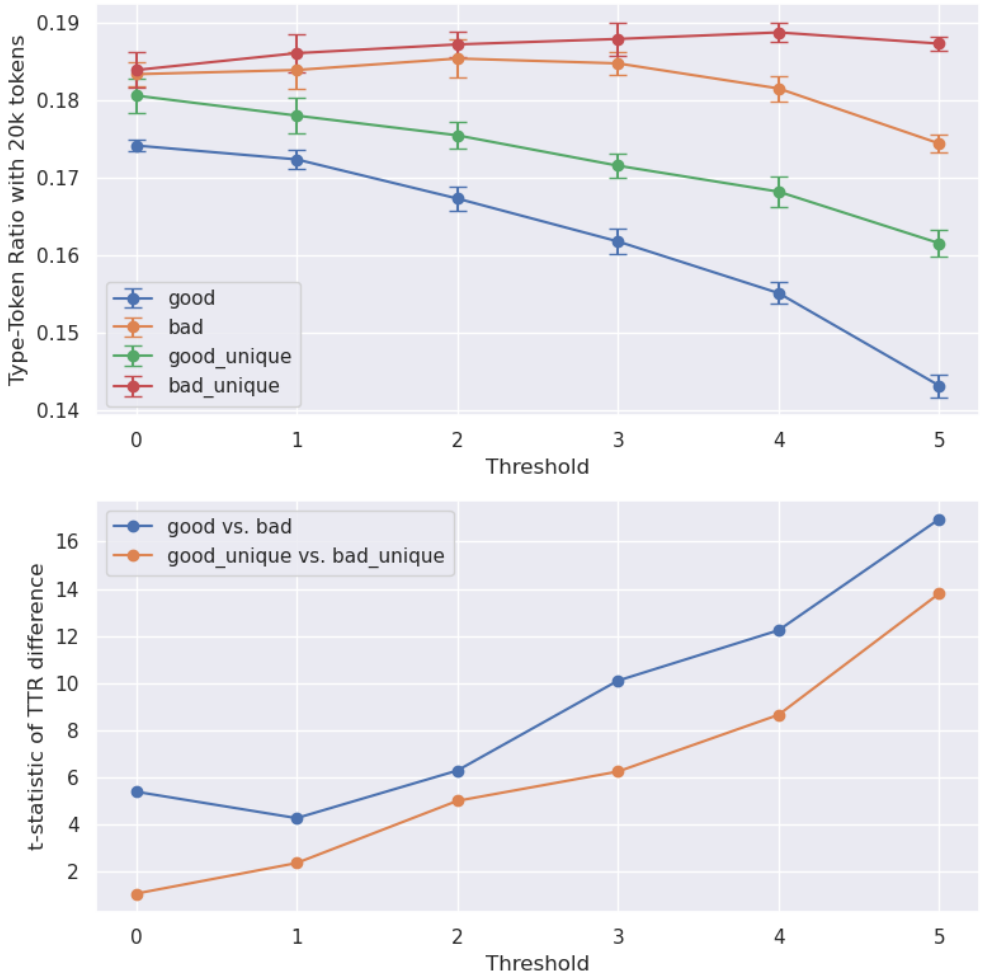


FIG. 9. The Type-Token Ratio with 20k tokens for different groups of headlines. The upper plot shows the measured TTR with different threshold settings, where the value is obtained by averaging 10 repetitions, and the error bar denotes one standard deviation among the 10 measurements. The lower plot shows the t-statistic of the difference in TTR between `good` vs. `bad` and `good_unique` vs. `bad_unique`.

word frequency table from all the unique headlines in the dataset. The word frequency of a certain word $w$ is then defined as

$$\texttt{freq}(w) = \frac{\text{Number of times the word } w \text{ appear in all headlines}}{\text{Total number of words in all headlines}} \tag{7}$$

An example of a few most frequent words and least frequent words is shown as a word-cloud in Fig. 10 and Fig. 11. The most frequent words are dominated by the pronouns and prepositions, and the least frequent words are usually more specific in meanings and sometimes include words that are misspelled or very confusing to people. The most frequent

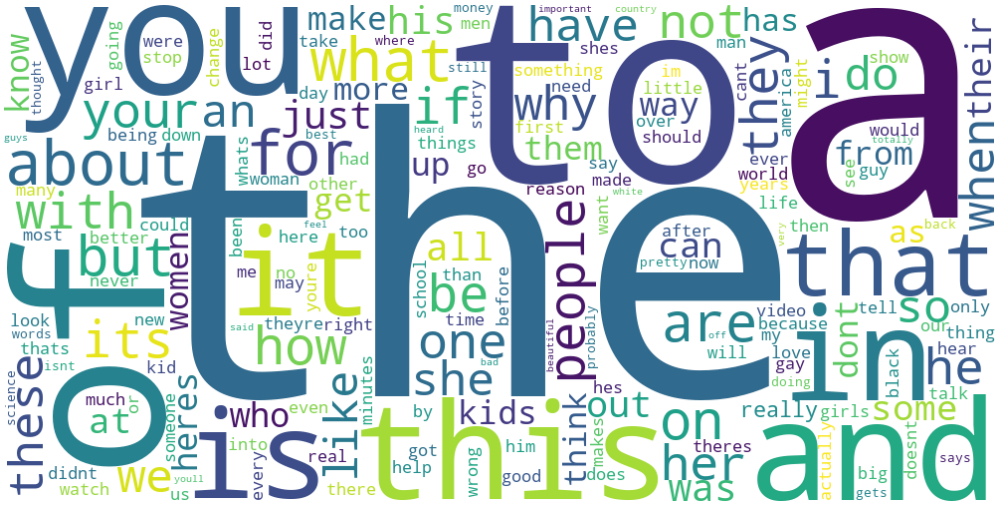


FIG. 10. The most frequent words in the headlines. The font size is ranked by the word frequency.


