What They Left Behind Is Wondtacular:

Examining the Impact of Emotional Headlines on Clickthrough Rates

Chris Yun, Spencer Weinstein, Sterling Williams-Ceci, and Stephanie Lim

COMM 4940: The Design and Governance of Field Experiments

Prof. J. Nathan Matias

May 15, 2020

Introduction

Digital marketing is a tactic that requires delicate mastery for today's generation of audiences. One of the biggest challenges is successfully enticing people to click on an article in order to maximize its influence. Today's audience is unique from the past; it has an abundance of available resources of information online and is given the power to choose which sources of information to read. One way people consume information in the digital environment is by clicking on story headlines. The headline of an article is what the audience sees first, so the attention articles get is often influenced by whether the headline was captivating enough to be clicked. Hence, organizations often measure an article's popularity with headline click rates. Click rates refer to the ratio of people who click on an article's headline to the total number of people in the audience. Many organizations seek to maximize their articles' click rates to captivate and mobilize supporters (Karpf, 2016).

Upworthy is a website that was dedicated to increasing the click rates for articles by changing the articles' headlines. Upworthy's innovative technique of manipulating the language of a given article's headline, and running A/B tests to observe their popularity, was a turning point for increasing click rates, and it highlighted the importance of a good headline. Furthermore, it opened up the potential for studying how headlines could be used to target certain audiences and increase click rates. For example, behavioral scientists have studied how headlines' subjectivity (i.e., how neutral or opinionated a headline is) and polarity (how emotionally positive or negative it is) impact how many people click and share it.

From the literature summarized below, we hypothesized that subjective headlines would get more clicks than objective ones and that negative headlines would outperform positive ones. Using a dataset of Upworthy's A/B tests on thousands of headlines, we put these hypotheses to test.

Background

To gain insight into how text polarity influences interest, we turned to the literature on sentiment analysis and framing effects on metrics of interest in online content. In a formative study, Jonah Berger and Katherine Milkman (2012) looked at features of newspaper articles as predictors of virality. Using a combination of automatic sentiment analysis (for basic polarity scores) and content analysis techniques (for coding of physiological arousal), the authors found a basic correlation between positive emotion score and virality. Content that was more positive in sentiment was more likely to be shared, and while content of both emotional valences (overall subjectivity) was significantly associated with increased virality, the effect size was larger for positive than negative content. However, this effect was moderated by the arousal provoked by

the content: that is, content of either emotional valence was more likely to be shared if it caused higher arousal (such as anger or awe). In addition, the authors noted that both positive and negative content were more likely to be shared than neutral content.

More recently, Kuiken and colleagues (2017) looked at the classic characteristics of clickbait in relation to the clickthrough rates of online newsletter headlines. "Clickbait" is a term used to describe a writing style meant to grab readers' attention and draw them to click on a headline (Munger et al., 2018). Kuiken and his team used machine learning to predict how different clickbait features, including sentiment polarity, influenced the rate at which headlines from a site called Blendle were clicked. Again, they found that subjectivity (the presence of emotionally-charged words, regardless of valence) predicted higher clickthrough rates than objectivity of headlines (Kuiken et al., 2017), supporting one of Berger and Milkman's main findings. They also found, however, that stories with negative words predicted significantly higher clickthrough rates than stories without them (Kuiken et al., 2017). It is worth noting that this finding didn't represent a comparison between stories with negative versus stories with positive words: instead, it was simply comparing those with the presence of negative words to those with their absence. The researchers did not appear to test the effects of stories with positive content as their own category. This makes it hard to draw conclusions about whether there are meaningful differences in expected clicks on stories with negative versus positive words.

Negative sentiment has also been associated with increased headline attention in other work. In a study of headlines retrieved from The New York Times and The Guardian, headlines were scored on subjectivity, positivity, and negativity using the sentiment analysis tool SentiWordNet. The popularity of headlines was measured as the number of cites they received across Facebook and Twitter. The researchers found that negativity was significantly and positively correlated with the number of cites received on social media for headlines from The Guardian (Piotrkowicz, Dimitrova, Otterbacher, & Markert., 2017). The researchers also discovered that overall subjectivity significantly predicted increased cites of headlines on social media from both news outlets (Piotrkowicz et al., 2017). The results of this study align with those seen in Kuiken and colleagues' paper.

