Investigating The Use of Pronouns And Punctuation Marks In Headlines

Team 13

November 2021

1 Non-Technical Executive Summary

Our findings suggest that pronouns and punctuation, in particular, the use of question marks, has some impact on a headline's click rate. We see overall that they have a negative impact on the probability of a reader, given that they have read the headline, will click on the article.

2 Technical Exposition

Firstly, in order to evaluate the impact of different variables on the viewers, we will standardise the number of clicks using the number of impressions on each article. So, throughout this report we will consider the ratio, r:

$$r = \frac{\text{clicks}}{\text{impressions}}$$

This will be useful in ensuring that conclusions are not skewed by natural growth of the websites over time, or annual/monthly trends in page impressions.

One of the variables we considered was the use of punctuation in headlines. In particular, the set we considered was:

$$\{"?", "!", "...", " - "\}$$

.

We quantified whether each headline contained any of these punctuation marks be adding a column to the data frame, containing a value of 1 if any of the punctuation marks listed above were included in the headline, and 0 otherwise.

Since we were not interested in experiments where only the images were varied, we removed rows of the data frame which had repeated headlines. This will be important later on when conducting tests, since such a small proportion of the data is assigned a value of 1. Before removing the repeated headlines, around 24.7% of the headlines contained any of the punctuation marks, whereas afterwards, this rose slightly to 27.5%.

Next, we can visualise this data is various ways. The boxplots shown in Figure 1 (without outliers) show that the set of headlines containing punctuation, and not containing punctuation are distributed very similarly.

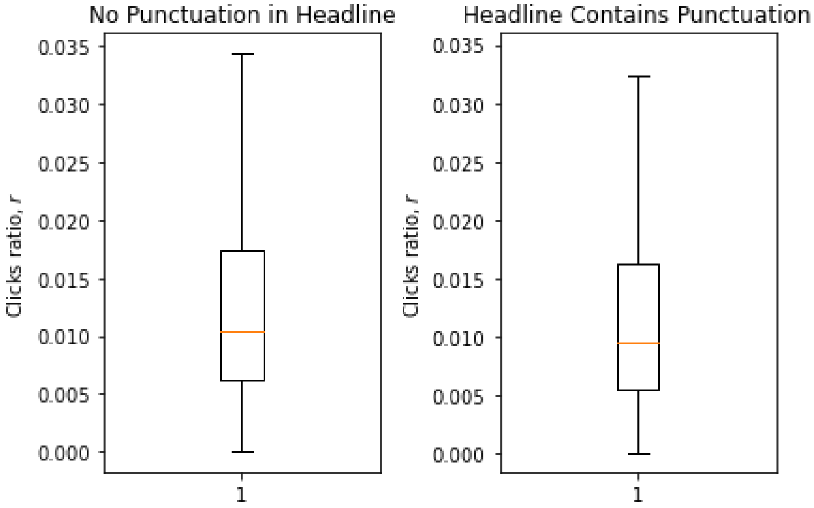

Figure 1: Boxplots of headlines not containing the specified punctuation marks (left) and headlines containing the specified punctuation marks (right). Outliers have been removed from these diagrams.

The histograms shown in Figure 2 also highlight this relationship, but we can see that the mean of the data containing punctuation marks, μ_p, is slightly lower than of mean of those without punctuation marks, μ_{np}. To be exact:

$$\mu_p = 0.0127 < 0.0137 = \mu_{np}$$

We may hypothesise that the use of punctuation marks is off-putting to a reader, and that given that someone has read a headline, the probability of them clicking on it is lower if it contains the specified punctuation.

One way to investigate this further is to consider a logistic regression model. The scatter plot shown in Figure 3 also suggests that the click ratios, r of data points with punctuation marks are more concentrated closer to 0.

By applying logistic regression onto a training subset of the data, we obtain constants of $\beta_0 = -4.50419434$ and $\beta_1 = 0.0594598$. So our logistic regression line is:

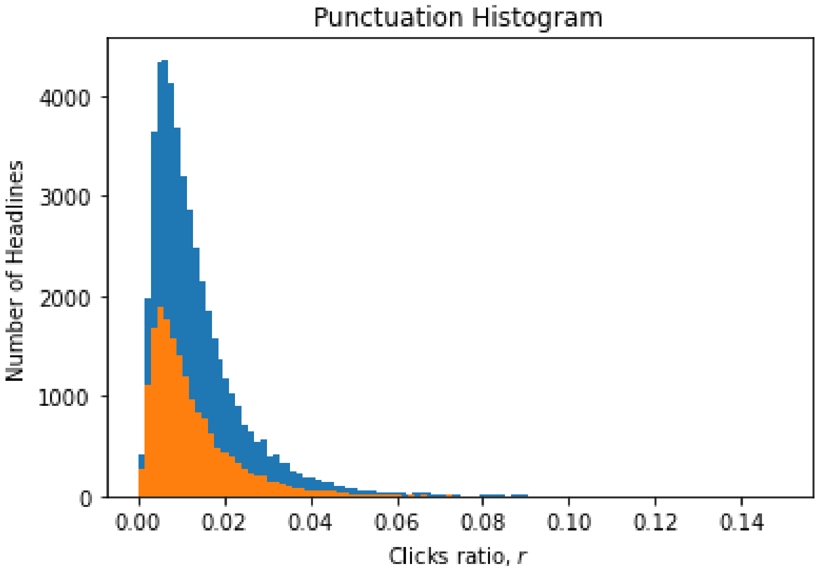

Figure 2: Histograms of headlines not containing the specified punctuation marks (orange) and headlines containing the specified punctuation marks (blue).

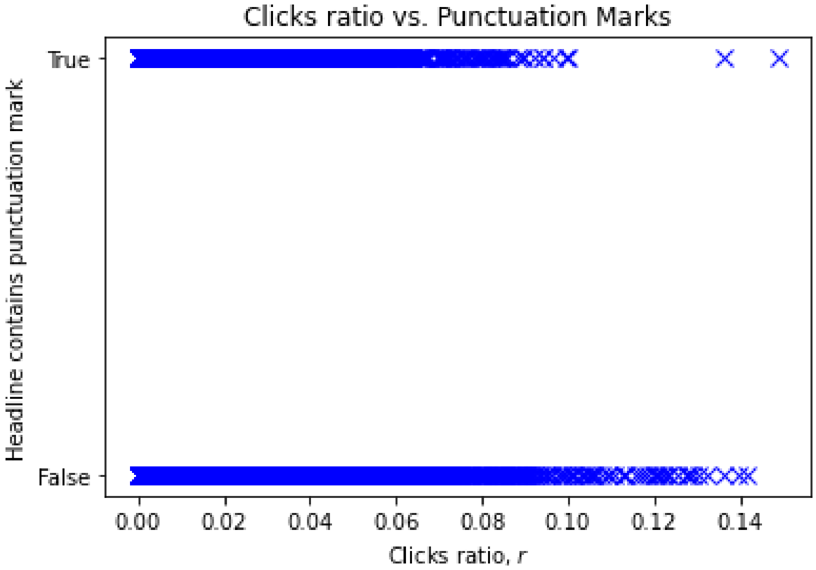

Figure 3: Scatter plot showing the click ratio, r, for headlines which do and do not contain the specified punctuation marks.

$$y = \frac{1}{1 + e^{-(-4.50419434x + 0.0594598)}}$$

Since we have $\beta_0 < 0$, we can see that if there were any significant relationship between punctuation and click ratio, it would be a negative relationship.

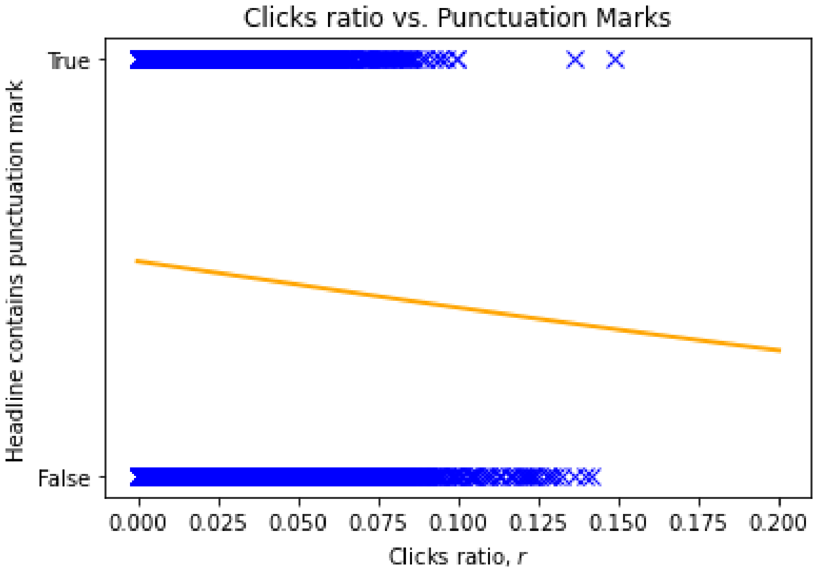

Figure 4: Scatter plot with linear regression line showing the click ratio, r, for headlines which do and do not contain the specified punctuation marks.

