

A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

Driving Engagement through Emotional Content: A Data-Driven Analysis from The Upworthy Research Archive

Clickbait Detection and Its Implications Using a Machine Learning Approach Applied to the Upworthy Archive

Tomás Martins Amorim e Sá

Work project carried out under the supervision of:

Professor Michael Kummer

23/01/2024

Abstract

In a digital landscape where capturing reader attention is crucial to news platforms, the Upworthy Research Archive enables the study of how article headlines influence engagement. This research assesses how the components of a headline impact user clicks through machine learning methods to extract features related to emotional and phrase-meaning topics to predict the level of engagement on each piece of news. Results show that timing dominates engagement, headline structure plays a key role, and emotional traits have limited influence.

Additional analysis explores clickbait detection and analysis of its impact in the Upworthy Research Archive.

Keywords

Digital Content, Engagement, Upworthy, MIND, Prediction, News Articles, Digital Content, Headline, Excerpt, Categories, Classification Algorithms, Machine Learning, Natural Language Processing, User Interaction, Content Strategy, Interaction Trends, Patterns, Insights Clickbait Trends, Plots, Statistical Analysis, t-Test, Regression Analysis, Hypothesis Testing, Model Evaluation

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

1. Introduction

In the modern digital era, one of the major challenges content producers face is standing out among the overwhelming deluge of information. Digital content is being uploaded as never before, so it is harder than ever to capture and hold audiences' attention. Creating content that connects with viewers on a deeper emotional level and incentivizes them to share it is key to breaking through the clutter. Platforms like Upworthy are renowned for their highly engaging and emotionally charged content and can provide insightful lessons in this regard. A unique opportunity to analyze what makes some content engaging or not is the Upworthy Research Archive, a comprehensive dataset compiled by Matias et al. (2021), with information on the realm of audience engagement, headlines, and content performance. This study delves into the data to discover how Upworthy's headlines contain psychological and emotional triggers that affect engagement.

The main objective of this investigation is to analyze and explore how the emotional effect of content affects audience engagement. Content that is charged with meaningful and impactful emotions, such as joy, rage, surprise, or urgency, tends to pique people's curiosity and encourage viewers to connect with and share it. This can be a great tool for content producers as putting the right amount of a certain emotion into content might lead to greater engagement. However, it can be very challenging to pinpoint the precise emotional cues that promote engagement and sharing. This research analyzes the emotional and structural components that are most indicative of high engagement by examining the headlines and excerpts from the Upworthy Research Archive. Additionally, this study delves into machine learning models that use these emotional traits and other attributes to forecast the performance of a piece of content.

This research focuses on two main research questions. Firstly, what aspects of the headlines, such as the word choice, emotional tone, and structure, are most relevant to predict a strong engagement in Upworthy material? Headlines are very important in the success of digital content as usually are the first thing a viewer sees, how well they grasp attention directly affects whether a reader clicks through the content, or simply ignores it. This research investigated the aspects that contribute to a headline's effectiveness, including sentiment (positive, negative, or neutral), emotional triggers, and whether specific structures (such as the use of questions, figures, and length) might increase engagement.

The second main research question of this investigation was, how can machine learning models predict a piece of content's engagement level based on its headline and preliminary content description? By building predictive models with Upworthy Research Archive data, it can be assessed how accurately these elements can predict the performance of content, and what is the best way to do it. The goal was to create predictive models that can identify trends in both the headline and emotional structure and how that directly influences engagement from the audience. This is insightful for content producers to perfect the art of headline creation, maximizing audience reach in a very congested digital market.

To perform this analysis on the Upworthy Research Archive data, various processes were employed such as data understanding, collection, cleaning, feature engineering, and others. Ensuring the data is reliable and valid for analysis is of the utmost importance, dealing with outliers and invalid data, and the collection of more data are crucial steps. To get the most out of the data, it was also collected information from the Microsoft News Dataset (MIND) that granted the creation of a new and important feature in the original data, the categories of each article. Furthermore, to define the engagement of digital content, the Click-Through Rate (CTR) was used to represent the proportion of users who interact with the content by clicking on it. Furthermore, the textual elements of the Upworthy Research Archive were analyzed

using various natural language processing (NLP) methods and techniques, such as Sentiment Analysis (SA). This approach has allowed to pinpoint the emotional tone of each headline, assess the intensity of each emotion, and determine to what topic it relates. The insights extracted from this analysis were crucial to predict engagement, as emotional cues and the topic derived from the headlines can influence how viewers interact with digital content. All these techniques were decisive as they provided opportunities for new analysis and deeper interpretation of the data.

When focusing on the second research question of this study, various machine learning models such as CatBoost, Logistic Regression, Random Forest, and LightGBM were developed to forecast engagement through CTR. A vital part of assessing the performance of these predictive models was selecting the right metrics, such as precision, accuracy, recall, F1-score, and others. Furthermore, it was very important to fine-tune and optimize the models to achieve the best possible result.

The outcomes of this research have the potential to further the present knowledge of digital engagement while providing actionable insights for the optimization of audience engagement in the current cluttered digital landscape. Firstly, temporal features emerged as the most significant predictors of engagement, emphasizing the importance of delivering content at the right time. Headline structure also emerged as an essential characteristic, with word and character count both being vital in feature importance. Finally, emotional traits were found to have a more limited impact, with only a few emotions moderately contributing to performance. These insights offer actionable steps to improve audience engagement in the current cluttered digital landscape.

2. Literature Review

In the digital age, the way people consume media has shifted dramatically from traditional to internet-based platforms. The change in transmission channels has increased the availability of content, which has led to an overload of information. Research by Schmitt, Debbelt, and Schneider (2017) shows that younger audiences and those with lower self-efficacy in information-seeking have been particularly vulnerable to this overload. The constant inflow of real-time content via push notifications and repetitive site visits intensifies the issue, potentially impacting engagement.

2.1 Emotional Content and Engagement

Emotion plays a pivotal role in how individuals engage with digital content. Triggers like joy, fear, and surprise significantly drive audience interaction, especially in terms of attention and sharing. The study performed by Berger and Milkman (2012) found that highly engageable content typically evokes intense emotions, whether positive (like awe or excitement) or negative (such as anger or anxiety). After analyzing over 7,000 New York Times articles, the paper concluded that emotional arousal determines how widely it is shared. Further studies affirm that emotional intensity enhances engagement. Heath, Bell, and Sternberg (2001) explored urban legends, discovering that emotionally evocative stories spread more widely.

2.1.1 Psychological Theories and their application to Online Content Engagement

Several psychological theories provide insights into how emotional content drives online engagement. Affective Events Theory (AET), by Weiss and Cropanzano (1996), suggests that specific stimuli, like headlines, videos, or images, cause emotional responses that influence behavior. In digital settings, these emotional provocations often prompt actions such as clicks, shares, or comments by triggering natural emotional reactions.

Content that stimulates strong emotional responses can ripple through user interactions, amplifying its reach. Universal emotions like fear or joy are particularly effective at creating viral loops.

The Elaboration Likelihood Model explored by Petty and Cacioppo (1986) highlights that emotionally charged content is often processed via the peripheral route, relying on feelings rather than detailed cognitive evaluation. This quick emotional engagement leads to instinctive actions, such as likes and shares, which are critical in today's fleeting attention economy.

