,

FINAL REPORT: AD COPY EVALUATION ENGINE

MICHAEL GOH

Applied Data Analysis

Department of Computer Science
University of Chicago

01/29/24

Contents

1. Executive Summary

The ability to predict the success of advertisements remains a relatively unscientific process. Here, we analyze the likely success of headlines on clickthrough rates (CTRs). Specifically, we do so by leveraging the Upworthy Research Archive, a public dataset with over 20,000 A/B headline tests from Upworthy. Our research allowed us to determine if given some specific headline, it would outperform expectations, as measured by CTRs. We have managed to predict if a headline will outperform the median headline with a 69% accuracy rate at 72% precision. This was accomplished using gradient boosted trees, with just sentiment-analysis and text-complexity features. We also successfully correlate the importance of sentiment and text-complexity with CTRs. Our results allow marketers to better understand and predict the likely success of their ad-copy, allowing for better ex-ante decision making.

2. Introduction

Speak to any advertising executive, and they'd likely say that it's an art. It's the art of persuasion, engagement, and conversion. Over the years, however, advertising has become increasingly quantitative. Executives leverage focus groups, quantify outcomes using analytics tools, and measure and benchmark success. That being said, quantitative marketing is still primarily an 'ex–post' iterative exercise. There are few predictive tools to determine the success of advertisements. In this research project, I decided to tackle this challenge by developing an ad-copy evaluation engine to determine if a digital headline would be successful. For the purposes of this project, we conducted our analysis on the Upworthy Research Archive, a public dataset with over 20,000 A/B headline tests from Upworthy, spanning January 2013 to April 2015 [Matias et al.(2021)]. This rich dataset contains 20K+ randomized controlled trials (RCTs) from a period when Upworthy reached over 100M users monthly and conducted extensive A/B testing. In an ideal world, we would have access to RCTs for advertising copy. However, due to the commercial sensitivity of this data, we were unable to obtain it for the purposes of this research. Through my research, I have managed to predict if a headline will outperform the median headline with a 69% accuracy rate at 72% precision. This was accomplished using gradient boosted trees, and achieved with simply sentiment-analysis and text-complexity features.

While there was previous research on the predicted impact of sentiment on CTRs, there was limited research on predicting the efficacy of headlines. Existing research hypothesizes that highly polarizing content is likely to increase CTRs as they are more sticky and elicit greater physiological reactions [Rathje et al.(2021)]. Features commonly referenced to highly relevant to CTRs include sentiment, text-complexity, and emotions.

3. Existing Research

Research Paper	Summary	Methodology	Outcomes
Negativity drives online news consumption [Robertson et al.(2023)]	Explores the effect of negative words on news consumption. They also explore the impact of high-arousal emotions including anger and fear on CTRs. The paper concludes that negative language positively impacts CTRs.	Discarded experiments consisting of single headline variations and merged experiments where there are multiple treatment arms with identical headlines. Leveraged dictionary-based (TF-IDF) sentiment analysis using the LIWC dictionary, accounting for text-complexity (Gunning-Fog, length) and negation, to create unique headline features. Also considered impact of emotions on CTRs. The authors conducted a series of Logistic Regressions to determine the importance of previously mentioned features. The model's performance is assessed by evaluating the p-values for each of the features.	Concluded that negative sentiment is correlated with increase in CTRs. Each additional negative word increases CTR by 2.3%.

