Unveiling Clickbait: A Dive into Sentiments, Emotions and Topics

Ranking and quantifying user engagement with clickbait articles with NLP-created features

Luka Blagojevic
Department of Network and Data Science
Central European University, Vienna
lukablagojevic1995@gmail.com

November 14, 2023

Contents

1 Executive Summary

In this report, I analyze what properties of the clickbait articles make users want to click on them. I determined three features of the article text - 'sentiment', 'emotion', and 'topic' and then, by using their combination as a unique fingerprint of text, I ranked and compared how much users tend to click them on individual, group, and daily levels. With these rank metrics, I analyze how much each of the features contributed in different scenarios, leading to insights about their interplay and potential product ideas.

1.1 Introduction

Clickbait headlines, crafted to get people to click, have become widespread in the past decade. These catchy titles are designed to boost website traffic, a crucial measure for publishers and content creators looking to increase visibility and revenue. Additionally, the easy sharing of clickbait on social media makes content go viral, spreading across online social networks.

Despite its initial success, clickbait has negative effects. When the promised content doesn't live up to expectations, it damages user trust and hurts the credibility of both the content provider and the platform. Prioritizing sensationalism over accuracy also lowers the overall quality of information, contributing to the spread of misinformation and distracting users. That's why spotting and preventing clickbait is crucial for social media networks and tech companies to enhance the content quality users see.

Berger and Milkman's study on viral online content [1] shows that emotional arousal plays an important role in engagement with online content, making a difference between effects caused by positive and negative emotions. At the same time, some studies show that the level of engagement is different in different topics, and varies when different words are used. [4] I use some of these findings as a starting point. Hence, to understand how sentiment, emotion, and topics influence user engagement with clickbait articles, I employ Natural Language Processing (NLP) methods, as illustrated in Figure 1. Then, I measure user engagement by analyzing the correlation between these NLP-derived features and the click-through rates of the articles. By ranking these combinations of 'emotion', 'sentiment, and 'topic' (referred to as 'triplet'), I manage to capture different aspects of user engagement with the content. Finally, I introduce a product concept of a "Clickbait defender" in Figure 1, that could be utilized to detect and filter clickbait, using my methodology.

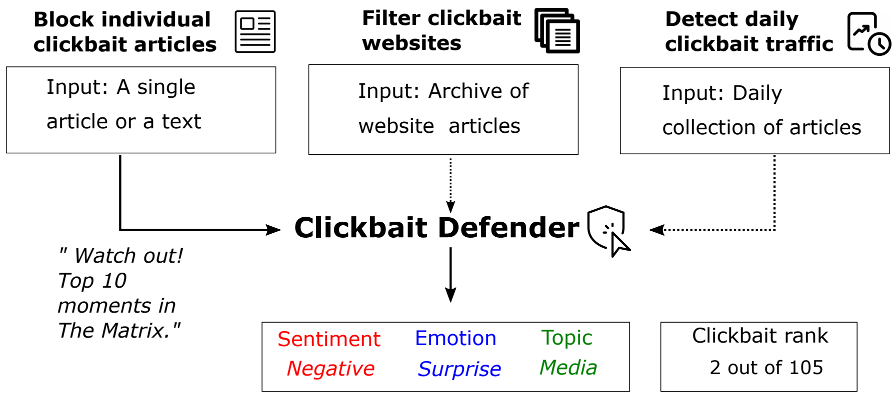

Figure 1: Product concept of a "Clickbait Defender", is analogous to my research methodology. I determine 'sentiment', 'emotion', and 'topic' on three different levels - individual articles, aggregated groups of articles, and daily traffic of articles, which I rank relative to each other, increasing the chance of identifying clickbait content.

1.2 Key Findings

Below I summarize my findings that apply to the context of Upworthy, a known clickbait website, in the period of 2013-2015. The findings apply to the textual aspect of the clickbait articles A/B tested on the website for a fixed time period:

1. The website traffic was composed of new users and driven by social media behavior (i.e. Facebook post sharing), but it is not related to user engagement with the website [Subsection.2.3]

2. There was a lot of daily variation in the proportion of Sentiment, Emotions, and Topics of the tested articles, but with stable long-term proportions. Both positive and negative were close in proportion, with joy being the dominant emotion and "Society and Social Issues" being the dominant topic. [Subsection.2.5]

3. Compared to other investigated options, Clickbait articles with "angry" or "sad" emotional tones, which are about "Society and Social Issues" or "Pop Culture, Media and Entertainment" were determined as the most "clickbaity". [Subsection 2.6]

4. Emotion and Sentiment seem to have a significant impact on user engagement, while this is not true for the Topic (which might be related to the vagueness of topics that were tested). [Subsection 2.7]

5. Some, but not all combinations of Sentiment, Emotion, and Topics have shown that they cannot arise by random user engagement or errors in classification. [Subsection 2.8]

2 Technical exposition

This section delves into the technical aspects of the research, including data context, assumptions, and detailed explanations of the methodologies used in data cleaning, processing, and analysis.

2.1 Data Context & Working Assumptions

It is important to contextualize Upworthy within the broader scope of its operational environment. As a website, Upworthy is not just a platform for content dissemination; it operates within a complex ecosystem influenced by various business, marketing, and technological factors. These elements could have a significant impact on the data I am analyzing.

For instance, Upworthy's content strategy, marketing tactics, and even the design of the website itself could influence user behavior. Changes in marketing strategies, shifts in content focus, or alterations in website layout and functionality can all have profound effects on how users interact with the site. Additionally, external factors such as the evolving landscape of social media, internet usage trends, and socio-political events could indirectly affect user engagement and traffic patterns.

For example, a key turning point in user traffic occurred with Facebook's algorithm change [3], leading to a decline in Upworthy's external traffic. This highlighted the importance of external sources in driving user engagement for such sites.