By leveraging these theories, content creators can design emotionally resonant articles that capture attention, encourage immediate responses, and spread widely, maximizing engagement in the competitive digital landscape.

2.2 Headlines and Content Structure

Headlines are a deciding factor in engagement and whether an article gets shared. A well-crafted headline has the power to grab attention, stir up emotions, and ignite curiosity, which can ultimately boost CTRs and shares. This chapter discusses how headlines drive interaction, breaking down the structural and emotional elements that encourage clicks.

2.2.1 Impact of Headlines on Engagement

The headline's structure greatly impacts its ability to attract attention. Headlines that elicit strong emotions like joy, surprise, or anger tend to achieve higher CTR and overall engagement. Iarovici and Amel (1989) show that they often favor nominal structures, reducing content to phrases or single words, which enhances their impact and memorability.

Blom and Hansen (2015) highlight the significance of clarity and curiosity in headline design. They found that using elements like surprise and curiosity such as question headlines or unusual information can boost CTR. Moreover, the inclusion of numbers in list headlines signals that the content will be relatively quick to read and easy to digest.

2.2.2 Predictive Elements in Headline Design

Understanding how effective headlines work requires analyzing language patterns and structural features. Studies on clickbait tactics have shown that exaggerated and emotional language strongly influences user engagement. Szabo and Huberman (2010) highlight the predictive power of early user interactions, suggesting that headline structures can determine the future popularity of content.

Headline effectiveness is assessed through sentiment analysis. Stories with a stronger emotional impact tend to engage more readers. Heimbach and Hinz (2016) demonstrate that sentiment and emotional tone significantly influence content virality, with positive or emotionally charged headlines more likely to be widely shared.

2.3 Methodological Approach for Predicting Engagement

Concerning the topic of the predictivity of content engagement, supplementary research was needed to comprehensively understand the possible implications of various factors on modeling efforts.

The Random Forest and XGBoost models play an important role and could be a great fit for this kind of analysis and understanding them will provide a solid foundation for this work. Both models have demonstrated their effectiveness in various fields, particularly in the realm of social media analysis.

Naïve Bayes classifiers are based on Bayes' theorem and have been commonly applied for text classification due to their simplicity and effectiveness. Despite the assumption of feature independence, Naïve Bayes often performs surprisingly well, particularly with smaller datasets. Schütze, Manning and Raghavan (2008) suggest that this probabilistic approach allows for straightforward interpretation and quick computation, making it a go-to model for many baseline tasks.

Moreover, Natural Language Processing (NLP) models have great importance when analyzing and extracting these insights, the prominence of these models can be observed when dealing with the realm of social media. NLPs enable the extraction of emotions from user-posted content.

Additionally, emotion detection frameworks, for example, Ekman's Emotion Model, by Ekman (1992), have enabled the understanding of emotional expressions contained in a text. This kind of method has allowed the overall ability to perform sentiment analyzes and interpret the user's emotional state more effectively.

One other model that should be mentioned in this context is Bidirectional Encoder Representations from Transformers (BERT). It has revolutionized NLP by allowing models to grasp the context of words in a sentence bidirectionally. Devlin et al. (2018) explore how this architecture has led to state-of-the-art performance in a wide variety range of tasks, including text classification, named entity recognition, and question answering.

For example, in the study by Roy and Pan (2021), this model was enhanced with domain-specific medical knowledge to improve its performance in clinical relation extraction. By fine-tuning BERT on a corpus of medical texts and integrating structured medical knowledge, the researchers aimed to accurately identify and classify relationships between clinical entities within the text. The enhanced BERT model demonstrated superior capability in understanding complex medical contexts and extracting relevant clinical information, thereby assisting healthcare professionals in accessing critical patient data more efficiently. This approach not only improved the precision and recall of clinical relation extraction but also highlighted the potential of combining advanced language models with specialized domain knowledge to tackle specific challenges in healthcare.

Hugging Face has become a cornerstone in the NLP community by providing easy access to powerful transformer models, including BERT, as shown by Wolf et al. (2020). Their transformers library offers pre-trained models that can be easily fine-tuned for various applications. Hugging Face's platform has democratized access to advanced NLP, enabling researchers and practitioners to leverage state-of-the-art models without extensive computational resources.

Some research has been able to deploy such a method, for instance, De Choudhury et al. (2013) utilized sentiment analysis to predict depression by analyzing Twitter data. The results showed that there were high correlation levels between the emotional tone of the message and user engagement levels. The importance of NLP can be outlined by these studies that exhibit the crucial information that is provided by the insights into the emotional tone behind online user content.

After deploying the models, it is essential to assess their usability and generalization. To make this possible, some general metrics are used widely across the community. Evidently AI (2024) explains that terms like Accuracy, Precision, and Recall are crucial in many machine-learning problems since they can reveal the performance under unseen data. Accuracy measures the proportion of true results, both true positives and negatives, among the total number of cases examined, if there is a large imbalance between classes, this metric should not be used since it can disguise a low-performing model with a high value of Accuracy. Precision is the ratio of correctly predicted positive observations to the total predicted positives, indicating the accuracy of positive predictions, this metric must be used when the cost of a false positive is low, otherwise, it should be used the Recall since it represents the ratio of correctly predicted positive observations to all the observations in the actual class, showing the model's ability to capture positive instances.

To be able to capture a different view of the classification task success, the F1-score emerges as a balance made between two different measures, precision, and recall. Achieving a higher score in the underrepresented categories means that those classes are being accurately predicted without any major trade-offs, as suggested by V7 Labs (2022).

2.4 Upworthy Dataset

The Upworthy Research Archive appears as one of the study's most interesting sources, evidencing how content goes viral and creates online engagements. The media company that produced this data, Upworthy, has a reputation for capturing and holding onto an audience's attention. The company leveraged A/B testing to systematically access different headlines for its stories, a key element of its approach. The study uses trial versions, and each headline is slightly modified to see which one generates the highest number of clicks and impressions, culminating in the dataset demonstrated in Table 1, by Matias et al. (2021).

Table 1 – Upworthy Data Dictionary

| Variable Name | Explanation |
|----------------------|--|
| created_at | Time of creation (timezone unknown) |
| test_week | Week the test was created |
| clickability_test_id | Unique identifier of the test (The same test has several headlines being tested) |
| impressions | Number of viewers who were assigned to this package |
| headline | Headline being tested |
| eyecatcher_id | Unique identifier of the image associated with the story (image not available) |
| clicks | Number of viewers who clicked |
| excerpt | The excerpt related to every article |
| lede | Opening sentence of the paragraph or of the story |
| slug | Internal name for the web address |
| share_text | Summary for display on social media when the article is shared |
| square | When used, part of the same social media sharing suggesting as share text |
| significance | Inconsistent calculation |
| first_place | Shown to editors, to help guide decisions |
| winner | Indication of if the test was selected to be used after the test |
| updated_at | Last time the package was updated (timezone unknown) |

Research using the Upworthy dataset has provided key aspects of the linguistic predictors of viral content, showing how minor changes in word choice, emotional tone, and headline structure can impact the spread of content through online platforms. For instance, Robertson et al. (2023) demonstrated that negative words in headlines, more specifically expressions that transmit low-arousal emotions like sadness, were more effective at driving

clicks when compared to words that express high-arousal emotions such as anger or fear. For a headline of average length (approximately 15 words), the inclusion of a single negative word could increase the CTR by 2.3%. This highlights a connection between the engagement and the emotional tone of an article.