We can see from Figure 4 that the relationship is negative, but very weak.

We can then apply the model to the test data. The confusion matrix for this is shown in Figure 5. We see that the false positive and false negative predictions are very significant, meaning that a prediction model based on whether the headline uses the specified punctuation marks is not a good one. The metrics of this model are given below:

Accuracy: 0.4507624983423949

Precision: 0.2571855648088832

Recall: 0.6448918397943885

We see that this model is not more accurate than randomly assigning true or false values to data points, and so the set of punctuation marks specified is not a good predictor of the click ratio, r.

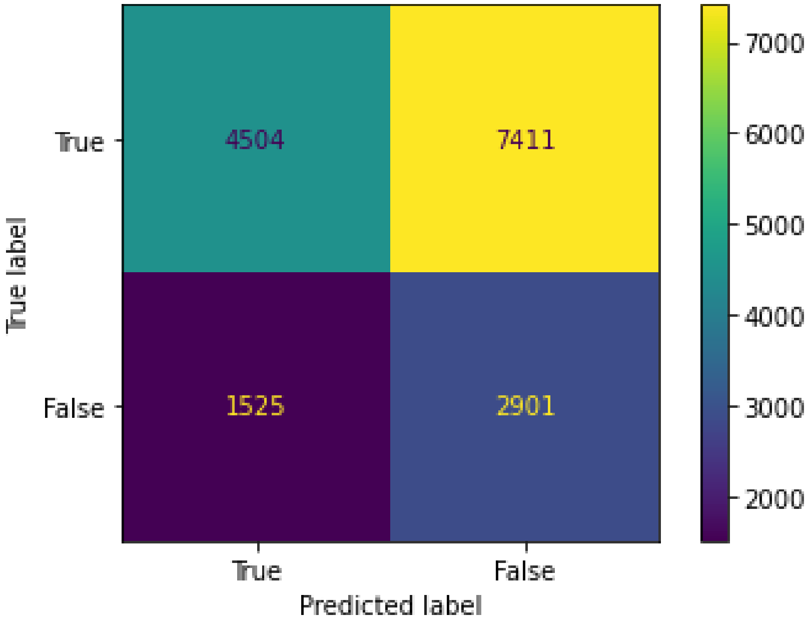

Figure 5: Confusion matrix corresponding to logistic regression model. From left to right and top to bottom, each box represents: true positive, false positive, false negative, true negative.

We also investigated further by looking at specific punctuation marks. This was done by summing impressions and counts over all tests with the same headline to determine the click rate.

The mean of click-rates for headlines with '?' was 0.0116 (3sf) while the mean of click-rates for headlines without '?' was 0.0137 (3sf). This can be seen in Figure 6.

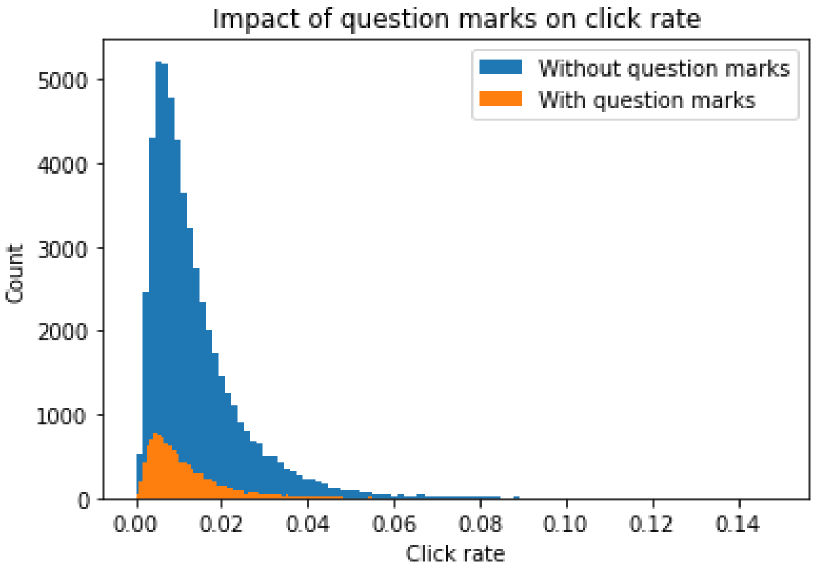

Figure 6: Histogram comparing distributions of titles with and without question marks

We also considered headlines with a quote mark ('). This could have signified a quotation or an abbreviation. This can be seen in Figure 7.