Regarding the dataset, I encountered some ambiguities. The exact lengths of the A/B test periods were unclear. While the 'test_week' column indicates when a test commenced, it does not specify its duration. Moreover, the 'created_at' date, which I use to approximate the start of each test, might not align with the actual beginning of these tests. Nevertheless, given the data limitations, I consider the 'created_at' date a reasonable approximation starting date of the test.

I also made a critical assumption about the nature of Upworthy's content. Given the site's operational context and its reputation, I presumed that the entire article corpus could be categorized as clickbait. This assumption is crucial as it frames my approach to analyzing the data and interpreting the results within the specific context of Upworthy's content strategy and user interaction patterns.

2.2 Data Cleaning & Processing

In my approach to cleaning and processing the dataset from Upworthy, I made several critical decisions to refine the data for more accurate analysis. These steps were essential to eliminate irrelevant variables and focus on meaningful metrics that could reveal insights into user engagement and content effectiveness.

Initially, I removed the 'excerpt' column as it predominantly contained the phrase "Things that matter", showing little variation (approximately 52%). I also excluded the 'share_text' and 'share_image' columns. These elements, determined by users who shared the article, introduced an unpredictable level of variation, and without additional context on user motivations or profiles, they were deemed extraneous for the study.

Further, I decided to drop the 'winner' and 'first place' columns. My focus was not on the editors' choices or their changing significance metrics over time. This decision was informed by the understanding that editorial decisions are influenced by a multitude of factors like team size, individual editor preferences, and overarching publishing strategies.

The 'slug' column, primarily serving as an internal web address, was also removed. The 'headline' and 'lede' columns provided sufficient text data for analysis.

To understand the variations in the A/B test data, I analyzed metrics such as 'impressions', 'clicks', the number of unique 'image_ids', and the combined text from 'lede' and 'headline' (merged into a 'combined_text' column). This exploration revealed significant variations in the data:

A/B Test Metric	**Value**
Average impressions relative to the total mean	0.061 ± 0.084
Average clicks relative to the total mean	0.521 ± 0.416
Average ctr relative to the total mean	0.522 ± 0.406
Median number of unique headline + lede per test_id	4.0 (IQR: $4.0 - 5.0$)
Median number of unique images per test_id	4.0 (IQR: $4.0 - 5.0$)

Table 1: **Upworthy A/B tests introduced a lot of variation between the test articles:** This table shows the summary of the A/B test metrics done by Upworthy website curators. They typically had four different article texts and images, which gained roughly equal user views (impressions), contributing to substantial variation in user engagement (clicks). Note: The measure of variation for the first three metrics is their standard deviation.

The results from 1 indicate that the A/B tests conducted by Upworthy indeed introduced significant variation in article content. This variation is sufficient to treat the test instances as distinct articles. Despite the lack of image data, I decided to treat each article as an individual entity based on its text, clicks, and impressions. In this context, the absence of image data is considered an external and unavoidable source of noise in the dataset. This approach assumes that each experiment was conducted independently and that random user assignment to different A/B tests negates any potential cross-influence between duplicate articles in the same test phase.

Moreover, while the exact test durations are unknown and there's no clear correlation between the date and other metrics, I assumed that the tests lasted for approximately equal durations, as suggested in the Upworthy archive. This assumption, while critical, is based on limited information and could be a potential area for further inquiry.

2.3 User Behavior Analysis

To gain a comprehensive understanding of user behavior, I analyze the provided Google Analytics data, along with daily clicks and impressions data, which are intertwined with the users.

The bulk of users were from the US and other English-speaking countries (Canada, the UK, and Australia). Given that the website's primary language was English, this demographic distribution simplified the task of text analysis, as both linguistic and cultural contexts were predominantly influenced by American English, with 64% of total users hailing from the United States, as shown in Figure 5:

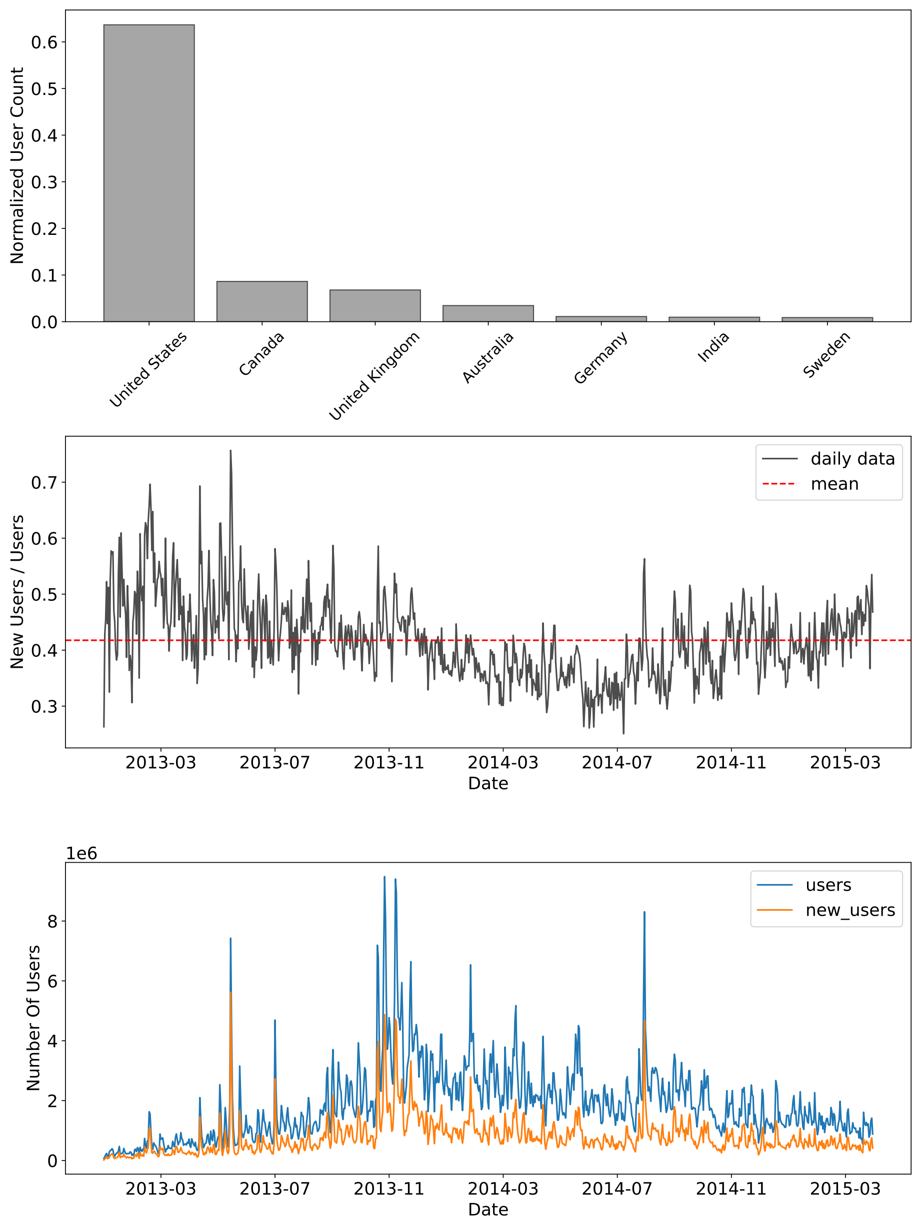

Figure 2: **First-time visitors from the United States make most of the website traffic:** **(Top)** The majority of users come from English-speaking countries, with the United States as the dominant user location. **(Middle)** The ratio of new users to all users is approximately half, sometimes dipping to a quarter. **(Bottom)** There is a notable correlation in the daily traffic patterns of new and returning users.

In the context of the analysis, each visitor to the website was treated as a distinct user. This categorization included labeling individuals accessing the site from a device for the first time as new users. Key metrics employed to gauge user engagement included the bounce rate and pages per session. The bounce rate represented the proportion of sessions that were either very brief (under ten seconds), did not result in any conversion, or involved viewing just a single page. Pages per session, on the other hand, measured the average number of pages a user navigated through during a single visit

The mean ratio of new to all users was 0.42 ± 0.07, peaking at 0.76 and dipping to 0.25 at its lowest. Users typically read 1-2 articles per session, as indicated by the mean number of pages per session being 1.33 ± 0.13, and the mean bounce rate at 0.58 ± 0.35. This pattern aligns with the site's reputation as a clickbait platform, heavily reliant on viral external traffic, particularly from social media platforms like Facebook.

To analyze the relationships between various user engagement metrics, I calculated the Spearman rank correlation (as it doesn't require normally distributed variables) in Table 2:

	Users	Pageviews	Daily CTR	New Users	Impressions
Users	1.00***	0.99***	-0.01	0.96***	0.53***
Pageviews	0.99***	1.00***	-0.01	0.93***	0.54***
Daily CTR	-0.01	-0.01	1.00***	-0.01	-0.23***
New Users	0.96***	0.93***	-0.01	1.00***	0.50***
Daily Impress.	0.53***	0.54***	-0.23***	0.50***	1.00***

Table 2: **More website traffic leads to more views, but not to more clicks and user engagement:** Spearman Rank Correlation Coefficients between Google Analytics metrics - Users, New Users, and Pageviews and Upworthy A/B test metrics - Daily Impressions and Daily Click-Trough-Rate (CTR). Significance levels denoted as follows: $***p < 0.001$, $**p < 0.01$, $*p < 0.05$.

My analysis revealed a strong and significant correlation between users, new users, and page views, while the click-through rate did not correlate with any other analytics metric. This lack of correlation adds credibility to the click-through rate as a reliable metric of user engagement, independent of the website's visitor count.

It's crucial to note that the accuracy of the Google Analytics data is contingent on the correct implementation of its tracking code on the website, a factor I couldn't verify. Nonetheless, the findings from this analysis suggest that many external factors, such as events or platforms like Facebook, could have influenced user arrival on the website. The data supports the hypothesis that users were primarily drawn in by clickbait content, engaging superficially with the site's articles.

2.4 User Engagement Metrics

When I referenced the daily peaks with a database of external events, only some of them corresponded, These events might drive the article headlines or topics, but it wasn't strong evidence. External averages for clickthrough rates

Man Whitney test was done to compare the means and it was statistically significant for daily ctr and individual article ctr, meaning that the distributions are different. Although similar when visually inspected, I cannot just aggregate them.