Furthermore, the influence of headline simplicity was studied by Shulman et al. (2024), which supported the "simpler-is-better" hypothesis, finding that audiences are more likely to click on and engage with simple headlines. By comparing CTR between simple and complex headlines, the study concluded that online readers show a preference for less complicated language.

Similarly, a broader examination of over 32,000 A/B tests was conducted by Gligorić et al. (2023), which aimed to identify the specific linguistic features that contribute to the success of news headlines. The research revealed that headlines containing negative-emotion words like "worst" or "scary" were more likely to generate clicks, whereas positive-emotion words had no significant effect on engagement. Additionally, the study also discovered that longer headlines generally performed better, as they provided more context and information for readers, contradicting the notion that readability alone drives engagement.

2.5 Microsoft Mind Dataset

Microsoft MIND (Microsoft News Dataset) is a large-scale dataset for news recommendation research. This contains anonymized user-article interaction data, such as clicks and impressions. Additionally, each article has some metadata associated with it like its title, abstract, and list of categories as shown in Table 2. This dataset has been widely adopted to study user engagement and preferences for news consumption, allowing researchers to delve into the difference between news-type and recommendation models that affect user behaviors and engagement.

Table 2 - MIND Data Dictionary

| Variable Name | Explanation |
|-------------------|---|
| newsId | Unique identifier of each news article |
| Category | Category of the article |
| SubCategory | Subcategory of the article |
| Title | Title of the article |
| Abstract | Abstract of the article |
| URL | URL of the article |
| Title Entities | Metadata about named entities extracted from the article's title |
| Abstract Entities | Metadata about named entities extracted from the article's abstract |

Even though the current thesis focuses on how engagement is impacted by the emotional tone in headlines, some results given by results under the MIND dataset could have some use. For example, Wu et al. (2020) implemented models that were the attention-based neural model Neural News Recommendation with Multi-Head Self-Attention and News Attention Modeling with Long Short-Term Memory to capture the semantics from both news content and user interests for better recommendation performance. Those models could predict the quality of articles, and along with an NLP model user preference to first suggest an article that will be clicked on. The use of the MIND dataset has broken the barrier to the research for content recommendation and user engagement, benefiting a wide range of recommendation systems.

2.6 Gaps in Current Literature

Although interest in emotional framing and headline generation for predicting online engagement is growing, there are still significant gaps that hinder the development of comprehensive predictive models. These gaps suggest the need for a more expansive, integrated approach combining psychology, linguistics, and data science to better understand content virality and user interaction.

2.6.1 Lack of Comprehensive Models

What is missing from current research under the Upworthy Research Archive is an end-to-end model that successfully combines emotional features with headline structures to predict

engagement using machine learning methods. Although there's extensive literature on both the individual effects of emotional triggers and headline features, these elements are typically studied in isolation, without examining how their interactions influence user behaviors. For example, while many studies highlight that emotions like awe, anger, or excitement boost viral potential, research on headline styles often targets linguistic tactics like clickbait or curiosity hooks to maximize initial engagement. Few models attempt to merge these psychological insights with advanced machine learning techniques to create robust predictive frameworks that consider both emotional and structural factors.

Additionally, many existing predictive models rely on traditional sentiment analysis, which classifies emotions in a basic positive-negative-neutral framework, rather than capturing the nuanced levels of emotional responses linked to engagement. Emotional content is complex, and its intensity, context, and the combination of emotions can significantly influence user reactions to digital media. High-arousal emotions like anger or enthusiasm tend to prompt more immediate and impulsive sharing than low-arousal emotions like sadness or calmness. The absence of models that integrate these subtler distinctions into machine learning algorithms limits the precision of engagement predictions. To bridge this gap, various predictive models were tested to identify one that delivered strong performance.

2.6.2 Limited Research on Upworthy Dataset and Large-Scale Analysis

A significant gap is the relatively limited use of the Upworthy dataset in large-scale, quantitative studies on engagement metrics. The Upworthy dataset is particularly valuable as it includes extensive engagement data related to changes in headline wording, emotional appeal, and content types. However, most studies using this dataset have been anecdotal or focused on specific case studies, lacking the rigorous statistical or machine learning methods

needed for broader conclusions. This represents a missed opportunity to analyze large-scale data to develop generalized models applicable across different digital platforms.

This challenge makes it difficult to develop detailed theories about what kinds of content will be most engaging in the rapidly changing digital media landscape. Engagement behaviors and patterns are fluid, they evolve with new user preferences, platform algorithms, and the broader digital ecosystem. A robust dataset like Upworthy's could be instrumental in tracking these changes over time, offering valuable insights into the shifting trends in emotional appeal and headline effectiveness. To combat this gap in current literature, the focus of this research was always on the Upworthy Research Archive, bringing to light new insights about how to predict engagement in this platform.

2.6.3 Need for Integration of Advanced Emotional Metrics in Machine Learning

While sentiment analysis and emotional recognition have advanced significantly, there remains a gap in integrating these insights into machine learning models to account for nuanced emotional metrics like emotional contagion, sentiment intensity, and multi-dimensional states. Current models often oversimplify emotions, neglecting how combinations or gradients of emotions affect engagement. For example, a joyful headline might prompt sharing, but a mix of joy with surprise or nostalgia could dramatically enhance its viral potential. These complex emotional interactions are rarely addressed by current predictive algorithms, resulting in an overly simplistic view of user behavior.

Emotional contagion, where emotions spread rapidly across social networks, is rarely quantified in machine learning models designed to predict engagement. Although there's theoretical support suggesting that emotionally striking content can quickly influence many people, few practical models integrate this effect into engagement predictions. Developing

models that simulate or predict emotional contagion could be crucial to significantly improving content virality forecasts.

To deal with this gap in current literature, sentiment analysis was performed allowing for a nuanced identification of sentiment and its intensity. This sentiment was then evaluated in terms of its predicted importance, providing deeper insights into the role of sentiment.

3. Data

3.1 Initial Dataset – Upworthy Research Archive

The initial dataset is composed of a total of 150,817 records of 32,487 A/B tests created between January 2013 and April 2015 in the Upworthy platform. The data was obtained directly from the Upworthy Research Archive platform via a CSV file format. The dataset comprises information about the tests, their characteristics, and results in several features that are very relevant to the objective of the present research.

3.1.1 Test-Related Data

Regarding each test the dataset provides a unique *clickability_test_id* that refers to the whole test, this is each unique identifier can have several tests in it. There is also information about the headline, excerpt, opening sentence, the designated web address name and the summary shown on social media when the article is shared, and the *eyecatcher_id* which is the unique identifier of the image used on the test, even though there is no database with said images.

3.1.2 Temporal-Related Data

For each test, temporal information includes when the test was created, the week it was tested, and the date of its most recent update in the Upworthy system.

3.1.3 Results Related Data

To evaluate how effective a test was, the dataset provides several indicators such as the number of *clicks*, *impressions*, *significance*, if the test was the best out of the package, and if that test was selected by editors to be used on the Upworthy site after the test.

3.1.4 Data Cleaning

The data exhibited some inconsistencies, ranging from duplicates to records flagged by the archive that were deemed inadequate for analysis. This procedure commenced with the elimination of records from features that were not useful inputs for the models concerning the scope of future predictions.

