# Upworthy’s A/B Testing Preference For Clickbait Articles

By Shahnawaz Mogal, Baltazar Zuniga Ruiz, Sahil Rane, Karina Walker

## Abstract:

*Upworthy.com is a company that uses A/B testing for maximizing the clicks received by their published articles. Different packages consisting of a headline and image of the same article are randomly assigned to different users and based on the user response to the different packages, a certain package that is published to all users of Upworthy is chosen. This system can result in a long-term increase in clickbait articles being published as these articles are likely to maximize user clicks. This report employs a logistic regression model and an XLM-Roberta fine-tuned model to predict the ‘clickbaitiness’ of an article headline (a metric indicating whether an article is clickbait or not). The accuracy of the XLM-Roberta model was found to be greater (0.99625) than that of the logistic regression model (0.9697) and thus it was used as a model to find the ‘clickbaitiness’ of the packages published by Upworthy. We then ran a T-Test to understand whether the difference in ‘clickbaitiness’ of winners and non-winners is statistically significant. We found that the difference in the ‘clickbaitiness’ of winners and losers in the A/B tests was statistically significant, thus indicating that winners tend to have more clickbait headlines, thus resulting in more clickbait articles being published by Upworthy in the long term.*

## Background and Research Question:

Upworthy.com is a company that focuses on positive story-telling that aims to “change what the world pays attention to”. The data they use to improve user interaction is created by showing readers the same article but with a different package consisting of a headline and an image. Upworthy then determines the effectiveness of each package for a specific article through A/B testing — common within the media industry. A/B testing is the process of improving business metrics by identifying the version of a webpage or article that would have the highest impact on visitors. Different versions of a website are presented to two audiences and the webpage or article that has favorable outcomes for the business is selected to be the desired website. This testing method allows Upworthy to measure user behavior for different packages of a particular article. Based on the user responses, either more A/B tests can be conducted or a certain package can be finalized as a ‘winner’ and be published to all users.

Upworthy’s success has been attributed to the idea of ‘clickbait’ which despite improving business metrics has downsides. The ‘clickbait’ movement has led to sensational articles becoming mainstream, which often resulted in misleading headlines and a decrease in authenticity. As a result, there has been a growing movement for social media platforms to deliver more genuine content by altering their algorithms to reduce the visibility of clickbait.

For our study, we define a metric called ‘clickbaitiness’ (a number between 0 and 1) which describes the extent to which an article is clickbait or not. If the ‘clickbaitiness’ is 0 then the article is not clickbait and if it is 1 it is clickbait. This allows us to determine the varying extents of clickbait in the articles produced by Upworthy.

## Research Question:

Can we implement reliable models to quantify the ‘clickbaitiness’ of an article? Is there a statistically significant difference between the ‘clickbaitiness’ for winners and losers of the A/B tests conducted by Upworthy?

This question provides insight into whether A/B testing directs companies towards increased use of clickbait. If there is a statistically significant difference between the ‘clickbaitiness’ of winners and losers such that winners tend to have more clickbait, then we can conclude that A/B testing played a role in the increase of clickbait headlines published by Upworthy. If there isn’t a statistically significant difference between the ‘clickbaitiness’ of winners and losers we can conclude that Upworthy’s A/B testing was in fact an effective and useful way of improving user interaction without providing sensationalist and misleading headlines.

Moreover, by analyzing whether there is a statistically significant difference between the ‘clickbaitiness’ of winners and losers, we are able to get a glimpse into the underlying reasons behind the growing trend of sensational headlines and the eroding sense of trust towards media articles. If we find that A/B testing shows a preference for clickbait headlines, this helps expose flaws in the A/B testing methodology a publisher adopts.

## Executive Summary:

Our research process consisted of two parts. First, we considered the question: how can we define a metric for the ‘clickbaitiness’ of an article? Second, using this metric for clickbait, we explore whether there is a statistically significant difference between the ‘clickbaitiness’ of the winning article packages and the losing ones.

To answer the first question, we sourced an external dataset, Clickbait Dataset, composed of headlines of news articles and their classification into two categories: clickbait or not clickbait. The external dataset is an open-source dataset collated by Aman Anand shared via Kaggle. The clickbait headlines are from sites such as ‘BuzzFeed’, ‘Upworthy’, and ‘ViralStories’, and the non-clickbait headlines are collected from news sites such as ‘WikiNews’, ‘New York Times’, and ‘The Guardian’.