Another study came to the same conclusion as Kuiken et al. and Piotrkowicz et al. that headlines with more emotion (i.e., greater subjectivity) predicted increased clicks (Reis et al., 2015), which further supports the finding that subjective headlines become more popular than objective ones. However, this study also revealed that headlines with both extreme positive and negative sentiment scores gained the most clicks of all, suggesting that either type of emotion can be effective in gaining attention if it is very intense (Reis et al., 2015). This finding supports Berger and Milkman's that the arousal of headlines trumps their emotionality when predicting online virality. The authors also point out that, in their dataset of headlines from four news

sources, the proportion of headlines classified as negative (based on scores from a sentiment analysis program) vastly outnumbered those classified as positive (Reis et al., 2015). This is an important point because the differing proportion of headlines with each sentiment type can confound analyses of popularity: for instance, if there is a higher proportion of headlines with negative than positive language in the dataset to begin with, they may be more likely to be encountered and clicked. If this phenomenon occurred in other studies, it may have accounted for the finding that negative headlines garnered more clicks than positive ones, making it hard to deduce whether the sentiment of the headline was truly the cause of this difference (showing one of the weaknesses with non-experimental studies).

In short, this literature review reveals that there are some conflicting conclusions on how the sentiment of news headlines influences their popularity online. We analyzed a large sample of A/B tests from Upworthy.com to determine the effects of overall subjectivity and specific sentiment (positive compared to negative) on the number of clicks headlines received. Below, we present our hypotheses:

> H1: Headlines that are subjective (regardless of the direction of emotional valence) will receive significantly more clicks than those that are objective ("neutral").
>
> H2: This effect will be larger for headlines classified as negative than for those classified as positive.

Methods

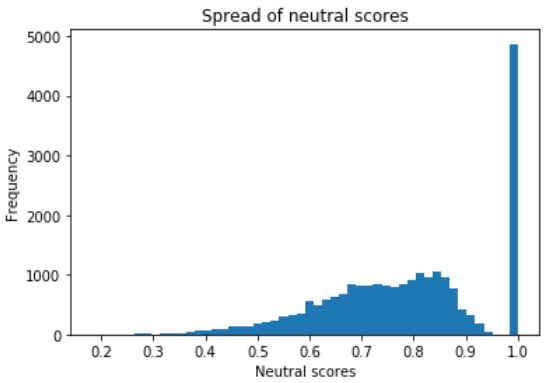
Sentiment analysis (also called Opinion Mining) has become a widely-used mechanism for producing metrics on the emotional tone of text corpuses (Serrano-Guerrero et al., 2015) (Asghar et al., 2014). The process of sentiment prediction attempts to measure the polarity of a given corpus (Asghar et al., 2014). Polarity is the common dependent measure computed by sentiment analysis that refers to the positive, negative, and neutral emotional valence in a text sample (Serrano-Guerrero et al., 2015). Other measures can also be generated by sentiment prediction techniques, including subjectivity of the corpus, which measures how subjective (opinionated) versus objective the given corpus' tone is (Liu, 2010). For the purposes of this project, we look at both the polarity and subjectivity of headlines. There are three general methods for performing a sentiment analysis that is capable of predicting the metrics of polarity and subjectivity for bodies of text: rule-based (naive classification), automatic (machine learning), and hybrid (combination) approaches (Rana & Cheah, 2015) (Ravi & Ravi, 2015). For the scope of this project, we use 3 distinct rule-based approaches for determining polarity and a single rule-based approach for determining subjectivity.