Furthermore, with the assistance of the isolation forest algorithm, outliers were removed, as these could negatively affect the model's performance. This model utilizes machine learning to eliminate outliers, presenting itself as a safer option than other methods since it could consider both clicks and impressions when filtering the records. Other alternatives, such as the interquartile range and Z-Score models, were also assessed. These methods, when applied to the data, detected faulty records too abruptly, thereby removing important information from the analysis.

3.2 Additional Data Collection - Microsoft News Dataset

Despite the initial Upworthy Research Archive containing a lot of valuable information, there was one big gap identified. The articles in this dataset are not categorized, something that would be very valuable for this study. To tackle this issue, there was a big process of research for a dataset from a good source that contained a title, abstract, and category. For this matter, the Microsoft News Dataset (MIND) was utilized. As shown by Wu et al. (2020) this is a large-scale dataset intended for news recommendations aimed to build recommendation systems, as this was not the objective of the present research, there was a process of data preparation and cleaning.

To get the data prepped, there was a collection of only the datasets from the Microsoft News Dataset that had information about news, ignoring the ones with data about behaviors from other users. Following this, there was a concatenation of the data from the training, test,

and validation set, which was already divided at the moment of collection, followed by a selection of the relevant features, in this case, *newsId*, *Title*, *Abstract*, and *Category*. Additionally, there was a duplicate removal per Title and Abstract, as the goal is to have the maximum number of unique Titles and Abstracts. After all these processes, the final dataset resulted in 120,148 records without any null value in any of the features.

3.3 Data Description – Upworthy Research Archive

3.3.1 CTR

One feature that was important to be calculated at the start of this chapter was the CTR since it is one of the metrics that shows the level of engagement of a certain article. It was calculated by dividing the values from *clicks* with the values from *impressions*, two features that are going to be studied in this section.

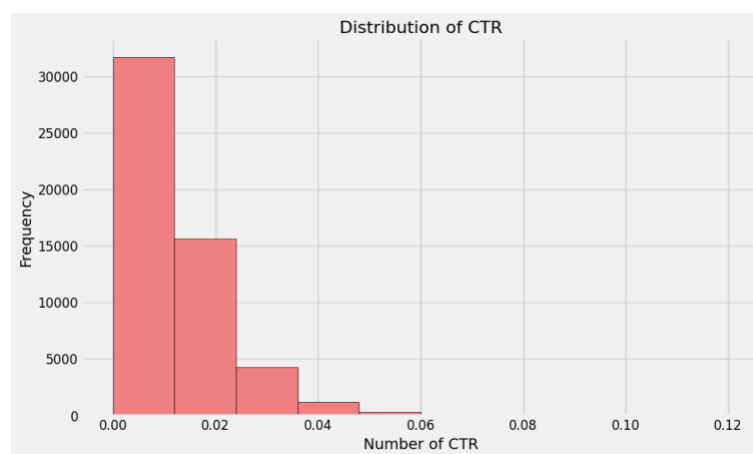


Figure 1 – Distribution of CTR

According to Figure 1, the values are skewed to the left, which creates an expectation of both clicks and impressions having similar distributions.

3.3.1 created_at

This feature provides the date of creation of the post. With this, the impact of things such as the timeliness of the possibility of the post becoming highly engaging. Most of this dataset's posts were made in 2014, and it can be seen a decrease in post frequency during the

weekends as displayed in Figure 2. Furthermore, a decrease can be observed in activity during the summer months. Even though the activity during the weekends seems to be lower, the impact of the *CTR* says otherwise since a light increase can be seen in comparison to the other weekdays, as shown in Figure 3.

3.3.2 updated_at

The plot shown in Figure 4 shows that all the data inserted into this feature was always a Saturday in April 2016. Since this feature adds no value to this work, it will not be used further on.

3.3.3 Headline

For the headline, two new features have been created, those being the *character_count* and the *word_count*. A normal distribution can be observed for the headline count, but the word count has a slight positive skewness, as seen in Figure 5. Furthermore, through the plot presented in Figure 6, it is possible to see a correlation between word count and *CTR*, where the headlines with the number of words between 13 and 16 have a higher tendency to have a higher *CTR*.

3.3.4 Excerpt

For the excerpt, the same features were created. Their distribution is highly imbalanced, as shown in Figure 7, this comes from the fact that the data provided in this feature after the cleaning continued with a vast number of duplicated records since several excerpts were re-used for a great number of different tests, as displayed in Table 3.

3.3.5 Impressions

This feature represents the number of users that have seen the post. The values have a vast discrepancy in values, where they tend to be lower since the distribution is negatively

skewed, similar to the *CTR* distribution. Also, the average impressions per post is 4836.38, but this is largely affected by the range of the minimal and maximal values, as shown in Figure 8.

3.3.6 Clicks

This feature represents the number of users who clicked on the post. The values exhibit a significant discrepancy, tending to be concentrated at lower levels, as the distribution is negatively skewed, similar to the *CTR* distribution. Most posts have fewer than 100 clicks, with a gradual decrease in frequency for higher values. The average number of clicks per post is 59.27, as shown in Figure 9. However, this average is likely affected by the large range between the minimal and maximal values.

3.4 Data Description – Microsoft News Dataset

3.4.1 Category

In the course of the research, the data gathered in the categories feature provided great insights into what were the most prominent types of articles published in Microsoft News. The data consists of 11 different categories, these being news, sports, finance, video, travel, food and drink, lifestyle, weather, health, and autos.

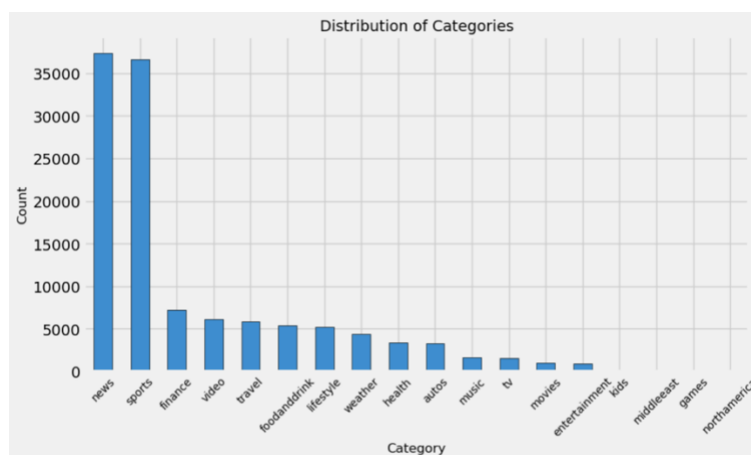


Figure 10: Distribution Categories MIND Dataset

The two categories most present in the data, as shown in Figure 10 are news and sports with 37,377 and 36,631 records, respectively, other categories all found themselves in between 7,253 records from the finance category and 3,334 records from the autos category.

3.4.2 SubCategory

The dataset contains 283 different subcategories, the top 3 being news USA, football NFL, and news politics, with 17,831, 13,317, and 6,352 records respectively. It makes sense that the top three subcategories are in the two major categories of the data, showing that these two topics, news and sports are truly more present in the data. Interestingly of the ten subcategories with more records, nine are from news and sports, with only weather top stories being from another category, as can be observed in Figure 11. The *SubCategory* feature provides a high level of granularity to the data by differentiating news articles per the most specific segment.