Linguistic effects on news headline success: Evidence from thousands of online field experiments [GligoriÄ et al.(2021a)]	Explores the influence of various linguistic traits such as emotion, length, readability, generality, and pronoun usage on headline success. The study concludes that negative sentiment, headline length, usage of indefinite articles, and usage of singular first-person and third-person pronouns improve headline success.	The authors controls for experiments by constructing headline pairs. Headline pairs are constructed such that each pair consists of two headlines from the same A/B test, with the same image and tested in the same week. This controls for external factors like content, author influence, and contextual presentation. In cases where there are more than 15 headline pairs in an A/B test, a random sample of 15 pairs is chosen to prevent bias from any single experiment. Leveraged dictionary-based (TF-IDF) sentiment analysis using the LIWC dictionary, accounting for text-complexity (Flesch Reading Ease) and negation, to create unique headline features. The model's predictive performance is assessed by calculating the accuracy, recall, and precision of the model on the confirmatory dataset.	The study finds that certain linguistic traits, including the use of negative-emotion words, headline length, generality (the use of indefinite articles), and the use of first-person singular and third-person pronouns, improves headline success. In contrast, the use of first-person plural pronouns is negatively associated with headline success. The paper also notes that predictability of headline success based on linguistic features is limited, suggesting that while certain features can be identified, crafting successful headlines is inherently challenging. The authors acknowledge that the findings might not generalize to other contexts outside of Upworthy, and they call for further research to test the applicability of their results in different settings and languages.

4. Data Preprocessing and Feature Engineering

4.1. **Dataset Filtration.** We begin by combining the exploratory, confirmatory, and holdout datasets from Upworthy into one large dataset with 128,151 unique experiments across 27,614 unique tests. Following which, we calculate the CTRs for each experiment. This is done so using the formula below:

$$\text{CTR}_{rl} = \frac{\text{clicks}_{rl}}{\text{impressions}_{rl}}$$

The probability of observing k clicks out of n impressions for some particular headline variation, rl, can is derived in

$$P(X = k) = \binom{n}{k} \theta_{rl}^{k}(1 - \theta_{rl})^{n-k}$$

where θ_{rl} is the CTR of a headline variation l for some article (RCT) r. Given that these tests contain images, we filter out all tests where there are variations in the image used for experiments in the test. We do this by removing all experiments where there are variations in 'eyecatcher_id'. This leaves us with 17,882 unique tests that we are able to draw from. After we specify the type of each field in the dataframe, we proceed to calculate the time-delta for which each experiment was live. Finally, we manually create a training ($n_{experiments} = 62,804$) and testing dataset ($n_{experiments} = 15,702$).

4.2. **TF-IDF Dictionary-based Sentiment Analysis.**

4.2.1. *Text Complexity.* We begin by calculating the Flesch Reading Ease scores for each headline by leveraging the textstat library. The Flesch Reading Ease score is calculated on a 100-point scale based on the average number of syllables per word and words per sentence. A higher score signifies that the text is easier to read. Scores between 90-100 are considered easily understandable by a fifth grader, wheras scores below 30 are best understood by university graduates. The Flesch Reading Ease score can be determined using the following formula:

$$RES = 206.835 - (1.015 * ASL) - (84.6 * ASW)$$

To be comprehensive, we also calculated the Gunning-Fog index, which was used by [Robertson et al.(2023)] in their analysis. The Gunning-Fog index is a readability metric that estimates the years of formal education a person needs to understand a piece of text on the first reading. It's used to measure the complexity of a text and is particularly useful in the field of education and readability research. A higher Gunning-Fog index implies reduced readability. Given that the Flesch Reading Ease score is on a fixed-point scale, we believed that it would function better in a regression analysis. In any case, our attempt at using the Gunning-Fog index did not yield substantially different results. Therefore, we did not end up using the Gunning-Fog index as a feature as the Flesch Reading Ease score was a great proxy for the Gunning-Fog index. For completeness, the Gunning-Fog index can be calculated using the following formula:

$$0.4 \times (ASL + 100 n_{w \geq 3} \sqrt{n_w})$$

Table 2. Using SentiWordNet on Example Experiments (RCT)

Headline Variation	Clickthrough Rates
He **dissed** a **rape victim** on **national tv** and now **needs** to **shut** the **entire hell** up	1.29%
I **wouldn't eat** that if I were you... It's not **poison** or **anything**, it's just...	1.79%
You **might call jacob** a **man** with **lipstick**. **Well**, **jacob** isn't a **man**. **Hear** them **explain** why.	2.00%

Bolded words are words used to determine the positive, negative, as well as overall sentiment score.

where $n_{w\geq 3}$ is the number of words with 3 or more syllables, n_w is the total word count, and ASL is the average sentence length.