To come up with a metric for the ‘clickbaitiness’ of the article, we initially used a logistic regression machine learning model that took the headlines of the articles as input and gave the ‘clickbaitiness’ of the article as an output. This had an accuracy of 0.9697. Thus, our logistic regression model is highly accurate in determining whether an article is clickbait or not and thus gives us a good scale for ‘clickbaitiness’. However, we decided to attempt to improve the accuracy of our model further to get a more robust metric for ‘clickbaitiness’ using a fine-tuned version of the large pre-trained XLM-Roberta model to predict the ‘clickbaitiness’ of the article. This model had an accuracy of 0.99625 when tested on the Clickbait Kaggle Dataset.

Following this, we used the provided dataset from Upworthy consisting of 32,488 A/B tests and 150,000 packages with a median of 4 packages per test. We used the data in packages.csv, specifically the headline column which had the headline of the package of each article, and the winner column which documented whether a particular package was chosen as the winner or not through the A/B testing. Since the XLM-Roberta model was more accurate than the logistic regression model, it was used to calculate the 'clickbaitiness' of the Upworthy articles. After we found the 'clickbaitiness' of the Upworthy articles, we ran a T-Test to determine whether there were more winning packages that were clickbaity than losing packages and whether this difference was statistically significant. We found the difference in 'clickbaitiness' of winning packages and losing packages is statistically significant. This means that there are more clickbait articles in winning packages, which indicates that Upworthy was more likely to publish articles that had clickbait in it as a result of the A/B tests.

## Technical Exposition:

### 1) A logistic regression model with multiple runs to test for clickbait

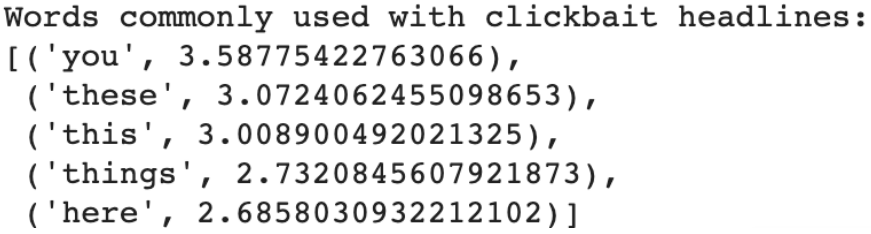
We start by applying logistic regression to the headlines of the Kaggle Clickbait dataset and their respective clickbait value (0 or 1). The process of applying logistic regression to the headlines and 'clickbaitiness' begins by converting every headline into different features — the most frequent features are extracted. To select these features, we experimented with combinations of word bi-grams, character bi-grams, word tri-grams, or character tri-grams. The formation of these features allowed us to vectorize the headlines and as such, use them as the input into the linear regression model. An 80-20 split of the data allows us to train our data with 80% of the available data and test on the remaining 20% of data that the model had not yet seen. The vectorized training headlines are the $X$ values and their respective clickbait values are the $Y$ values used to train the logistic regression model. Once trained, we predict the 'clickbaitiness' of our testing data by passing in the vectorized headlines and comparing the results of our models with the actual clickbait value given to us by the dataset.

Using this logistic regression model, we found that words that tend to be more specific, such as names of people, tend to predict lower 'clickbaitiness' values. On the other hand, words that are more general, such as 'you' or 'these' tend to predict higher 'clickbaitiness' values. This observation agrees with the idea that a clickbait headline is likely to be more sensational and appeal to a larger audience through the use of vocabulary that is broad and general. Correspondingly, the use of more specialized and specific vocabulary in headlines is likely to appeal to a smaller audience that is more well acquainted with the topic.