3.4.3 Textual Features – Title & Abstract

In terms of textual features, the MIND dataset contains two features *Title* and *Abstract*, both these features are of extreme importance as they enable the creation of machine learning models to predict the category based on text. The average word count for titles is 10.73 and for the abstract is 38.88, which shows that naturally, the title is shorter than the abstract. To better understand the words that are most present in both these features, two-word clouds were built as shown in Figures 12 and 13.

4. Topic Modelling

The Latent Dirichlet Allocation (LDA) was chosen to extract themes and topics from large text datasets. According to Blei, Ng, and Jordan (2003), this unsupervised probabilistic model enables the discovery of themes in unstructured data without needing prior labeling. Its probabilistic approach allows for nuanced interpretations by assigning each document a probability distribution across topics, capturing both primary and secondary themes. Additionally, LDA's scalability enables it to handle large datasets, while its interpretability offers clear, actionable insights into topics that may drive user engagement. These qualities make LDA a fitting choice for this study, supporting a data-driven understanding of user-engaging content.

In addition, the topic modeling process started with text preprocessing to refine the data for analysis. Preprocessing steps included tokenization, which splits text into individual words or tokens, and removing stop words like "the" or "is" that do not add meaningful information. This cleaned and tokenized data was then converted into a bag-of-words format, representing each document by its word count. This format preserved the structure of the text while simplifying it for LDA input. The LDA model then used this representation to group words that frequently appeared together across documents, allowing it to infer potential topics from these clusters. For each document, the model calculated the probability distribution of topics, effectively identifying which topics were most represented within that document. This approach has made it possible to identify dominant topics within the dataset, revealing the primary themes discussed in the text. The output of this typical LDA model is a representation of topics, where each line corresponds to a single topic and lists the most representative words and their weights. For example, *Topic 1: 0.013*"gay" + 0.011*"people" + 0.010*"know" + 0.009*"think" + 0.009*"women"*.

Moreover, the topics identified were “LGBTQ+ and Gender Issues”, “Positive Values and Parenting”, “Media and Creativity”, “Trends and Innovations”, and “Family and Relationships”. The first topic centers around discussions on LGBTQ+ and gender-related issues. Words such as "gay" and "women" imply conversations surrounding sexual orientation and gender. Terms like "people" and "think" suggest social opinions or attitudes toward these subjects.

The second topic revolves around positive values and possibly parenting. Words like "good," "reason," and "kids" indicate a focus on child-rearing, family values, or the promotion of moral reasoning. The term "people" may reflect discussions on societal perspectives regarding these values.

The third topic relates to media and creativity. The presence of words such as "video" and "make" suggests that this topic involves media creation or creative expression, potentially in the context of digital media. The inclusion of "women" may indicate a focus on gender representation in creative fields or women in media.

The fourth topic discusses trends and innovations. Words like "new", and "way" suggest a focus on progressive concepts or emerging trends. The term "people" implies societal impact or acceptance of these new developments.

The final topic centers on family dynamics or relationships. The word "kids" indicates familial contexts, while "one" and "get" suggest narratives around personal experiences or social interactions.

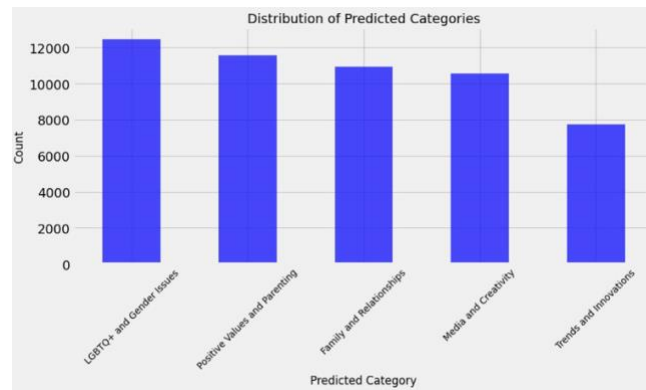


Figure 14: Distribution Categories Topic Modelling

As verified in Figure 14, the frequency of content themes shows that “LGBTQ+ and Gender Issues” was the most frequent category, followed by “Positive Values and Parenting,” “Family and Relationships,” and “Media and Creativity.” Conversely, “Trends and Innovations” had significantly less representation. This distribution highlights the emphasis on socio-relational topics, underscoring the platform's focus on engaging audiences through inclusive and meaningful content.

The LDA approach to topic modeling summarizes the prominent topics found within the data, but according to research made by Wallach, Mimno, and McCallum (2009), it does come with some limitations. For starters, LDA finds topics at a higher level, such as general patterns therefore, it does not line up well with engagement prediction. Finally, it weighs heavily on word frequency and co-occurrence, making it ineffective in picking up more nuanced contextual relationships or high-level user preferences.

Therefore, the LDA-based topic modeling approach was not sufficient for this study since the topics generated were overlapping conceptually, with terms such as “people” and “kids” appearing in multiple categories. Additionally, some topics, such as “Trends and Innovations” or “Media and Creativity” were quite abstract and lacked specificity. These challenges highlighted why transitioning to a supervised approach with predefined categories might turn out to be more advantageous.

5. Categories

One of the main problems identified after deciding to use Upworthy's Research Archive to predict the articles that could have high engagement was the lack of variables present in the initial dataset. Analyzing the MIND data, one of the columns identified as valuable for further study was the *Category* that represented the theme of the news article.

5.1 Data Cleaning

Looking at the number of entries in each of the 16 categories, class labels like "news" or "sports" that have 37,377 and 36,631 instances are significantly larger than classes like "entertainment", "movies", "kids" or "northamerica" with 994, 993, 111 and 1 records. It is clear how imbalanced the classes are, as shown in Figure 10, presenting a risk for the classification task to be biased towards categories with more entries.

To prevent overfitting from happening, some classes suffered transformations. The groups "kids", "northamerica", "games" and "middleeast" have a combined weight of 0.096% of the total dataset, so they have been removed, reducing the number of classes.

A large incongruity has remained present when comparing the registers of some classes to the average entries of all classes, 6,674.89. Therefore, the minority groups "entertainment", "tv", "music" and "movies" which summed have 5,114 records, were merged into one single category, "entertainment", due to the similar content on the *Title* and *Abstract* columns between these topics.

5.2 Data Augmentation

To reinforce the model's category prediction, a Data Augmentation algorithm has been performed. During this phase, the balancing was also performed due to the data discrepancy between categories. The final dataset contained 10,000 records for each category. The Natural

Language Toolkit (NLTK) was utilized extensively for text preprocessing and generating augmented records. Additionally, NLTK's resources, such as WordNet, performed synonym replacement.

There is evidence that supports the use of these methods, for example, Wei and Zou (2019) showed that the use of simple techniques like synonym replacement and random insertion can significantly improve text classification performance.

To reach the final process of data preparation of this chapter, other alternatives were considered besides the data augmentation process utilized. The first was undersampling the train data to achieve a balanced set, this was rapidly discarded since, by utilizing this technique, the amount of data loss would significantly decrease the learning capability and, therefore, the predictive power of the model. Moreover, other techniques, such as oversampling by repeating records in the minority classes in the train sets, did not achieve the same level of results as the synonym replacement technique since duplicating does not necessarily mean an improvement in learning performance, and could lead to an increase in overfitting. Even when reaching the technique of synonym replacement, there was a long process of understanding the right dimension for all classes, with 5,000 and 15,000 records being tested but achieving worse results. The 10,000 emerged as a balanced solution providing enough size for minority classes without an excessive number of artificial records.