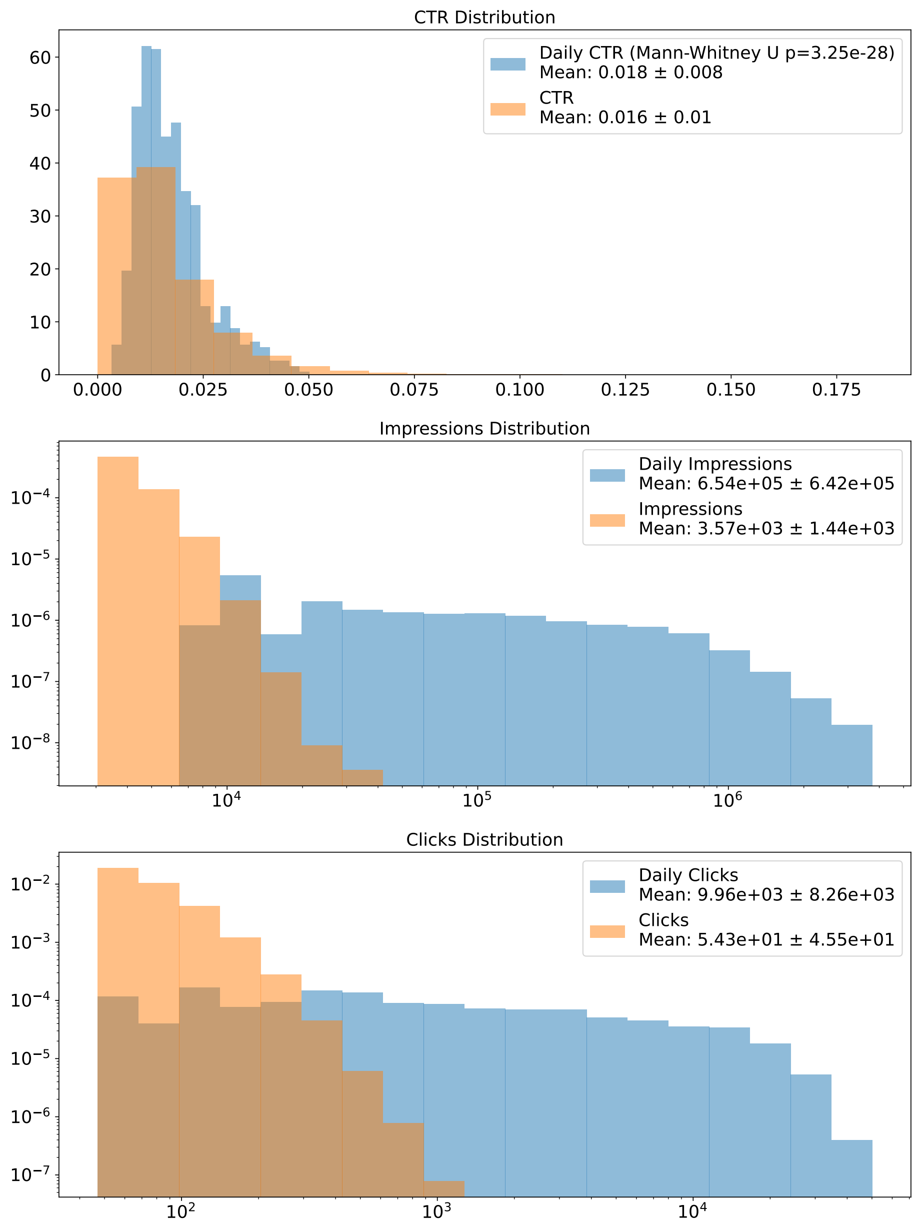

Figure 3: **Clicktrough rate remains a stable measure both on an individual article and aggregated daily levels: (Top)** Distributions of individual article CTR and daily CTR have similar, but statistically different means and variance **Middle** A log-log plot of the individual article and daily sum of impressions, which both have long tails and sharp peaks. **(Bottom)** A log-log plot of the individual article and daily sum of clicks which follow similar distribution patterns as impressions.

To further understand the relationship between these metrics, I conducted a Spearman rank correlation analysis, as shown in the following Table 3:

	ctr	impress.	clicks
ctr	1.0***	-0.13***	0.889***
impress.	-0.13***	1.0***	0.287***
clicks	0.889***	0.287***	1.0***

	daily_ctr	daily_impress.	daily_clicks
daily_ctr	1.0***	-0.231***	0.135***
daily_impress.	-0.231***	1.0***	0.91***
daily_clicks	0.135***	0.91***	1.0***

Table 3: **Spearman rank correlations for individual article and daily aggregated data:** All relationships are statistically significant at the significance level $p = 0.01$. The results show that if something gets more impressions, it is less likely to get clicked.

In addition, I employed the 'scipy.peaks' method with a set peak prominence of 20% to identify significant peaks in daily CTR. This analysis, combined with a rolling mean CTR plot, revealed discernible patterns of high and low user engagement.

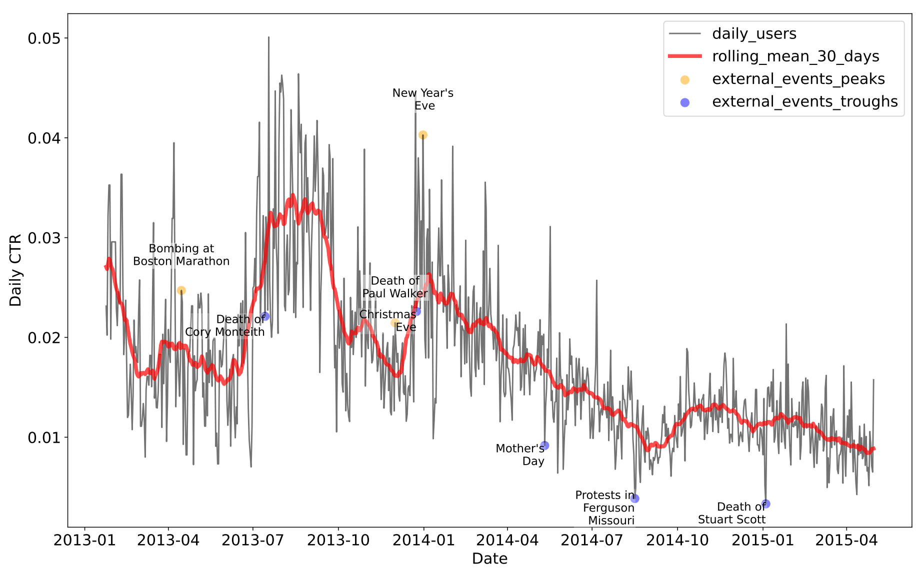

Figure 4: **Events and CTR:** Peaks in Daily CTR were matched against annotated data from Twitter highs and lows using the Hedonometer database [2], indicating a possible link with public sentiment.