Finally, we also calculated the sentence's character length, using it as a proxy for text-complexity.

4.2.2. *Data Preperation and Tokenization.* We begin by substituting contractions with full words. For instance, if the word "won't" exists in the headline, we substitute it with "would not", before removing all non-alphanumeric characters, extra spaces, as well as stop words (using the NLTK stopwords dictionary). Following which, we lemmatize the words using the Word Net Lemmatizer from the NLTK library and convert them into their base or root form. Following which, we tag each word in our sanitized headline with their word type (e.g. noun, verb, adjective, or adverb).

4.2.3. *First Pass Analysis without Negation.* We note that SentiWordNet does not handle negation by default. This first pass analysis using SentiWordNet therefore lacks negation. For each headline, we eliminate all words that are not nouns, verbs, adjectives, or adverbs. We proceed by lemmatizing the words in the headline, and remove all words that lack an available root word. As this method of sentiment analysis is inherently lossy, discarding potentially relevant words is a tradeoff that we need to accommodate. Following which, we determine the relevant synonyms for each word, before calculating the positive, negative, as well as overall sentiment score. We also calculate the polarity score, which is the absolute sum of both the positive and negative sentiment scores:

$$Polarity = abs(Sentiment_{Negative}) + abs(Sentiment_{Positive})$$

4.3. **VADER Dictionary-based Sentiment Analaysis.**

4.3.1. *Text Complexity.* We begin by calculating the Flesch Reading Ease scores for each headline by leveraging the textstat library. This is similar to the method used in our TF-IDF based sentiment analysis. We also calculate the total number of characters for each sentence, using it as a proxy for text complexity.

4.3.2. *Fine-grained Sentiment Analysis.* We utilize Valence Aware Dictionary and Sentiment Reasoner (VADER) to conduct sentiment analysis on our headlines. As VADER automatically handles negation, takes into consideration intensity modifiers, conjunctions, as well as capitalization, we use the VADER model 'out of the box' without making any parameter changes. We applied VADER to each headline without making any adjustments to the headline. We note that the sentiment score incorporates both the positive and negative sentiment, and is calculated using the following formula:

$$Sentiment_{rl} = Positive_{rl} - Negative_{rl}$$

5. Feature Selection and Output Determination

5.1. **VADER: Feature Selection.** Here, we conducted preliminary analysis on our proposed features. We define the dependent variable to be the CTR, and the independent variables to be the min-max normalized time delta, VADER sentiment score, Flesch Reading Ease score, and the min-max normalized length of the sentence. We obtain the following coefficient values for our analysis:

$$\begin{aligned} \beta_{timedelta-minmax} &= 0.0201 & p &= 0.000 \\ \beta_{vader} &= -0.0017 & p &= 0.000 \\ \beta_{flesch} &= 2.12e-06 & p &= 0.118 \\ \beta_{length} &= 0.0119 & p &= 0.000 \end{aligned}$$

Here, we see that no variables appear to have been shrunk to 0, although the impact of the Flesch Reading Ease score appears to have been shrunk close to 0. We hypothesize that this is likely due to the substantial similarities between the FRE score and the length of the headline. Our results generally aligns with existing research [Robertson et al.(2023)], which suggests that CTRs increase with increasing negativity and increased headline length. It also aligns with generally agreed upon marketing principles, that increased text complexity likely increases mental load, reducing interest and engagement [Gkikas et al.(2022)]. We note as well that coefficient for timedelta is also positive, which likely captures the survival bias of headlines that Upworthy keeps live on their website following initial A/B testing.