5.3 Data Preparation

Using the *train_test_split* from *scikit-learn*, the dataset has been split into training, validation, and testing. The data was distributed 60% into the training set, 20% to the test, and another 20% to the validation set. When splitting into these three sets, a stratification technique was utilized, as reinforced by Farias, Ludermir, and Bastos-Filho (2020), to ensure that the distribution of categories remained consistent. Performing stratified splits among significantly

imbalanced data ensures that a proportion of each class is preserved, preventing any bias that could arise if certain classes are underrepresented or overrepresented in one of the subsets.

5.4 Model Training

After selecting the best features according to the characteristics of the data, the Logistic Regression, Support Vector Machine, Multinomial Naïve Bayes, and Voting Classifier using the Term Frequency - Inverse Document Frequency (TF-IDF) algorithm and Bidirectional Encoder Representations from Transformers (BERT) models have been applied.

The choice of classification method was very important to maximize the results achieved. Apart from BERT, the methods have been used combined with TF-IDF because the algorithm usually outperforms Bag of Words methods with traditional machine learning models for non-binary classification like SVM or Logistic Regression, which is reinforced by the research of Akuma, Lubem, and Adom (2022).

To ensure the input has been truncated or padded to a uniform sequence length, tokenization of the text data was performed using BertTokenizerFast, allowing for the model to handle different text sizes effectively. To simplify the fine-tuning of the transformer-based model, the training was conducted using the Hugging Face Trainer API. The pre-trained approach includes arguments such as batch sizes of 16 for training and 32 for evaluation, 3 epochs to balance the trade-off between performance improvement, overfitting risks, and computational power availability. Moreover, weight decay penalizes larger weights during the training set to 0.01, and the evaluation strategy includes the performance of every epoch, of the one with the best results being saved based on the outcome of the validation set.

The Logistic Regression, Support Vector Machine, Multinomial Naïve Bayes, and Voting Classifier using the TF-IDF algorithm have been used by applying an L2 – Ridge technique to reduce overfitting that arises when applying these models. The C parameter

represents the inverse of the regularization strength. It directly controls the amount of penalty applied to the model's coefficients, so a high C , or low regularization, allows the model to learn more complex patterns by giving less penalty to large coefficients, which can improve performance on the training set but may lead to overfitting while a high regularization encourages smaller coefficient values, leading to a simpler model that helps avoid overfitting but can cause underfitting. The value was set to 0.2, being the best fit between ensuring there was no significant overfitting and maintaining the learning capabilities of the models.

5.5 Model Evaluation

The performance and robustness of various classification models have been assessed by analyzing accuracy, precision, recall, F1-score, and loss. The evaluation methods were chosen according to the type of models applied and the data specifications.

Naïve Bayes revealed itself to be the worst model of all, underperforming in every metric, it showed that it does not align well with the data complexity, with all metrics aligned at the 72% mark. Among the rest of the traditional models, the Voting Classifier slightly edged out the rest of the models, showing that cooperation improved results, as metrics reached the 76% and 77% mark. The Voting Classifier demonstrates itself to be a good solution that balances simplicity and performance. The BERT classifier outperformed every other model, as shown in Table 4, with all metrics reaching values of 86%, an improvement of more than 10 percentage points computational power availability, relative to the best traditional model.

In conclusion, as the previous analysis shows, BERT applied to resampled data was the best option to follow. This decision has been made due to the lack of trade-off consequences compared to the other models it has presented between loss and accuracy in the test set and between recall and precision in effectiveness to minimize false positives.

5.6 Model Deployment

The results reflect the performance of the model when predicting outcomes in a scenario where minimizing false positives is of higher importance. As it can be seen in the new category's column *predicted_category*, just like the MIND non-resampled data, the Upworthy Research Archive content distribution is highly imbalanced. Despite also being skewed, the distribution presented by each dataset is extremely different, as can be seen in Figure 15.

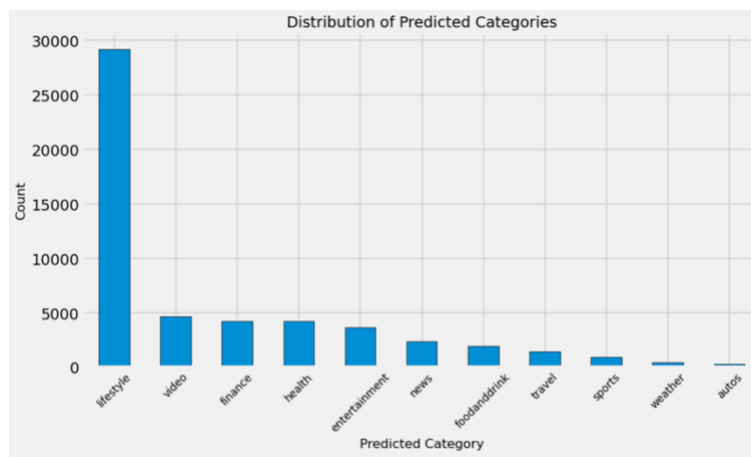


Figure 15 - Distribution Categories Upworthy Dataset

Although Figure 15 shows a large disparity between classes, the representation of each group being different indicates a good performance of the model due to how well adapted to new data, not replicating the distribution pattern seen in the data that has been used to train it.

A large portion of the data is represented by the “lifestyle” group, which portrays 54.81% of the dataset. Furthermore, the classes “weather” and “autos” had a representation below 1%, while all the other subsets represented between 1% and 10% of all the records.

This output can be further analyzed or used in downstream tasks, such as providing insights into content categorization, use as a feature for prediction, help in sentiment analysis, and in several other tasks within the scope of the research.

6. NLP

This chapter delves into sentiment analysis and how emotion intensity influences user engagement, providing valuable insights to guide strategies aimed at improving engagement metrics.

6.1 Sentiment Analysis

To conduct a thorough sentiment analysis of the headlines within the dataset, a multi-step approach was adopted, surrounding data loading, preprocessing, and the application of sentiment analysis through the Valence Aware Dictionary and Sentiment Reasoner (VADER) Sentiment Analyzer. The initial step involves loading the dataset, which is stored in a CSV format (cleaned before), into Pandas *dataframe*.

Before analyzing sentiment, the dataset undergoes a preprocessing phase. The focus of this preprocessing was on the *headline* column, which contains the text of interest. This phase includes removing newline characters that may exist within the text. Additionally, any unnecessary whitespace (extra spaces at the beginning or end of the text) has been stripped to standardize the format of each headline. It was needed to clean up the data to ensure the text was in its most optimal state for accurate sentiment analysis, removing any noise or artifacts that could distort results. Sentiment classification was performed using the VADER sentiment analyzer. It is a lexicon and rule-based tool tailored for analyzing social media and text content. It assigns a compound score to each text input based on the polarity of the words used. This compound score is an aggregated measure of the text's overall sentiment, making it suitable for classifying the sentiment of each headline.

The score that was outputted, on the other hand, provides different evaluations of the content's overall tone and attitude based on the parameters specified by this score. When an average of 0.05 or higher is given, the headline would be deemed positive. Thus, the compound

scores of headlines are grouped into two categories, namely the “negative” and the “neutral” categories, and are classified further into three ranges, less than -0.05, between -0.05 and 0.05, and more than 0.05. Through these intervals, the polarity of the text can be translated to the standard scale employed in sentiment analysis.