The daily data was characterized by significant noise, with notable peaks potentially linked to external events. Of the 49 external events analyzed, only 3 corresponded with CTR peaks and 5 with troughs. These events, associated with highs and lows in public sentiment on Twitter[2], underscore the potential emotional engagement of users with content on the website.

2.5 NLP Content Analysis

The headlines under analysis predominantly featured a combination of text and imagery. In this study, while image data was not accessible, I focused on the textual elements, specifically the headlines and ledes. These components were merged for a comprehensive NLP analysis. Additional textual content, such as shared text, was available but was not included in this study's scope to maintain a focused approach. For the classification and analysis of these textual elements, I employed a selection of pre-trained NLP models, as detailed in the subsequent Table 4:

Pre-training For	Models
Zero-Shot Classification	bart-large-mnli
DistilBert for Sentiment	distilbert-base-uncased-finetuned-sst-2-english
BERT for Emotion	bert-base-uncased-emotion

Table 4: **Summary of NLP Models Used for Text Classification:** These were pre-trained models designed for specific tasks as described, which I used for my own classification.

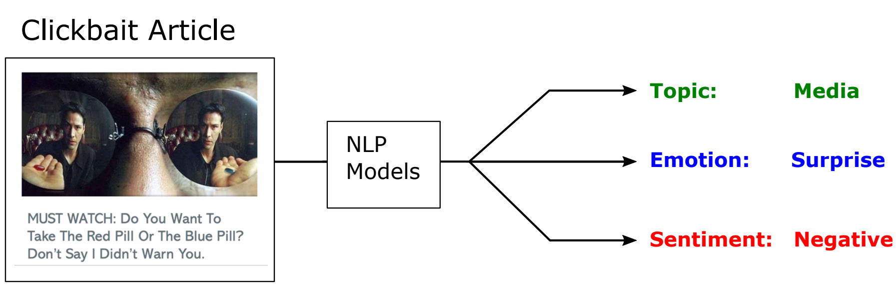

Figure 5: **Each article was classified by 'sentiment, 'emotion', and 'topic':** By using three different NLP models, I managed to classify text into three different categories of importance.

Due to time constraints, I used a manual approach to data exploration and analysis. Additionally, the zero-shot learning model used in the study lacked pre-defined topics, so I had to determine them myself. To address this, I conducted Latent Dirichlet Allocation (LDA) and analyzed the most frequent words. This analysis was supplemented by external information sourced from Upworthy's blog, specifically the article titled "The Most Upworthy Topics of 2013" [Upworthy, 2013]. Based on these findings, I introduced additional categories in my analysis, resulting in three distinct labels: unclassified sentiment, unclassified emotion, and unclassified topic. Notably, the classifiers for sentiment and emotion yielded more definitive scores than the topic classifier, which tended to be broader and encompassed multiple subtopics.

I looked at the rolling means (30 days) of the daily proportions of each class being present in articles, stationary patterns can be observed, which are all confirmed by doing the ADF test of stationarity for time-series, as shown in Figure 6.:

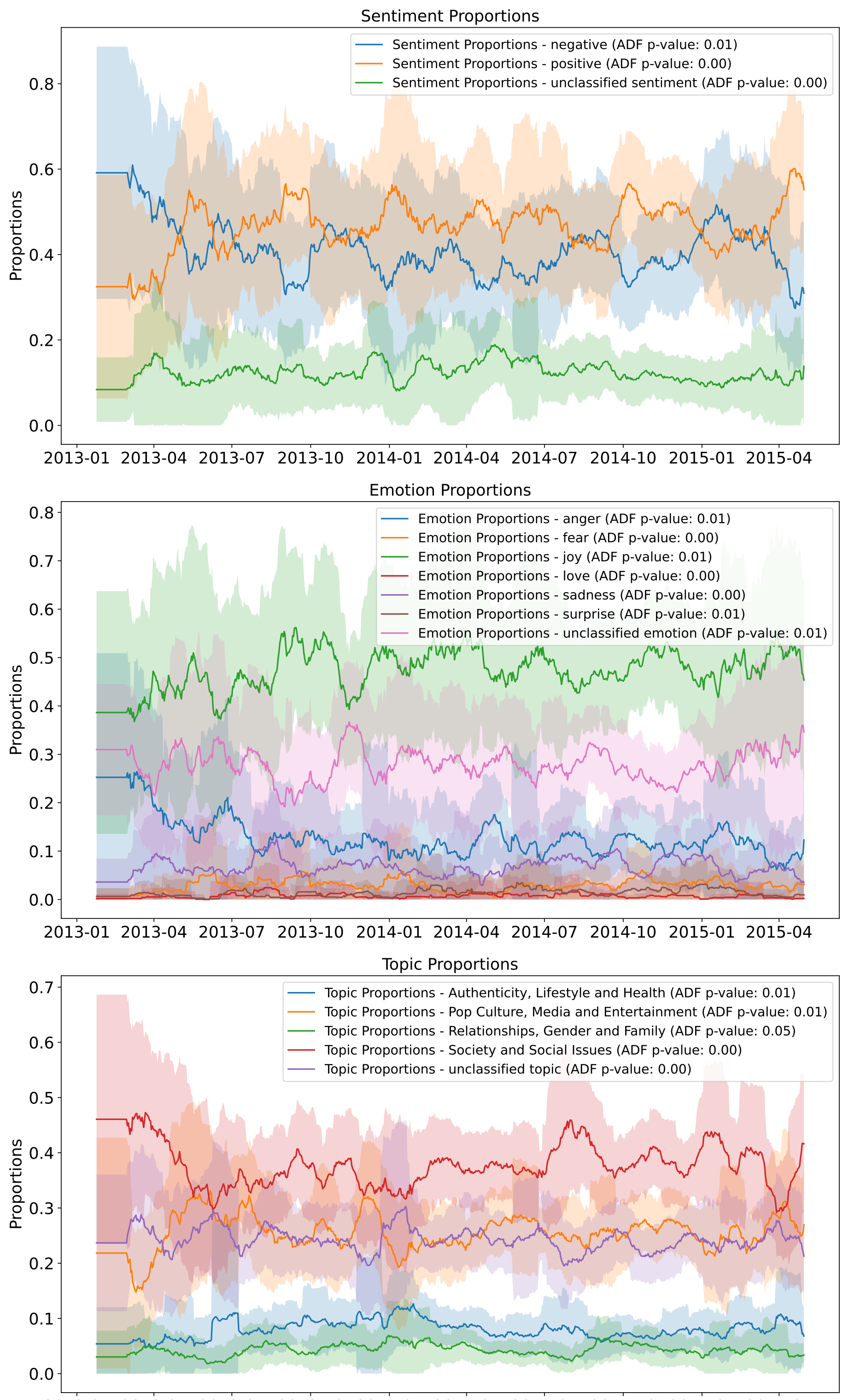

Figure 6: **Proportions of Sentiment, Emotion, and Topics on 30-day rolling means:**All time-series are stationary, yielding stable proportions across time.

The proportion time series results show stable temporal patterns in how clickbait articles were constructed, on a macro level. The insights from this NLP analysis formed the basis for the subsequent ranking and evaluation of the clickbait content on different levels - individual, aggregated and daily rankings of articles.

2.6 Ranking Clickbait

To investigate the textual complexities of clickbait articles on Upworthy, I developed a unique ranking system. This system was designed to evaluate how different articles stack up against each other based on their Sentiment, Emotion, and Topic, on three different levels of performance. The ranking was determined by answering three critical questions:

1. **Top Performers Analysis:** I first looked at the most frequent combinations of Sentiment, Emotion, and Topic among the top 5% of articles, sorted by their Click-Through Rate (CTR). This allowed me to rank these combinations based on their frequency in this high-performing group. The rank $rank_i$ assigned here is the individual rank.

2. **Overall Impact Assessment:** Next, I evaluated the entire collection of articles to identify which combinations of Sentiment, Emotion, and Topic had the highest mean CTR. This step was crucial in understanding which groups of triplets were, on average, more effective in driving click-throughs. The rank $rank_g$ determined from this analysis is the group rank.

3. **Daily Trends Evaluation:** Lastly, I analyzed the daily testing data to find out which Sentiment, Emotion, and Topic triplet was most frequently the largest contributor to clicks each day. This daily ranking adds another dimension to understanding the performance of these triplets. The rank is $rank_d$ from this daily analysis.

With these three ranks, I introduced a composite rank as a measure to capture all aspects. This composite rank is the simple mean of the three ranks:

$$< rank >= \frac{rank_i + rank_g + rank_d}{3}$$

For instance, the triplet combination of Positive, Joy, and Relationships/Gender and Family might have different individual, group, and daily ranks, (i.e $rank_i = 1$, $rank_g = 15$, $rank_d = 2$)which when averaged $< rank >= 6$, should provide a comprehensive view of its overall effectiveness in the context of attracting users.

Although I could have used absolute measures, this ranking method offers a more robust comparison since all three metrics are constructed differently. Additionally, the composite measure is interpretable and evenly weighted. However, there is an inherent sensitivity in this method as it involves comparing aggregated measures and groups. Further refinement and additional tests could potentially enhance the accuracy of these rankings.

$< rank >$	emotion	sentiment	topic
17.7	anger	positive	unclassified topic
18.7	anger	unclassified sentiment	Society and Social Issues
18.7	anger	positive	Society and Social Issues
19.0	anger	negative	unclassified topic
20.0	sadness	positive	Pop Culture, Media and Entertainment
21.3	anger	positive	Authenticity, Lifestyle and Health
22.3	anger	unclassified sentiment	unclassified topic
22.3	sadness	positive	Society and Social Issues
22.7	anger	negative	Pop Culture, Media and Entertainment
23.0	anger	negative	Society and Social Issues

Table 5: **Top 10 triplets by** $< rank >$: Emotions of anger and sadness in conjunction of Society and Social Issues dominate the rankings.

2.7 Aggregated group statistics of triplets

I wanted to understand the interplay between these elements and their collective impact on the effectiveness of clickbait content, as there might be general trends to be observed if triplets are aggregated into groups per class.

An essential part of my statistical exploration involved the use of the Kruskal-Wallis H test. This decision was based on the realization that the data did not satisfy the normality and equal variance prerequisites for an ANOVA test. The Kruskal-Wallis H test is a non-parametric method, ideal for situations where the data may not follow a normal distribution, thus offering a more suitable and reliable way to discern differences across multiple groups, as shown in 7:

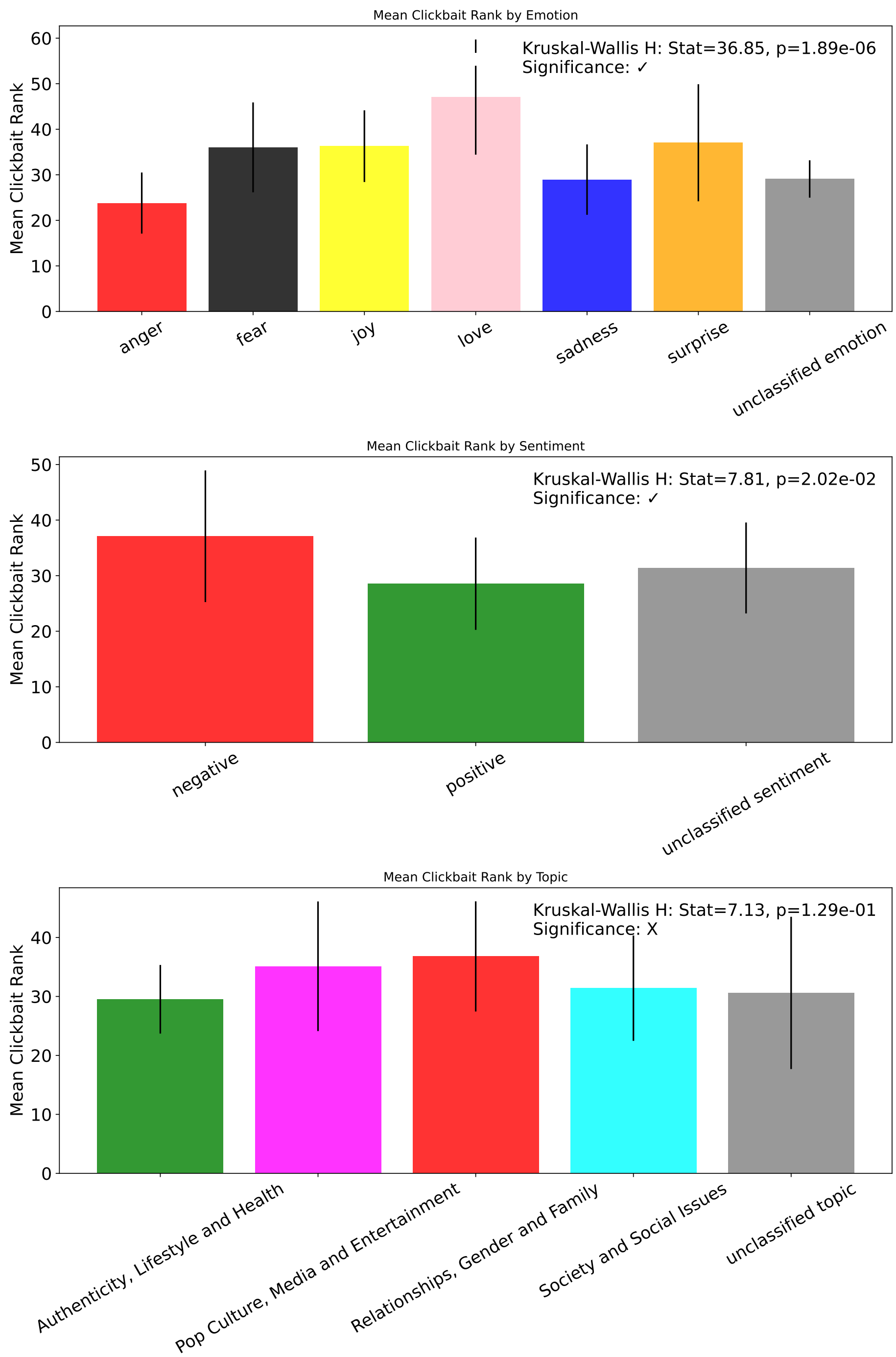

Figure 7: **Aggregated group statistics of combinations (triplets) of Sentiment, Emotions, and Topics:** Aggregated group statistics of triplets show a statistically significant difference for emotions and sentiment.

The findings from this test showed a significant impact of Emotion and Sentiment of distribution of mean clickbait ranks, highlighting their importance in determining how content is perceived and interacted with in terms of clickbait potential.

On the other hand, Topic did not exhibit a statistically significant difference in median clickbait ranks. This suggests that, unlike emotional and sentimental factors, the topic of content may not be as critical in influencing its effectiveness as clickbait. This could be due to the broad categories that were selected during the NLP classification, which potentially diluted their impact.

The results suggest that focusing on the emotional and sentimental aspects of content could be more effective in enhancing its appeal or effectiveness in attracting clicks. However, the analysis also confirmed the limitation in the effectiveness of the NLP classifier used for categorizing topics. This suggests that employing a more refined NLP classifier or utilizing more specific topic categorizations could potentially yield more insightful results regarding the influence of content topics on clickbait effectiveness.

2.8 Null Models for Ranks

Recognizing the limitations inherent in NLP classification, such as the potential for generating a high volume of unclassified text or external events influencing user traffic, I developed two null models using the same dataset categories to emphasize the significance of 'emotion', 'sentiment', and 'topic' triplets compared to random distributions, obtained by feature permutation (resampling without replacement). The primary goal was to determine if the observed mean clickbait rank for each triplet in the original dataset significantly differed from what might be expected by chance (as determined by shuffled data), as explained in 8:

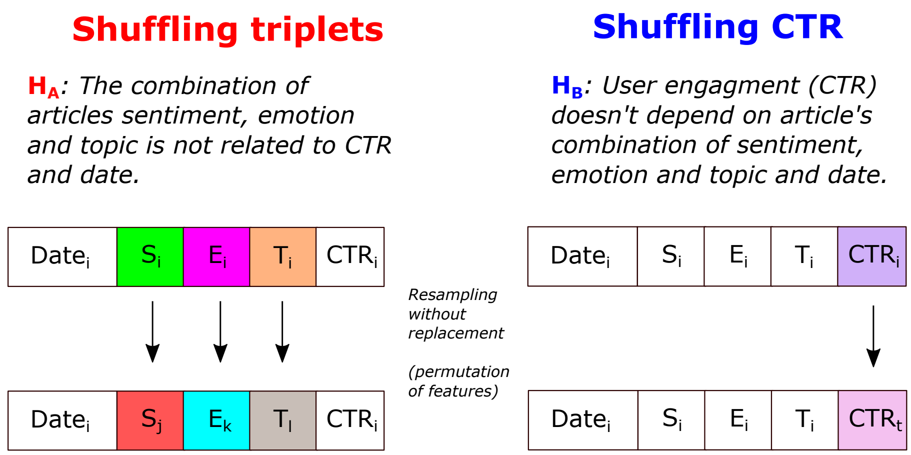

Figure 8: **Schema of Null models for classification errors and random user assignment**: two null models I used, with their corresponding hypotheses. The entire dataset is permuted in the described manner and this procedure is repeated for $n = 50$ trials. For each trial, mean clickbait ranking was computed, thus yielding a null model with 50 data points, to which the original value was compared for statistical significance.