The mean of click-rates for headlines with (') was 0.0132, while without had mean 0.0134.

Headlines with a number digit had a mean clickrate of 0.0135, while headlines without had a mean clickrate of 0.0133.

Another variable we considered was whether the headlines included pronouns. The list of pronouns we considered were ['i', 'you', 'me', 'she', 'her', 'he', 'him', 'it', 'we', 'us', 'they', 'them', "we're", "they're", "you're"] (uppercase / lowercase was ignored). Headlines that contained or didn't contain one of these pronouns both had a mean click rate of 0.0133 (3sf), but the distributions appear slightly different. This can be seen in Figure 10.

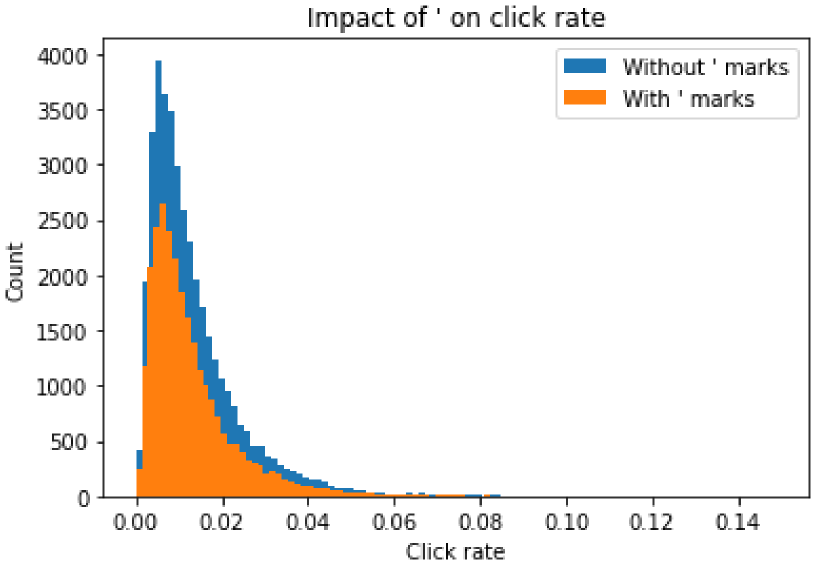

Figure 7: Histogram comparing distributions of titles with and without quotation marks

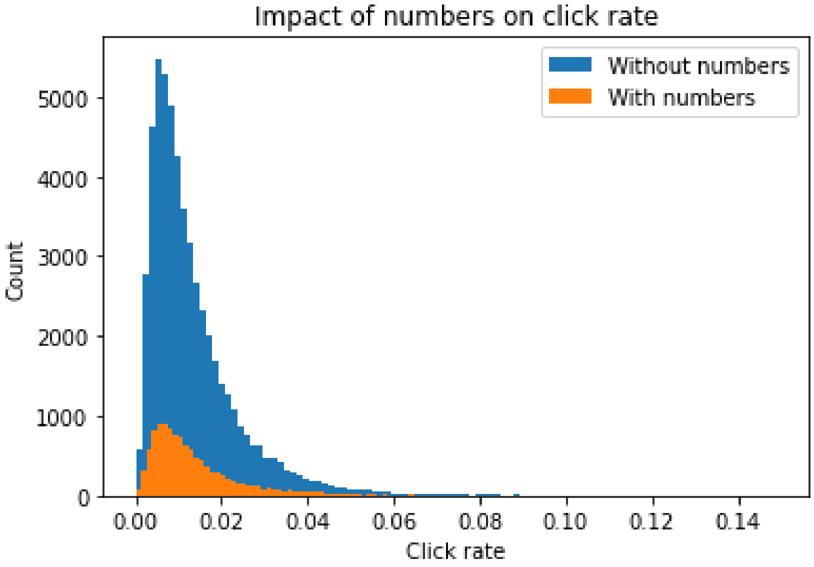

Figure 8: Histogram comparing distributions of titles with and without numreical digits