*Table 1: Classification Report for Different Feature Combinations:*

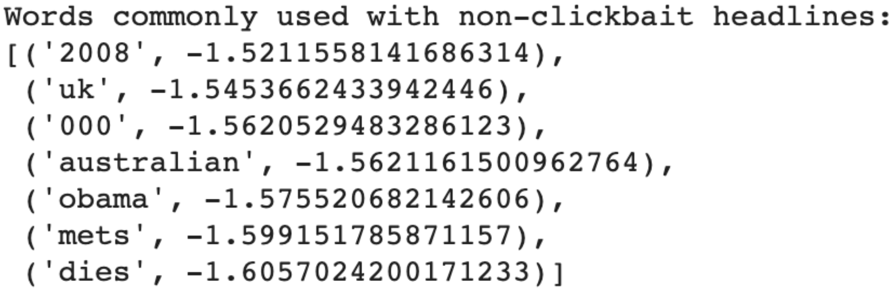
| *Features used* | *Accuracy* | *F1* | *Precision* | *Recall* |
|---|---|---|---|---|
| 100k word bi-grams | 0.969 | 0.968 | 0.957 | 0.979 |
| **300k word tri-grams** | **0.970** | **0.969** | **0.958** | **0.9799** |
| 100k character bi-grams | 0.938 | 0.936 | 0.930 | 0.9415 |
| 300k character tri-grams | 0.965 | 0.964 | 0.961 | 0.967 |

*Figure 1: Largest positive regression coefficients for words*



```
Words commonly used with clickbait headlines:
[('you', 3.58775422763066),
 ('these', 3.0724062455098653),
 ('this', 3.008900492021325),
 ('things', 2.7320845607921873),
 ('here', 2.6858030932212102)]
```

*Figure 2: Most negative regression coefficients for words*



```
Words commonly used with non-clickbait headlines:
[('2008', -1.5211558141686314),
 ('uk', -1.5453662433942446),
 ('000', -1.5620529483286123),
 ('australian', -1.5621161500962764),
 ('obama', -1.575520682142606),
 ('mets', -1.599151785871157),
 ('dies', -1.6057024200171233)]
```

The numbers assigned to each word represent the regression coefficient. The higher the coefficient value is, the more it correlates with ‘clickbaitiness’ and vice versa. As the images above show, the words that are more specific, such as ‘australian’, ‘uk’, or ‘2008’ have negative regression coefficients. This means that the inclusion of those words in headlines signifies that the headlines are not clickbait. On the other hand, words that are vaguer, such as ‘these’, ‘this’, and ‘things’, have positive regression coefficients, meaning that the inclusion of these words in headlines signifies higher ‘clickbaitiness’. The word ‘you’ also has a high regression coefficient as it is often used in a sensationalist manner and aimed to appeal to the viewer’s own sense of self-importance.

This makes sense since words such as “this” and “you” commonly occur in clickbait titles such as “Kim Jong Un Would Really Hate For **You** To Watch **This** Which Is Exactly Why **You** Should”.

**Understanding the classification report for our logistic regression:**
The accuracy of the logistic regression model is `0.9697(4 d.p.)`. Since accuracy is a measure of all currently identified instances, this suggests that the model was very good in identifying clickbait or not clickbait articles. Since the data we used to train our model contains 50% of clickbait and 50% of not clickbait, accuracy is a good measure for the model.

The size of our testing data for our logistic regression model is 6400. The precision of our model was `0.9576(4 d.p.)`. This informs us that our model correctly predicted whether the headline was clickbait or not 96% of the time. High precision relates to a low false-positive rate.

The recall value for our model was `0.9796(4 d.p.)`. This suggests that our model found 98% of clickbait headlines. A high recall indicates the model labeled a high proportion of clickbait headlines.

The f1 score is the weighted average of precision and recall where the best score is 1 and the worst is 0. The f1 score for our model is `0.9685 (4 d.p.)`. As the f1 score accounts for both false positives and false negatives, this suggests that 97% of our positive predictions were correct.