We see that all variables are statistically significant at 15% LOS. With a Durbin-Watson statistic of around 2, there is likely to also be no significant autocorrelation in the residuals. In addition, we see that the model is statistically significant, with the p-value of the F-statistic = 0. That being said, we obtain a relatively poor R^2 value of 0.184, suggesting that there are a number of other missing features. This makes sense, given that other variables such as the placement of the article on Upworthy, content relevance, audience type, etc likely affects CTRs substantially. Indeed, research has shown that placement of advertisements, the relevance of content, and the type and intent of audience, affects engagement, interest, and clickthroughs on digital media [Shahbaznezhad et al.(2021)]. It is also interesting to point out that our R^2 value compares favorably to existing research that leverage different features [GligoriÄ et al.(2021a), GligoriÄ et al.(2021b), Robertson et al.(2023)].

Despite the reduced significance of the FRE score compared to other variables, we opt to include it in our actual regression analysis, given 1) the lack of features available, and 2) the added variance it could potentially capture beyond the length of the text. Here, we establish that the timedelta, VADER, FRE, and length of a headline will be used henceforth for our predictions. We continue by applying the Standard Scaler to the following variables:

(1) timedelta_numeric: The total number of seconds between when the headline was published and when the analytics were collected
(2) vader: A fine-grained sentiment score derived using the VADER package
(3) flesch_reading: The Flesch Reading Ease score derived from the the NLTK package
(4) length: The length of the headline

5.2. **Pivot and Output Determination.** Recognizing the challenges posed by the model's high significance yet low R-squared value, we pivoted our approach towards evaluating the potential of headlines to surpass average engagement levels. This pivot aims at discerning whether a particular headline has the propensity to secure a click-through rate (CTR) above the median (CTR = 1.19%). Such a methodology, while not without its flaws, notably its lack of precise predictive power and the reductionist nature of its binary outcome (above or below median CTR), offers an entry point into the complex dynamics of headline performance.

This recalibrated approach acknowledges the diverse ways different industries and audience segments engage with content, underlining that median CTR benchmarks are not universal but vary significantly across different contexts. Despite these variations, our goal is to unearth underlying patterns and key components that drive headline engagement above the median threshold, effectively cutting through the clutter of raw engagement data that can be muddied by external variables like fluctuating market conditions or the idiosyncrasies of algorithmic content distribution.

Moreover, this approach serves as a foundational step towards a deeper, more granular analysis of headline efficacy. It allows for a more focused examination of engagement metrics by sidestepping some of the broader challenges associated with measuring raw engagement, thus providing a clearer pathway to understanding how to craft headlines that resonate more profoundly with target audiences.

5.3. **TF-IDF: Feature Selection.** Given that we have pivoted to simply predicting if a specific headline is likely to outperform the median, we define the dependent variable to be whether some headline's CTR is above the median, and the independent variables to be the min-max normalized time delta, positive and negative sentiment scores, overall sentiment score, polarity score, Flesch Reading Ease score, and the min-max normalized length of the sentence. We obtain the following coefficient values for our analysis:

$$\begin{aligned}
\beta_{timedelta-numeric} &= 9.893e-09 & p &= 0.000\\
\beta_{polarity} &= 0.0142 & p &= 0.000\\
\beta_{positive-sentiment} &= -0.0241 & p &= 0.000\\
\beta_{negative-sentiment} &= 0.0383 & p &= 0.000\\
\beta_{flesch} &= 0.0001 & p &= 0.033\\
\beta_{length} &= 0.0025 & p &= 0.000
\end{aligned}$$

Here, we see that all variables are statistically significant at 5% LOS, once again with a Durbin-Watson statistic of around 2, implying there is likely to also be no significant autocorrelation in the residuals. In addition, we see that the model is statistically significant, with the p-value of the F-statistic = 0. Once again, we obtain a relatively poor R^2 value of 0.103, suggesting that there are a number of other missing features. Similar to our coefficients derived on VADER coefficients, we see that the articles with positive sentiment are more likely to have lower CTRs. We also see that more complex texts are likely to have lower CTRs, and longer texts are likely to have higher CTRs. Despite the lack of negation, the results that we derive from using TF-IDF is remarkably similar to those that we obtain from VADER.