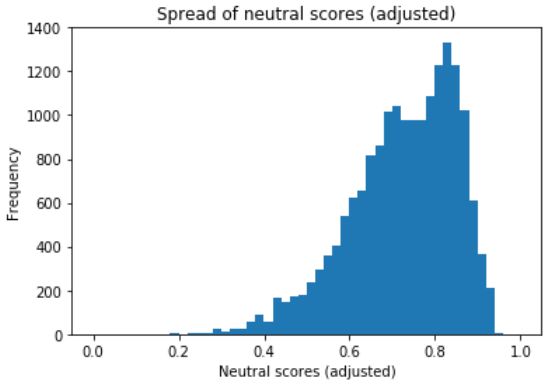
The first method for classifying the polarity of each headline was using a naive rule-based method that takes two pre-existing dictionaries of words provided by LIWC software where one dictionary is known to be positive and one is known to be negative. The naive algorithm counts how many words from each dictionary appear in each headline. Some simple and fairly obvious approaches for classifying these headlines are labeling any headline with more positive words than negative words as positive and vice versa; labeling any headline with at least n positive words and at most m negative words as positive and vice versa; or labeling headlines as positive or negative depending on the ratio of positive (or negative) words to total words by setting a threshold.

An obvious limitation of this naive algorithm described above utilizing LIWC dictionaries is that it can't properly account for a number of factors in text. Little alterations can drastically change the meaning of a body of text; for example, "good" might be a positive word, but "not good" is most certainly not a positive phrase. Punctuation can change the meaning of a piece of text as well; for example, "amazing!!!" is bound to be more positive than "amazing!", which should be considered more positive than "amazing". Our next step in improving our sentiment classification model beyond the barebones approach was to utilize pre-existing python packages that provide APIs for various Natural Language Processing (NLP) tasks.

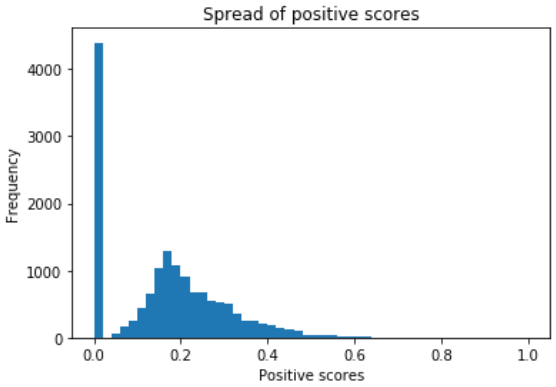
We used two different python packages to assign sentiment scores to the Upworthy headlines. One python package, VADER, generates four scores given a piece of text: neg, neu, pos, and compound. The first three scores, ranging from 0 to 1, represent how likely the text is to be negative, neutral, or positive. The fourth score, compound, is a metric ranging from -1 to 1 that represents the output of various syntactic rules. Another python package, TextBlob, generates two scores given a piece of text: polarity and subjectivity. Polarity ranges from -1 to 1 and represents a range that goes from negative to positive, while subjectivity ranges from 0 to 1 and represents a range that goes from neutral to highly opinionated/subjective. Unlike TextBlob, VADER does not have a subjectivity feature.

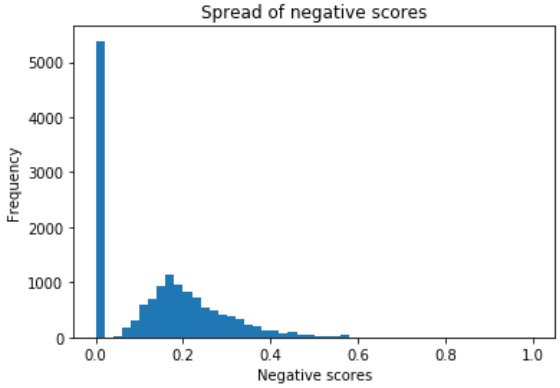
In order to classify headlines as positive, negative or neutral using VADER, we take two approaches. First we classify the headlines based on the pos, neg and neu scores provided by VADER's output. First, we looked at the spread of the neutral scores:



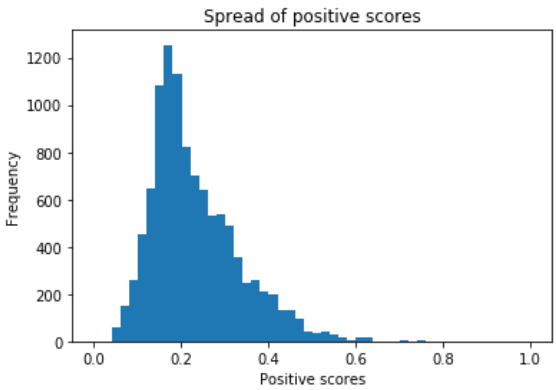


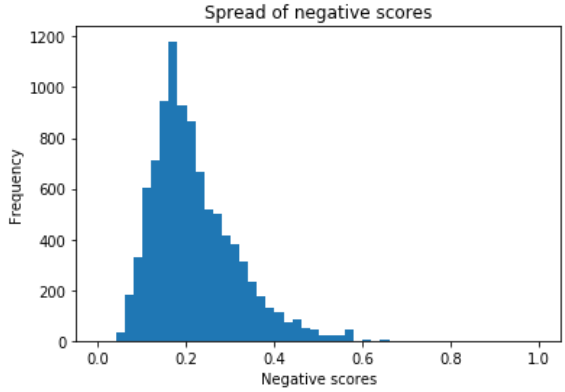
We see that there are a number of neutral scores of 1. Adjusting for this, we get left-skewed distribution. Unadjusted, we get a median of 0.791 and adjusted we get a median of 0.743. Thus, we find it appropriate to have a threshold of 0.85 for classifying headlines as neutral. Discounting these neutral headlines, we get the following distributions for negative and positive scores:





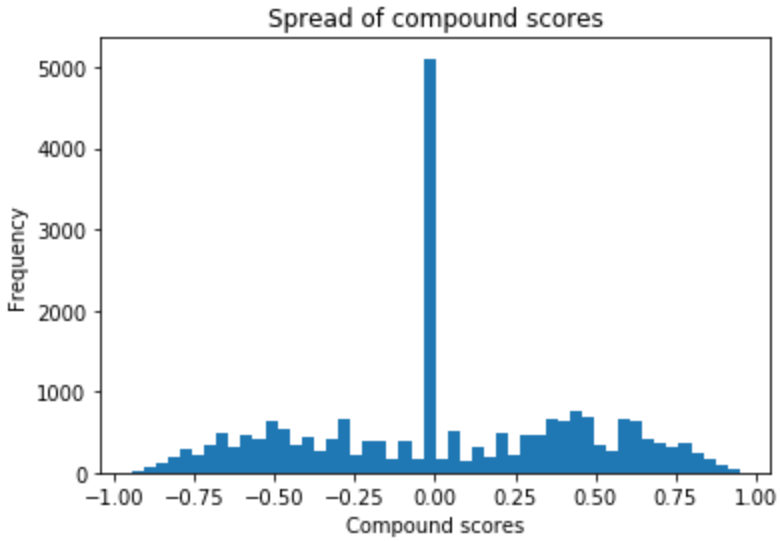
Here, there are a number of scores that are 0. Adjusting for that, we get the following distributions:





Note how there is a peak at around 0.15 for both positive and negative scores. Thus, we feel justified in setting a threshold of 0.15 for classifying headlines as either positive or negative. Otherwise, we classify headlines as neutral.

Next, we classified headlines using the compound score provided by VADER's output. First, we looked at the distribution of the compound scores:



Note that this distribution is relatively symmetrical. Typically, classification of compound scores is as follows: neutral for scores between -0.5 and 0.5, positive for scores above 0.5, and negative for scores below -0.5.