## 2) Fine-tuning the XLM-RoBERTa model

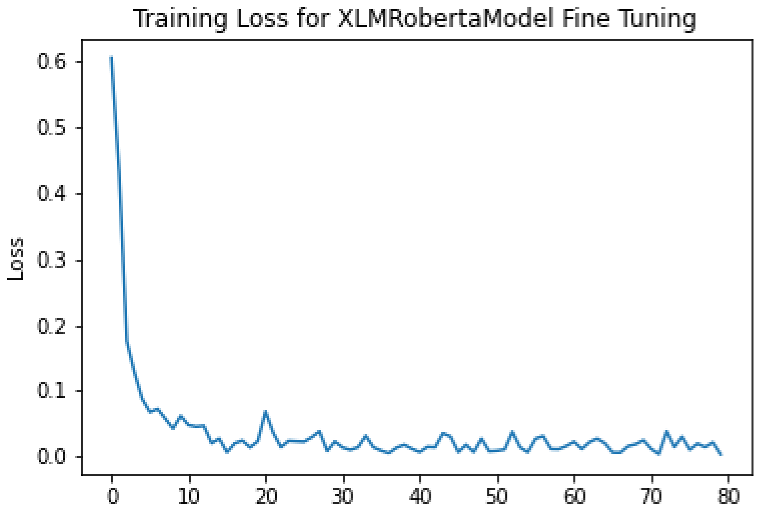
Even though our logistic regression model performed well, we thought we could make use of large pre-trained models in order to better distinguish between clickbait and non-clickbait headlines. The drawback of using large pre-trained models is the non-interpretability of the results. With logistic regression, we were able to figure out the correlation coefficients of each one of the features. However, with a large pre-trained model, we are unable to do so. That said, these large models have been trained on millions of parameters and a large amount of data which makes them more comprehensive.

We decided to use XLMRoberta, a large multi-lingual language model trained on 2.5TB of filtered CommonCrawl data, and fine-tune the model for our specific use-case: predicting the 'clickbaitiness' of headlines. We used the XLMRoberta model because the model performs almost as well as the other language models. Moreover, for future research, the model can also be extended to detect clickbait headlines in other languages.

We experimented with fine-tuning the model with different number network nodes, training epochs, and training label ratio, etc.

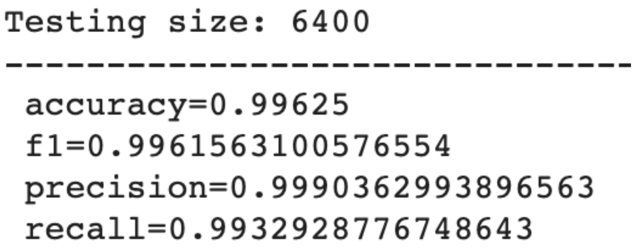
We found that fine-tuning using 32 nodes for 1 epoch on an even split of positive and negative samples gave us the best results.

*Figure 3: Training Loss for the XLM-Roberta model fine-tuning*



The model performed extremely well to predict classes for 6400 test headlines with the following results:

*Figure 4: Accuracy of the fine-tuned XLM-Roberta Model on test headlines*



```
Testing size: 6400
--------------------------------
 accuracy=0.99625
 f1=0.9961563100576554
 precision=0.9990362993896563
 recall=0.9932928776748643
```

In future work, we would like to use the perturbations of the headlines in order to better understand the features that this model is picking up.

### 3) Applying fine-tuned XLMRoberta to the original dataset

We then used the fine-tuned XLMRoberta Model to obtain ‘clickbaitiness’ scores for the original set of headlines used in the A/B testing process conducted by Upworthy.
We found that all headlines in the A/B testing that were labeled ‘winner’ had a mean ‘clickbaitiness’ of 93.54% and all headlines that were not labeled as ‘winners’ had a mean ‘clickbaitiness’ of 92.74%.