6. Logistic Regression

6.1. **Methodology.** Given that we have changed our prediction variable, we will also z-normalize our non-scaled features prior to training our model. Depending on the method of sentiment analysis, we apply the scikit-learn StandardScaler package to different variables. We proceed by running a logistic regression using the scikit-learn Logistic Regression module on these specified features for the test dataset to predict if a headline is likely to outperform the median CTR in the Upworthy dataset.

6.2. **Outperformance Prediction with TF-IDF for Sentiment Analysis.** We conduct a logistic regression using the scikit-learn LinearRegression module, using the z-scaled features as dependent variables. Here, we see that a simple logistic regression allows us to obtain an accuracy score of 64%. This is a statistically significant result, and is better than the base score of 50% (given that half of sample is expected to be above the median). In addition, we see that the precision of identifying if a headline is likely to outperform is 64% (weighted average).

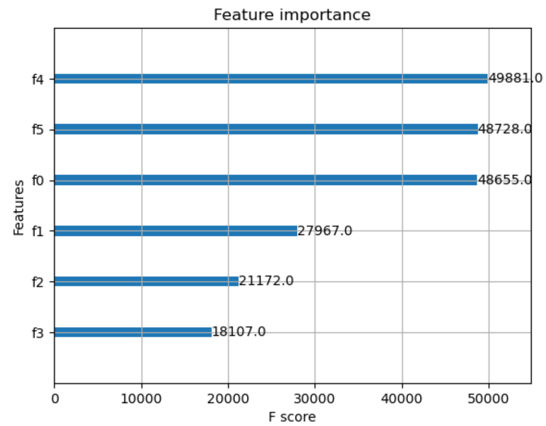
6.3. **Outperformance Prediction with VADER for Sentiment Analysis.** We conduct a logistic regression using the scikit-learn LinearRegression module, using the z-scaled features as dependent variables. Here, we see that a simple logistic regression allows us to obtain an accuracy score of 65%. This is a statistically significant result, and is better than the base score of 50% (given that half of sample is expected to be above the median). In addition, we see that the precision of identifying if a headline is likely to outperform is 65% (weighted average).

6.4. **Outperformance Prediction with Logistic Regression.** While we are able to obtain a statistically significant predictive power using a Logistic Regression (on whether a headline is likely to outperform the median), we do not see major differences between features generated using TF-IDF or VADER. We continue by evaluating the efficacy of other methods to predict the performance of a headline.

7. Gradient Boosted Trees

7.1. **Methodology.** In order to leverage Gradient Boosted Trees (GBTs) efficiently, we utilize the xgboost library instead of the scikit-learn library to create the trees. We utilize the RandomizedSearchCV library from scikit-learn to determine the most effective parameters for our GBTs. Here, we vary the number of estimators, learning rate, as well as the maximum tree depth to determine the best possible model for use in our training. The number of estimators that we test varies from 100 to 1,000, learning rate varies from 0.01 to 0.5, and the maximum depth

FIGURE 7.1. GBT Feature Importance (TF-IDF)



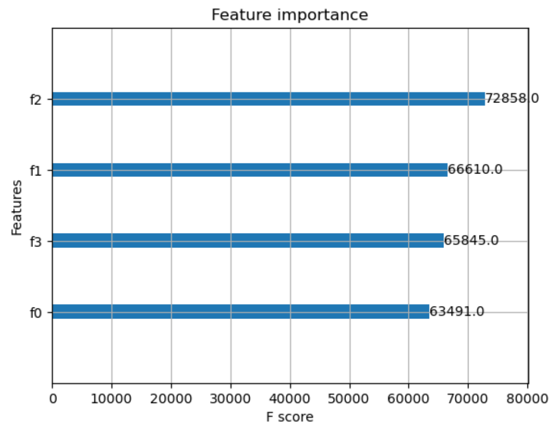
where f0 = timedelta_numeric, f1 = polarity, f2 = positive_sentiment, f3 = negative_sentiment, f4 = length, and f5 = flesch_reading.

of the tree varies from 3 to 15. We favor random search over grid search in this hyperparameter tuning phase, especially given the expansive size of our datasets. Random search offers a more pragmatic and computationally viable approach by sampling a subset of the hyperparameter space, thereby providing a comprehensive yet efficient exploration of possible model configurations. Here, we apply 3-fold cross validation within the training dataset to determine the best possible model based on accuracy. We then apply the best model onto the specified features on our test dataset, before proceeding to determine the efficacy of the model based on accuracy and precision.