Results

After undergoing initial OLS regression estimates, we find that our most robust models are from the VADER sentiment models and the TextBlob models. While the OLS regressions are a nice start, they run into a major flaw: they do not account for between test variation nor do they account for possible variation from using different images with the same headline. One possible solution of capturing these effects is through the fixed effects model. The results are as follows:

```
==============================================================
Unbalanced Panel: n = 4873            R-Squared:      0.0015615
T = 2-14
N = 22666
==============================================================
                          Estimate  Std. Error t-value  Pr(>|t|)
polarity_vaderNegative  0.00051583  0.00013292  3.8806 0.0001046
polarity_vaderPositive -0.00018325  0.00011978 -1.5298 0.1260731
```

Figure 1: VADER fixed effects model using neutral, positive, and negative scores as thresholds

```
==============================================================
Unbalanced Panel: n = 4873            R-Squared:      0.001059
T = 2-14
N = 22666
==============================================================
                          Estimate  Std. Error t-value  Pr(>|t|)
polarity_vaderNegative  0.00031973  0.00015845  2.0179 0.0436147
polarity_vaderPositive -0.00047512  0.00013388 -3.5489 0.0003879
```

Figure 2: VADER fixed effects model using compound score as threshold

```
==============================================================
Unbalanced Panel: n = 4873            R-Squared:     0.00058477
T = 2-14
N = 22666
==============================================================
                           Estimate  Std. Error t-value Pr(>|t|)
polarity                -0.00039465  0.00016700 -2.3632  0.01813
subjectivity             0.00039116  0.00015584  2.5100  0.01208
```

Figure 3: TextBlob fixed effects model

For both VADER, models, the classifications of positive, negative, and neutral are categorical. In order to avoid the dummy variable trap, which is the multicollinearity of categorical variables, the neutral headline variable is dropped out and used as the baseline for the other variables.

In the VADER classifier where raw polarity scores were used as thresholds, headlines classified as negative got about a 0.00051583 additional clickthrough rate than headlines classified as neutral and headlines classified as positive got about a 0.00018325 smaller clickthrough rate than headlines classified as neutral. Only the negative predictor was statistically significant.

In the VADER model using compound score as the threshold, headlines classified as negative got about a 0.00031973 additional clickthrough rate than headlines classified as neutral and headlines classified as positive got about a 0.00047512 smaller clickthrough rate than headlines classified as neutral. Both negative and positive predictors were statistically significant.

In the TextBlob model, the clickthrough rate decreases by 0.00039465 for every unit increase in polarity. Since negative scores indicate negative headlines, headlines classified as negative do better than neutral headlines. Conversely, since positive scores indicate positive

headlines, headlines classified as positive do worse than neutral headlines. Polarity here is statistically significant. Here, as opposed to the standard OLS regression, subjectivity is statistically significant, where a unit increase in the subjectivity score is associated with a 0.00039116 increase in the clickthrough rate.

Note that the R^2 of all our models are extremely low. However, we justify our research findings despite the low R^2 for the following rationalization: there are many different variables that could affect the clickthrough rate, not just headline sentiment. One could conceive that headline topic, headline length, and so on as relevant variables if we wanted to measure all the variation in clickthrough rates. Since we are looking at the effect of only one variable, the R^2 is largely irrelevant to the focus.

Conclusion

Our research findings partly supported our hypothesis that emotionally negative headlines are more likely to see significantly higher CTRs than the emotionally neutral headlines. Additionally, our team project found that emotionally positive headlines are more likely to see significantly lower CTRs than emotionally neutral headlines. This was a contradiction to our initial hypothesis, as we originally thought that emotionally positive headlines would have significantly higher CTRs than emotionally neutral headlines. Our hypothesis was the emotionally negative headlines would have the most clicks while emotionally neutral headlines would have the least clicks. Our research findings conclude that emotionally negative headlines did have the most clicks, but emotionally positive headlines had the least clicks.

Discussion & Limitations

One of the main limiting factors of our study is that we suspect the headlines deemed "neutral" by the sentiment analysis methods were not truly neutral/objective. By nature, Upworthy's headlines were subjective: the editors carefully crafted them to provoke emotion and curiosity to capture readers' attention (Karpf, 2016; Sobel Fitts, 2014). This means that all the headlines in our dataset likely had some degree of subjectivity. While we observed that subjective language significantly predicted increased clicks on headlines compared to the objective language, we cannot conclude that this difference was truly due to whether a headline was subjective or objective: rather, we can conclude that this difference occurred between headlines that were highly subjective and those that were subjective to a lesser extent.