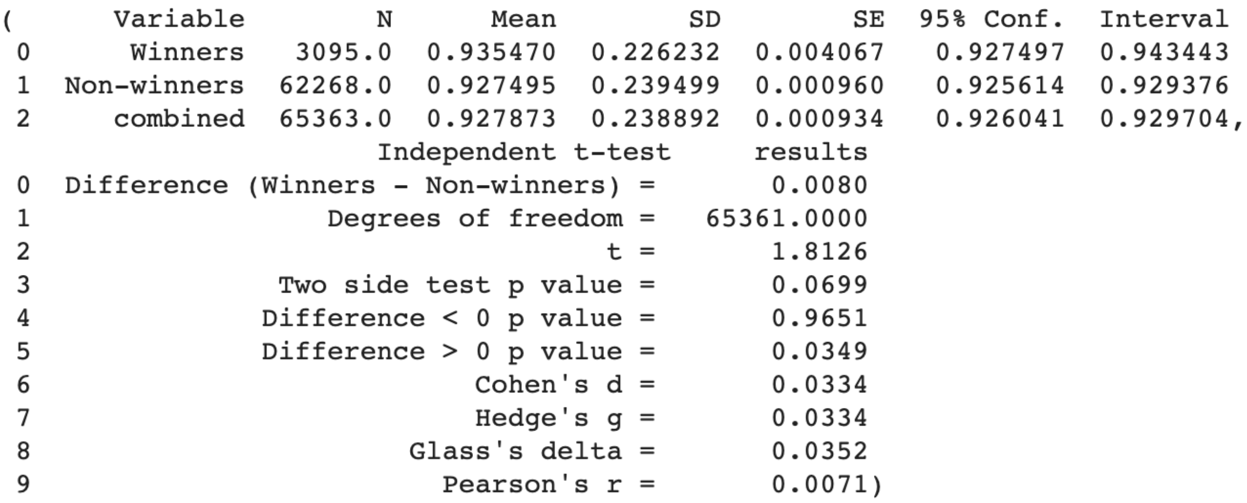
**One-tailed Two sample T-Test:**

We use the winning headlines as our sample and compare the ‘clickbaitiness’ mean to our population’s ‘clickbaitiness’ mean.

Let µ1 = the true mean ‘clickbaitiness’ of the winning headlines
Let µ2 = the true mean ‘clickbaitiness’ of the losing headlines

Our null hypothesis states that there is no difference between the true mean ‘clickbaitiness’ of our winning headlines and our losing headlines. Our alternative hypothesis states that the true mean ‘clickbaitiness’ of our winning headlines is greater than the true mean ‘clickbaitiness’ of the losing headlines.
Null Hypothesis: µ1 - µ2 = 0
Alternative Hypothesis: µ1 - µ2 > 0
Significance level: 0.05

*Figure 5: T-Test Results*



```
(      Variable        N      Mean        SD        SE  95% Conf.  Interval
 0      Winners   3095.0  0.935470  0.226232  0.004067   0.927497  0.943443
 1  Non-winners  62268.0  0.927495  0.239499  0.000960   0.925614  0.929376
 2     combined  65363.0  0.927873  0.238892  0.000934   0.926041  0.929704,
                       Independent t-test     results
 0  Difference (Winners - Non-winners) =       0.0080
 1                  Degrees of freedom =   65361.0000
 2                                   t =       1.8126
 3               Two side test p value =       0.0699
 4              Difference < 0 p value =       0.9651
 5              Difference > 0 p value =       0.0349
 6                           Cohen's d =       0.0334
 7                           Hedge's g =       0.0334
 8                       Glass's delta =       0.0352
 9                         Pearson's r =       0.0071)
```

We use the “Difference > 0 p-value” as our p-value since that is the one that corresponds with our alternative hypothesis.

The Conclusion from T-Test:
Since the p-value (0.0334) is less than our significance level (0.05), we reject the null hypothesis that there is no difference between the true mean ‘clickbaitiness’ between winning and losing headlines. We have convincing evidence that the true mean for the ‘clickbaitiness’ of winning articles is greater than the true mean for the “clickbaitiness” of losing articles.

**Conclusions:**

Through our logistic regression model, we were able to identify whether an article is clickbait or not with an accuracy of 0.9697. Using this model, we found that more specific words tend to lead to lower ‘clickbaitiness’ values whereas words that are vaguer tend to lead to higher ‘clickbaitiness’ values. Through the XLM-Roberta model, we got an accuracy of 0.99625, thus this model was more accurate than the logistic regression model and was used to calculate the ‘clickbaitiness’ values for the Upworthy article packages.

To answer the second part of our research question, we perform a one-tailed two-sample T-Test between the true mean ‘clickbaitiness’ of the winning and losing headlines. Through the t-testing, we obtained a p-value of 0.0334. Since this p-value is less than 0.05, our significance level, we have strong evidence that the winning headlines have a higher true mean ‘clickbaitiness’ than the losing headlines.

Because winning headlines have statistically higher ‘clickbaitiness’ when compared to losing headlines, we can conclude that Upworthy’s A/B testing has increased the frequency of clickbait packages being published by them.

Because A/B testing is a widely used method to compare headlines, it is likely that other companies have performed similar data analysis and reached a similar conclusion; the more “clickbaity” a headline is, the more likely it is to win an A/B test and thus gain higher online user attention. Therefore, the use of data analysis on these A/B tests is likely responsible to some degree for the higher prevalence of “clickbait” headlines in the news.

**References:**

Aman Anand. (2019). *Clickbait dataset*. Kaggle: Your Machine Learning and Data Science Community.

https://www.kaggle.com/amananandrai/clickbait-dataset