7.2. **Outperformance Prediction with TF-IDF for Sentiment Analysis.**

Utilizing the aforementioned RandomSearchCV parameters, we arrive at a GBT with a learning rate of 0.116, maximum depth of 14, and 376 estimators. Surprisingly, when we apply Gradient Boosted Trees (GBTs) in conjunction with our TF-IDF-based sentiment analysis approach, we observe a notable decrease in performance. In particular, the accuracy rate drops to 61%, and precision similarly declines to 62%. This outcome is unexpected given the typically robust performance of GBTs in various machine learning tasks. We hypothesize that the underperformance could be due to overfitting, or a lack of negation handling, leading to more pronounced differences in prediction performance between the 2 datasets. The GBT model's ranking of feature importance presents an intriguing narrative, particularly when juxtaposed with prior research. The model prioritizes 'Length' as the most critical feature, followed by 'Flesch Reading Ease', 'Timedelta', 'Polarity', 'Positive Sentiment', and finally 'Negative Sentiment'. This order is especially noteworthy as it appears to challenge the conclusions from our regression analysis, which suggests that text complexity is less important as features compared to sentiment. Such a discrepancy invites a closer examination of the underlying factors that might contribute to these divergent findings.

Figure 7.2. GBT Feature Importance (VADER)



where f0 = timedelta_numeric, f1 = vader, f2 = flesch_reading, f3 = length.

7.3. **Outperformance Prediction with VADER for Sentiment Analysis.**

Utilizing the aforementioned RandomSearchCV parameters, we arrive at a GBT with a learning rate of 0.0880, maximum depth of 13, and 558 estimators. Unlike our dataset with TF-IDF-based sentiment analysis, GBTs outperform Logistic Regressions in a statistically significant fashion when applied to our VADER-based dataset. In particular, the accuracy rate increases to 69%, and precision similarly increases to 69%. We theorize that our VADER-based dataset outperforms as it may capture non-linear relationships and interaction effects between words and sentiment more effectively for the specific dataset. GBTs, being non-linear models, can leverage non-linear relationships to model complex relationships and interactions between features, leading to better performance compared to linear models like logistic regression [Yang and Zhai(2022)]. In addition, the preprocessing steps involved in generating VADER scores might also play a role in reducing noise and highlighting the most important sentiment-related information in the text, which can further enhance the performance of GBTs. Here, we see once again that text complexity, as proxied by the Flesch Reading Score, is the most important feature, followed by the sentiment score obtained with VADER, the length of the headline, and the length of time which the headline was live on the site. Once again, this aligns with our analysis conducted on our TF-IDF dataset, and contracts the feature importance as illustrated in through our regression analysis.

8. Practical Implications

Our research serves as a foundational effort in forecasting the likelihood of success for headlines and, by implication, advertising copy. It helps us identify shared characteristics that might be influential across different domains. This knowledge can be leveraged by content creators to optimize their titles for enhanced engagement and by platforms aiming to recommend content more effectively. Moreover, understanding these linguistic predictors aids in designing communication that aligns with the audience's preferences, potentially increasing the visibility and impact of important messages in domains such as public health or education.