Still, our results are interesting in that they largely lend support to the preexisting literature's conclusions. We also recognize that there are further directions this type of research can take. For instance, a potential future study could take subsets of the headlines that fit the sentiment categories most strictly and analyze the results of just these headlines' language on

click rates. In our sample, the classifier may have categorized headlines as neutral when their scores were on the cusp between the neutral and negative categories' cutoff. These types of headlines may not be as representative of truly neutral headlines as the ones whose scores fell closer to the center of the distribution. While limiting the sample size for each group this way would limit the statistical power and make it more difficult to detect significant effects if they are small, it would allow for clearer conclusions about click rates on the basis of these discrete sentiment categories.

By analyzing the Upworthy Archive in this way, we improve upon the previous literature pertaining to the impact of headline sentiment on clickthrough rates in a very meaningful way. To our knowledge, the Archive is the only publicly available dataset in which multiple headlines were created for each of many articles and experimented on to determine clickthrough rates for each article's various headlines. These tests made sure that only one feature of a story's frame was manipulated at a time (in our case, the headline's language). Because of this, unlike most previous research in this field, we are able to infer causality by ruling out confounding variables. Our analysis is especially relevant to the study of how sentiment impacts clickthrough rates: after all, these headlines were designed to yield the maximum possible number of clicks by harnessing the power of emotion and curiosity. As this dataset becomes more widely available to academic researchers, we hope that more conclusions about headlines' sentiment can be drawn in future work.

References

Asghar, M. Z., Khan, A., Ahmad, S., & Kundi, F. M. (2014). A Review of Feature Extraction in Sentiment Analysis. *Journal of Basic and Applied Scientific Research, 4*(3), 181-86.

Berger, J. & Milkman, K. (2012). What Makes Online Content Viral? *Journal of Marketing Research, 49*(2), 192-205. https://doi.org/10.1509/jmr.10.0353

Karpf, D. (2016). *Analytic activism: Digital listening and the new political strategy.* Oxford University Press.

Kuiken, J., Schuth, A., Spitters, M. & Marx, M. (2017). Effective Headlines of Newspaper Articles in a Digital Environment. *Digital Journalism, 5*(10), 1300-1314. 10.1080/21670811.2017.1279978

Liu, B. (2010). Sentiment Analysis and Subjectivity. *Handbook of natural language processing*, *2*, 627-666.

Munger, K., Luca, M., Nagler, J., & Tucker, J. (2020). The (Null) Effects of Clickbait Headlines on Polarization, Trust, and Learning. *Public Opinion Quarterly.* https://doi.org/10.1093/poq/nfaa008.

Piotrkowicz, A., Dimitrova, V., Otterbacher, J., & Markert, K. (2017). The Impact of News Values and Linguistic Style on the Popularity of Headlines on Twitter and Facebook. In *AAAI Conference on Web and Social Media.* Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/viewPaper/15775

Rana, T. A. & Cheah, Y. (2015). Hybrid rule-based approach for aspect extraction and categorization from customer reviews. In *2015 9th International Conference on IT in Asia (CITA)* (pp. 1-5). IEEE.

Ravi, K. & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. *Knowledge-Based Systems, 89,* 14-46. https://doi.org/10.1016/j.knosys.2015.06.015.

Reis, J., Benevenuto, F., Olmo, P., Prates, R., Kwak, H., & An, J. (2015). Breaking the News: First Impressions Matter on Online News. In *The International AAAI Conference on Web*

and Social Media (ICWSM). Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/viewFile/10568/10535

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, H. (2015). Sentiment analysis: A review and comparative analysis of web services. *Information Sciences, 311,* 18-38. https://doi.org/10.1016/j.ins.2015.03.040.

Sobel Fitts, A. (2014). The king of content: How Upworthy aims to alter the Web, and could end up altering the world. *Columbia Journalism Review.* https://archives.cjr.org/feature/the_king_of_content.php