Once VADER had been configured, it was applied to each headline within the dataset. For each headline, the compound score was calculated and based on the previously defined thresholds, assigned a sentiment category. This sentiment classification, positive, neutral, or negative, is then stored as a new column. After this process, an additional sentiment column was included that stores the sentiment classification for each headline. This classification enables granular sentiment analysis, allowing cross-referencing with other engagement metrics to understand the relationship between sentiment and user interaction with the content.

6.2 Emotion Intensity Analysis

The emotion intensity analysis uses the NRC Emotion Lexicon, a lexicon that categorizes words based on emotional associations, such as "joy", "anger", and "sadness". According to Mohammad and Turney (2013), the NRC Emotion Lexicon was developed as part of their work on lexical resources for sentiment analysis, specifically aimed at capturing fine-grained emotions in text, which is especially useful in contexts like news, which is the target of this chapter.

To apply the lexicon, a function called *load_nrc_lexicon* was created to read a tab-delimited text file that links each word to specific emotions with a score of either 1 (indicating a strong association) or 0 (indicating no association). Only words with a score of 1 have been included, ensuring that each word fully represents its assigned emotion, which enhances the precision of emotional scoring. This function also included error handling to alert users if the lexicon file is missing, preventing potential disruptions.

Once the text had been cleaned, the emotional intensity of each headline was calculated using the *analyze_emotions* function. This method tokenizes each headline into individual words and then checks each word against the NRC Emotion Lexicon to determine its associated emotions. If a word appears in the lexicon, its emotional scores are recorded and added to an overall *emotion_scores* dictionary, which keeps a running total for each emotion represented in the text. This aggregation provides a view of the intensity and variety of emotions present in each headline.

In addition, the emotional intensity analysis is then applied to each headline in the dataset, creating a new column in the *dataframe* called *emotion_scores*. This column contains a dictionary of emotional scores for each headline, capturing the emotional landscape of the text data. The final output of this process was the *emotion_scores* column, which presented the emotional intensity scores in a dictionary format. This enabled a detailed analysis of the correlation between emotional intensity and engagement metrics, allowing for a more granular investigation into how strong positive emotions are.

6.3 Comprehensive analysis of NLP techniques results

6.3.1 Statistical overview

This chapter provides a statistical perspective of the data after applying those methods. The sentiment analysis revealed a distribution of sentiments, with “positive” sentiment occurring 22,692 times, “negative” sentiment 17,672 times, and “neutral” sentiment 12,882 times. This indicates that a significant portion of the content analyzed generates positive sentiments. The standard deviation of emotion counts is 10,231, suggesting variability in how different emotions resonate emotionally with users. The analysis revealed a diverse range of emotional expressions across the dataset, with the most frequent emotions being “positive” (43,795 occurrences) and “trust” (28,018 occurrences). “Anticipation” (21,620) and “fear”

(20,156) were also prominent, indicating that both hopeful and cautionary tones are common. Negative emotions, including “anger” (15,356), “sadness” (15,093), and “disgust” (10,703), as well as “surprise” (9,935), were observed in significant numbers, suggesting a balanced spectrum of emotional appeals. Notably, the “negative” category had a total count of 29,928, underscoring the impact of more intense or complex emotional tones. Note that these emotional intensity categories can have scores bigger than one in each headline (to see the predominant one).

6.3.2 Emotions correlation

To better understand the relations between the different emotional features, a correlation matrix was performed, as demonstrated in Figure 16.

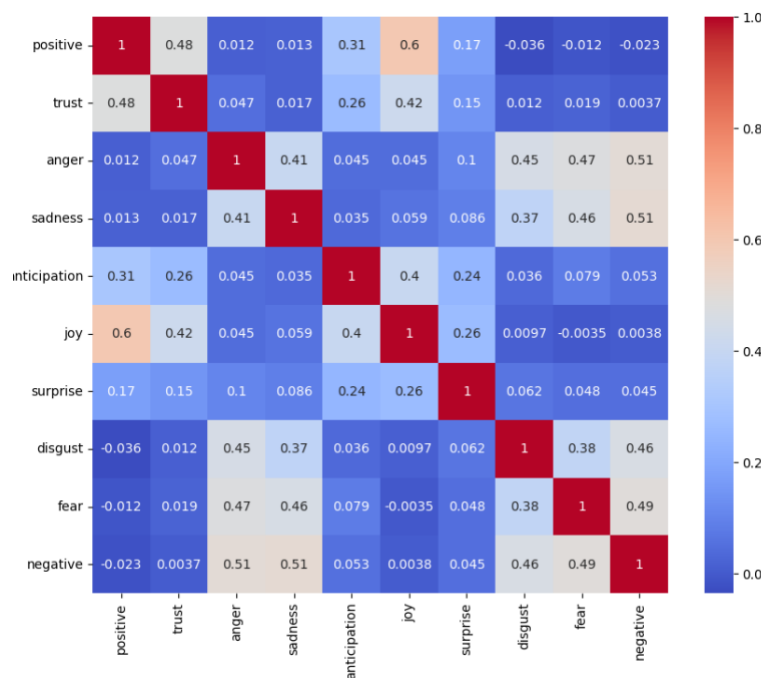


Figure 16 – Correlation Matrix between Emotions

As shown in Figure 16, the relationships between emotional expressions have been captured with some meaningful information regarding them from correlation as depicted. “Positive” sentiment is correlated with “joy” (0.597) which could mean content that is tagged as positive tends to be joyful. So, positive and hopeful content generally elicits a happy

response from the audience. Furthermore, the correlation between “positive” sentiment and “trust” (0.482) was a moderate positive one: meaning that content with a positive tone can foster some degree of building up trust which may be important to engagement and credibility in content as well. “Negative” emotions are correlated with one another, especially with “anger”, “disgust”, and “fear”. There are strong correlations of “anger” with “disgust” (0.454), and “fear” (0.474), while “sadness” also correlates reasonably well with “disgust” (0.373) and, at about the same level as “anger” and “fear” (0.457). Such co-occurrences imply that negative emotions are correlated negatively or positively to each other, and hence, the presence of a pair together as content may lead to an additive or synergistic emotional response. In addition, further confirming the clustering of negative emotional responses is that “negative” sentiment shows very strong correlations with “sadness” (0.511), “fear” (0.491), and “anger” (0.505). However, the moderately high correlation of anticipation to “joy” (0.404) implies that content that sparks anticipation can elicit joy as well and could suggest that themes around future outcomes or possible experiences are positively engaging audiences. On the other hand, “positive” sentiment has almost no correlation with “sadness” (0.013) and “anger” (0.012), just a subtle negative correlation was found for “disgust” (-0.036), “fear” (-0.012), and total “negative” sentiment (-0.023). These low correlations indicate that positive emotions are relatively dissimilar to negative emotions, consistent with the observation that positively toned content is less likely to elicit negative reactions. In conclusion, the correlation analysis provides good insights into emotions’ inter-correlation in contents. It unveils both positive and negative affect clusters, some of which (such as trust and anticipation) have been charted for driving engagement.