9. Future Research

We aim to include additional variables in our models, such as the use of various pronouns and the topical relevance of headlines, while also exploring advanced language-model-based approaches to enhance click-through rate (CTR) predictions. Unfortunately, due to existing limitation of resources (time and financial), I was unable to leverage LMs for sentiment analysis and CTR predictions. The intention is to develop a sophisticated model that not only captures the nuances of language used in headlines but also adapts to the evolving ways in which users interact with content in the digital space. In addition, we could also consider developing a ensemble model, leveraging more than one model to predict the performance of headlines.

10. Conclusion

Our studies have demonstrated that features like text complexity, text length, and the overarching sentiment have the capacity to predict headline success with statistical significance. Moreover, a nuanced sentiment analysis and the application of non-linear modeling techniques, such as Gradient Boosted Trees, have shown to enhance predictive performance considerably. These findings underscore the power of linguistic cues in influencing user engagement and open up new avenues for content optimization, offering a more nuanced understanding of how different elements of text can sway the success of digital content.

References

[Gkikas et al.(2022)] Dimitris C. Gkikas, Katerina Tzafilkou, Prokopis K. Theodoridis, Aristogiannis Garmpis, and Marios C. Gkikas. 2022. How do text characteristics impact user engagement in social media posts: Modeling content readability, length, and hashtags number in Facebook. *International Journal of Information Management Data Insights* 2, 1 (2022), 100067. https://doi.org/10.1016/j.jjimei.2022.100067

[GligoriÄ et al.(2021a)] Kristina GligoriÄ, George Lifchits, Robert West, and Ashton Anderson. 2021a. Linguistic effects on news headline success: Evidence from thousands of online field experiments (Registered Report Protocol). *PLOS ONE* 16, 9 (2021), e0257091. https://doi.org/10.1371/journal.pone.0257091

[GligoriÄ et al.(2021b)] Kristina GligoriÄ, George Lifchits, Robert West, and Ashton Anderson. 2021b. Linguistic effects on news headline success: Evidence from thousands of online field experiments (Registered Report Protocol). *PLOS ONE* 16, 9 (2021), e0257091. https://doi.org/10.1371/journal.pone.0257091

[Matias et al.(2021)] J. Nathan Matias, Kevin Munger, Marianne Aubin Le Quere, and Charles Ebersole. 2021. The Upworthy Research Archive, a time series of 32,487 experiments in U.S. media. *Scientific Data* 8, 1 (2021), 195. https://doi.org/10.1038/s41597-021-00934-7

[Rathje et al.(2021)] S. Rathje, J. J. Van Bavel, and S. van der Linden. 2021. Out-group animosity drives engagement on social media. *Proc Natl Acad Sci U S A* 118, 26 (2021). https://doi.org/10.1073/pnas.2024292118 1091-6490 Rathje, Steve Orcid: 0000-0001-6727-571x Van Bavel, Jay J Orcid: 0000-0002-2520-0442 van der Linden, Sander Orcid: 0000-0002-0269-1744 Journal Article Research Support, Non-U.S. Gov't United States 2021/06/25 Proc Natl Acad Sci U S A. 2021 Jun 29;118(26):e2024292118. doi: 10.1073/pnas.2024292118..

[Robertson et al.(2023)] Claire E. Robertson, Nicolas PrÃ¶llochs, Kaoru Schwarzenegger, Philip PÃ€rnamets, Jay J. Van Bavel, and Stefan Feuerriegel. 2023. Negativity drives online news consumption. *Nature Human Behaviour* 7, 5 (2023), 812–822. https://doi.org/10.1038/s41562-023-01538-4

[Shahbaznezhad et al.(2021)] Hamidreza Shahbaznezhad, Rebecca Dolan, and Mona Rashidirad. 2021. The Role of Social Media Content Format and Platform in Usersâ Engagement Behavior. *Journal of Interactive Marketing* 53, 1 (2021), 47–65. https://doi.org/10.1016/j.intmar.2020.05.001

[Yang and Zhai(2022)] Yanwu Yang and Panyu Zhai. 2022. Click-through rate prediction in online advertising: A literature review. *Information Processing Management* 59, 2 (2022), 102853. https://doi.org/10.1016/j.ipm.2021.102853