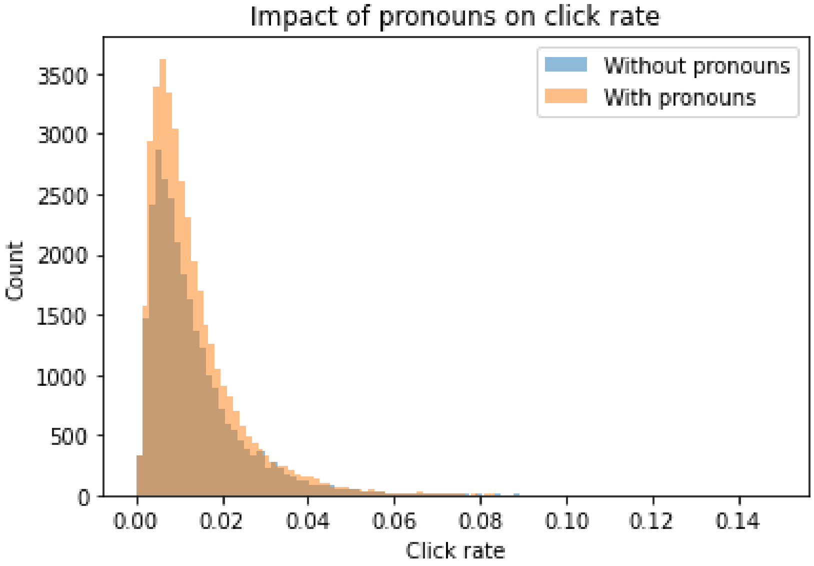

Figure 9: Histogram comparing distributions of titles with and without quotation marks

For the data in figures 6 to 10 tested whether the distributions were different by a Mann-Whitney-Wilcoxon test. This test was chosen because the only assumptions are that all observations are independent of each other (likely to be close to true since we have grouped tests with same headlines together) and that the results are ordinal (true). For each of these, the null hypothesis was that the distributions were equal. For the data in figures 6, 7, 8 (for question marks, quote marks and numbers), the tests returned a p-value of 0. For the pronouns data the p-value returned was 0.011. As all of these were below 0.05, the null hypotheses were rejected for all of them. These suggest that pronouns and punctuation play a role in determining click rate.

3 Appendix

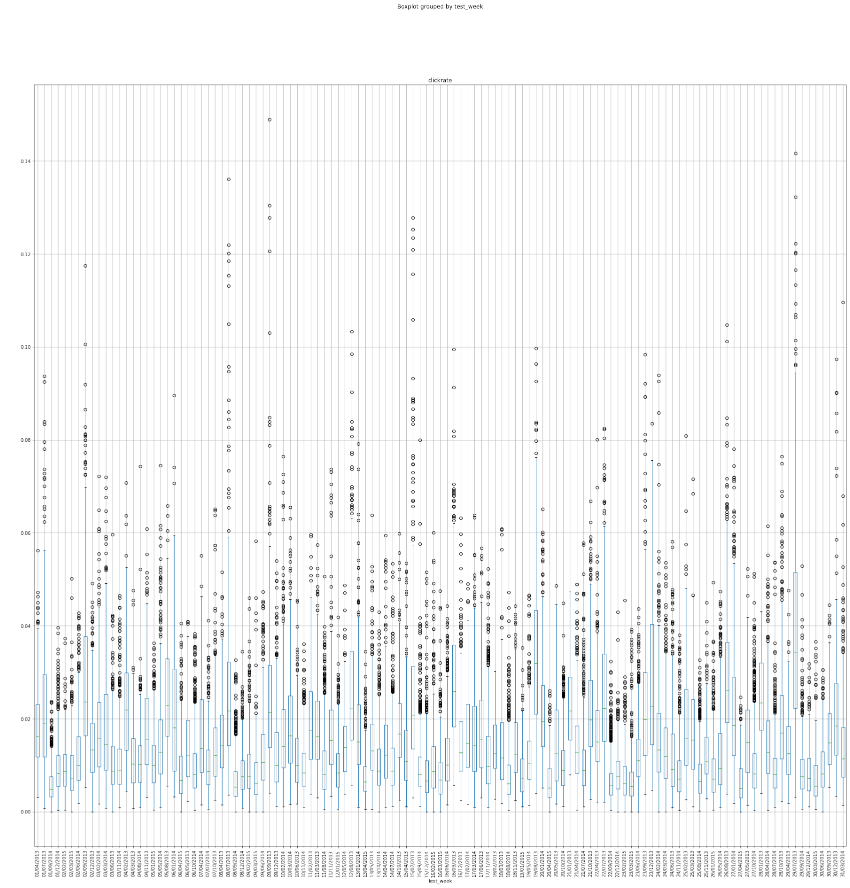

Figure 10: Another variable we briefly considered was whether there are any seasonal trends in the clickrates.