FIG. 11. The least frequent words in the headlines. the font size is ranked by the inverse of word frequency.

3000 words among the headlines constitute over 90% of the total word counts.

We define a headline "infrequent" score as the minimum value of log of the frequency of the words occurred in the text

$$\texttt{score}(headline) = \min_{w \in headline} \{\log(\texttt{freq}(w))\} \tag{8}$$

This score is related to the frequency of least frequent word in a headline, so if the headline includes a very infrequent word, its score will be lower. For every pair of headlines, we can

define a score difference as the difference in the score between the two headlines

$$\texttt{score\_diff}(A, B) = \texttt{score}(A) - \texttt{score}(B) \tag{9}$$

The distribution of headline scores among the good and bad headline is shown in Fig. 12, where the threshold is set to 0. The histogram shows a slight trend that the bad headlines are more likely to have lower score.

To study if the difference in the "infrequent" score explains the difference in click rates, we plot the scatter plots of the score difference and pairwise t-statistic in Fig. 12 and perform a least square fit $Y \sim X$ where $X$ is the score difference, and $Y$ is the pairwise t-statistic in click rates. The fit has an slope of 0.0954 and an $R^2 = 0.01$, which means that about 1% of the variation in the t-statistic can be explained by the score difference. Although the $R^2$ is small, the positive slope is significant above 0 with a t-statistic of 38.6, indicating a positive correlation with high confidence.

## V. CONCLUSION

From the Upworthy Research Archive dataset, we collected 140,621 pairs of packages that are different only by their headlines. By analyzing the good headlines (with higher click rates) and bad headlines, we find that the good headlines are composed with a less diversified vocabulary, which is reflected on their relatively lower Type-Token Ratio. We also find that the variation in the click rate difference can be partially explained by the "infrequent" score of a headline, which further strengthens the claim that the appearance of a less frequent word negatively affects the performance of a headline. Such finding indicates that *simplicity* is a common feature of good headlines.

In the future, more study could be carried on to find more features of good headlines.

[1] J. N. Matias, K. Munger, M. A. Le Quere, and C. Ebersole, The upworthy research archive, a time series of 32,487 experiments in us media, Scientific Data **8**, 195 (2021).

[2] D. Jurafsky and J. H. Martin, Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.

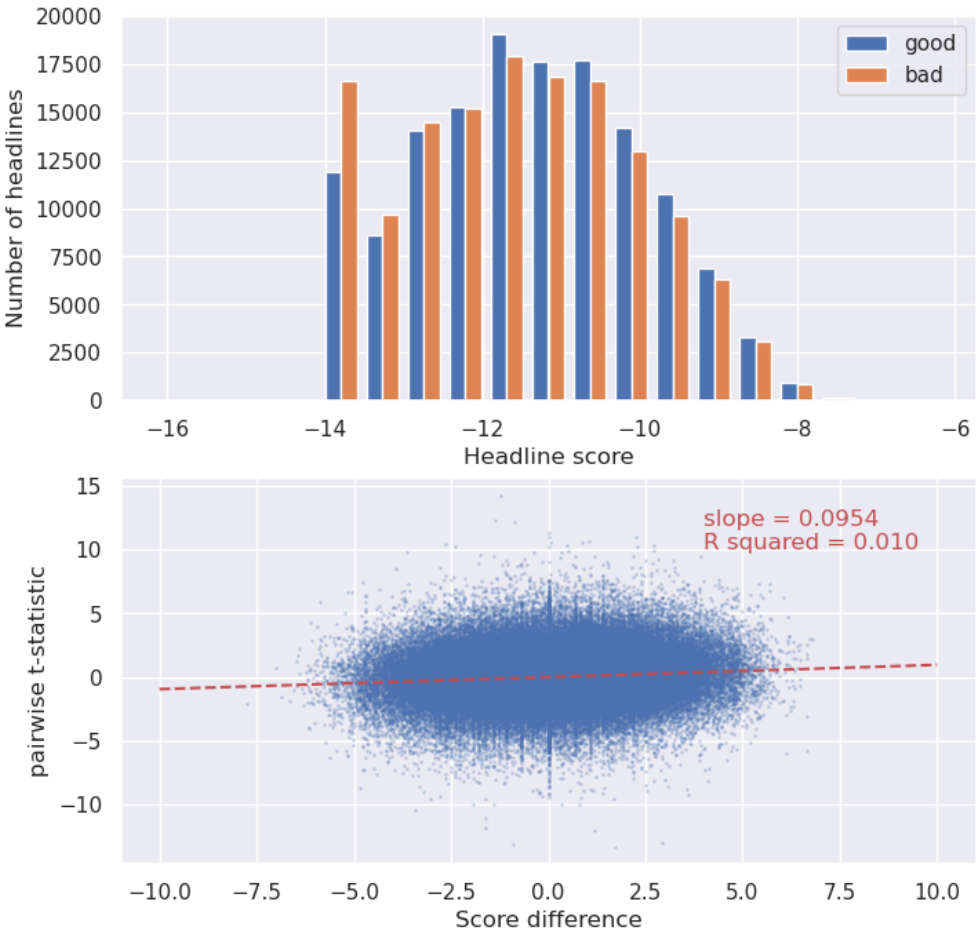


FIG. 12. The upper plots show the distribution of headline scores among `good` headlines and `bad` headlines if setting the threshold to be 0. The lower plot shows the scatter plot of score difference between a pair of headlines and their pairwise t-statistic of the click rates. The red dash line is a fit by linear regression with ordinary least square.

[3] W. E. Nagy and R. C. Anderson, How many words are there in printed school english?, Reading research quarterly , 304 (1984).