This methodology aims to assess the impact of various 'emotion', 'sentiment', and 'topic' combinations on the mean clickbait rank. It's crucial to determine whether the observed mean clickbait rank for each triplet in the original dataset significantly deviates from what might be expected in random scenarios, as illustrated and explained in Figure 8, while the entire procedure is outlined below:

1. **Shuffling Procedure**: I shuffled each triplet in the original dataset, randomizing the 'sentiment', 'emotion', and 'topic' columns. This procedure aimed to break any existing association between the triplets and the mean clickbait rank, simulating a null scenario.

2. **Shapiro-Wilk Test**: To determine whether the distribution of mean clickbait ranks from the shuffled data follows a normal distribution, I used the Shapiro-Wilk test. A p-value less than the threshold indicates a non-normal distribution.

3. **Bonferroni Correction**: Considering I conducted two tests for each triplet, I divided the significance level $\alpha = 0.05$ by Bonferroni correction, yielding $\alpha_B = \frac{0{,}05}{2}$. This adjustment was necessary to reduce the likelihood of type I errors (false positives) and was particularly relevant since each triplet is unique and can be considered independent and should provide a conservative estimate

Through these statistical tests, I aimed to identify how many of my ranking measures were different from the aforementioned randomizing scenarios, and their results are shown in **??**:

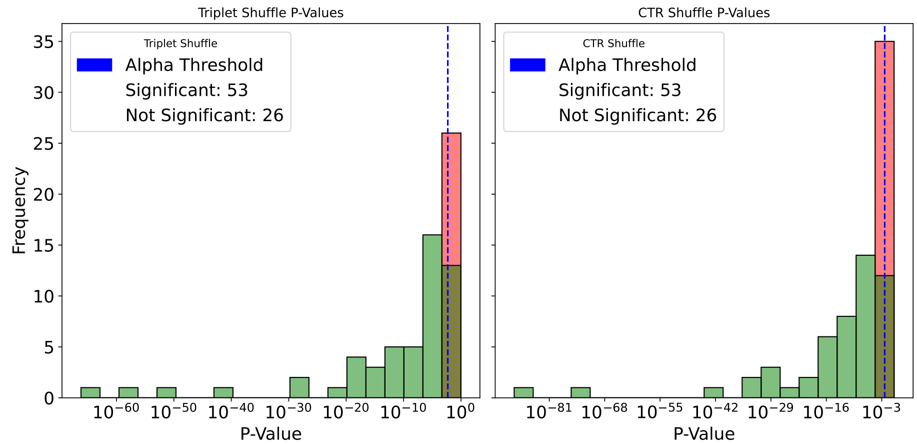

Figure 9: **Null Model Testing Results: (Left)** Triplet Shuffling Null Model (**Right**) CTR shuffling null model. A significant result in either the Z-Test or the Mann-Whitney U Test indicates a meaningful association between a triplet and its clickbait rank. The results indicate that some triplet ranks can be reproduced by random effects (null models).

3 Conclusion

This study has provided a multifaceted examination of clickbait content, particularly analyzing its emotional, sentimental, and topical dimensions. By employing advanced Natural Language Processing techniques and robust statistical analyses, I have gained novel insights into the mechanics of user engagement in the context of clickbait.

Key findings of this research indicate a significant influence of emotional and sentimental elements on user interaction with clickbait content. Particularly, emotions such as anger and sadness, when coupled with topics like Society and Social Issues, exhibited a higher propensity to engage users. Interestingly, the topic of an article, while an integral component of its content, did not show a significant impact on user engagement as compared to emotional and sentimental aspects. This emphasizes the pivotal role of emotional resonance in driving user behavior online, aligning with prior research on online content virality.

The introduction of a unique ranking system, based on the combined metrics of sentiment, emotion, and topic, presents a novel approach to analyzing clickbait content. This system not only offers a more nuanced understanding of what makes clickbait effective but also lays the groundwork for developing tools and strategies to manage such content, as exemplified by the proposed 'Clickbait Defender' model.

However, this study is not without its limitations. The reliance on specific NLP models and the constraints of the dataset, particularly in terms of the unclassified categories and the absence of image data, point to areas where further refinement is needed. Additionally, the study's focus on a single clickbait-heavy website may limit the generalizability of the findings. Future research could expand this analysis to a broader range of websites and content types, possibly incorporating more sophisticated NLP techniques and a more diverse array of data sources.

In conclusion, the insights garnered from this research contribute significantly to our understanding of clickbait, shedding light on the intricate interplay of emotions, sentiment, and content topics in driving user engagement. These findings not only have implications for content creators and marketers seeking to optimize user engagement but also offer valuable insights for technology companies and platforms aiming to improve the quality and credibility of online content. Looking forward, this study paves the way for more in-depth investigations into the dynamic landscape of digital content and user engagement.

References

[1] Jonah Berger and Katherine L. Milkman. What makes online content viral? *Journal of Marketing Research*, 49(2):192–205, 2012. Accessed on 10-12-23.

[2] P. S. Dodds, E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, K. Megerdoomian, M. T. McMahon, B. F. Tivnan, and C. M. Danforth. Human language reveals a universal positivity bias, 2014. Preprint available at http://arxiv.org/abs/1406.3855.

[3] Facebook. News feed fyi: Click-baiting, 2014. Accessed on 10-12-23.

[4] Joon Hee Kim, Eun Kyung Lee, Soo Min Kim, Sang Hoon Lee, and Meeyoung Cha. How to compete online for news audience: Modeling words that attract clicks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 487–496. ACM, 2016.