6.3.3 Influence of emotions on engagement

The analysis of engagement metrics, in this case, it is CTR due to being the most comprehensive understanding metric of engagement and recording various behaviors, aligns

with the findings of Yang, Zheng, and Xiao (2021), who emphasize that improving click-through rates can not only enhance user experience but also boost advertiser outcomes. As shown in Table 5, positive and negative emotions show the highest average CTRs, at 0.0087 and 0.0070, respectively. Emotions such as trust and anticipation also display meaningful influence, highlighting that both positive and trust-based emotions drive significant engagement. This aligns with the idea that audiences are drawn to content that fosters a sense of trust or anticipation, in addition to emotionally charged or value-based content. Emotions like fear, sadness, and anger appear less frequently but are associated with moderate CTRs, indicating they, too, play a role in drawing attention, albeit to a lesser extent. Lower CTRs are observed for disgust and surprise, suggesting these emotions may have a more limited impact on engagement.

7. Prediction of Engagement

7.1 Definition of Target Feature

The objective of the current study is to predict engagement with machine learning models informed by the Upworthy Research Archive, which raises the question, what is engagement in the context of this dataset? Another goal of this research was to create as much impact as possible, so the main goal of this thesis was to predict high-engagement pieces of digital content, as it proves itself to be a more challenging but rewarding task, expected to present itself with problems such as data imbalance and others.

To define high engagers, the feature click-through rate has been utilized as it is the only feature that balances the two metrics of performance in the dataset: clicks and impressions. The importance of this feature cannot be undermined, so high engagement is defined by the following formula:

$$(1) \quad \textit{High Engagement} = \textit{Top 20\% CTR}$$

This feature has allowed this study to create the utmost impact, being able to now predict which pieces of digital content will achieve the most.

7.2 Data Preparation

To deal with the problem of high data imbalance, coming from the fact that the target feature has a ratio of 0.2 between high engagement records to low engagement records, the technique of post-hoc analysis has been adopted, as seen in works such as Sabia et al. (2014), this means that, the evaluation of the machine learning models employed has only been made after the first testing. This allowed the reformulation of the target variable into four bins, solving the issue of data imbalance, as shown in Figures 17 and 18. The thresholds for segmentation have been determined using specific CTR deciles: the minimum value, the 5th decile (median), the 8th decile, and the maximum value in the dataset, with the categories being low, medium, high, and very high, respectively.

Additionally, the feature *predicted_category* has been converted into dummy variables to be ready for modeling, and the target feature, *engagement*, has been encoded with values from 0 to 3. From the column, *created_at* derived four new features, the *created_year*, *month*, *weekday*, and *hour*, as these could prove to be relevant later, and the features *created_at* and *updated_at* were removed. From the columns, *excerpt*, and *headline* derived five new features, these being *word_count* and *character_count* for both textual features and *punctuation_count* for the *headline* following the logic found by Gligorić et al. (2023), the two original textual features were also removed. The numeric features have been standardized, and irrelevant ones have been removed, leading to the final dataset containing the features present in Table 6.

7.3 Modelling

To start the modeling process, the first step has been to divide the data into training and test sets, with a ratio of 70% and 30%, respectively, to allow the evaluation of the model. Afterward, undersampling was employed in the training set to allow for a balanced dataset throughout all four classes, resulting in four classes, each with 7,439 records. This allowed the training to occur in a balanced environment while being tested in the original disproportions of the dataset.

7.3.1 How to Evaluate and Measure Performance

To ensure that the predictive models align with the present research's objective, the right choice of evaluation metric was of vital importance. In the specific task of predicting the most engaging headlines, recall emerges as the most important measure, as the opportunity cost of missing high-engagement pieces of digital content could mean lost opportunities for visibility, revenue, or strategic gains that might come from a higher marketing bet on these contents, this makes it essential to minimize false negatives, which recall measures explicitly. This happens in Foody (2023), where identifying the positive class is more valuable than reducing false positives, as shown in the following formula:

$$(2) \quad \text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

Other metrics, such as accuracy, precision, and F1-score, are also crucial for simultaneously assessing the equilibrium of the models when predicting both classes.

7.3.2 Model Evaluation

The choice of what classification model to use was of the utmost importance to maximize the results. The first employed model was the logistic regression as it poses itself as a great option as a baseline model thanks to its simplicity and interpretability, as shown by Hardt, Hovy, and Lamprinidis (2018). Unfortunately, this model has not achieved the best results, having high precision and recall for the other class (non-high engaging class) and an overall accuracy of 75.36%. These metrics are misleading since the model reveals a poor performance to the high engagement records, with low recall 26% and low precision 35% and other metrics as shown in Table 7, the significant class imbalance of the test set has created a bias towards the majority class. While the baseline model provided a straightforward analysis, its performance was revealed to be limited, to address these shortcomings, more complex models were employed.

To improve upon the previous results, three methods were applied: the Random Forest as seen by Ali et al. (2023), LightGBM, as shown by Fathima, Devi, and Faizaanuddin (2023), and CatBoost as demonstrated in Jhaveri et al. (2019).

The Random Forest model has been implemented using the *scikit-learn* library, and the key hyperparameters were chosen and selected to control model complexity to prevent overfitting, promote diversity among the different trees for better generalization to the test set, handle class imbalance, and ensure computational efficiency without compromising the performance this model can achieve. These hyperparameters can be analyzed in Table 8.

The LightGBM model has been implemented using the *lightgbm* package, the key hyperparameters were carefully chosen to optimize for multi-class classification, prevent

overfitting through constraints on tree size, and balance model convergence speed and accuracy with a moderate learning rate. These hyperparameters can be analyzed in Table 9.

The CatBoost model has been implemented and configured to perform multi-classification, with the accuracy metric set as the evaluation metric for the model and with a random seed defined to ensure the reproducibility of results. To take the most out of this powerful model, a grid search has been employed to fine-tune the model, resulting in the parameters shown in Table 10. All three models have been improved from the baseline model, LightGBM achieved the highest precision for class 3 of all models with a value of 57%, meaning it makes fewer false positive predictions for the high-engagement pieces of digital content. Simultaneously, this model has the lowest recall for class 3, which ultimately makes this machine learning model not ideal for the objective of the present study. On the other hand, the random forest has been able to achieve the highest recall for class 3 of all models, with a value of 71%. However, precision for class 3 is only 46%, indicating a lot of false positives, this model achieves a great recall for class 3 but lacks balance throughout all metrics. CatBoost has been the best model, with the highest F1-score for class 3 with a value of 59%, showing it is the most balanced model, recall for class 3 also emerges as a strong characteristic of this model with a value of 67%. Overall accuracy and the macro F1-score are also great, with values of 81% and 73%, respectively. The performance of the model can also be evaluated by the confusion matrix of the predictions, where the CatBoost showed a reasonably good performance, as shown in Figure 19. All three models have their strengths, however, looking into the objective of this research, CatBoost emerges as the best model, as analyzed previously in the highlighted metrics and as shown in Table 11.

7.3.3 Feature Importance

To better understand which features have the highest impact on the performance of the best model, a feature importance analysis has been conducted, as shown in Figure 20. The

standout features were the temporal ones: *testweek*, *created_hour*, and *created_month*, suggesting that timing plays a pivotal role in the performance outcome. *Excerpt* and *headline* metrics, such as character and word count, are among the moderately important features, indicating that length and structure also have a noticeable impact. Simultaneously, emotional traits such as *trust*, *positive*, and *negative* contribute to predictions but not as significantly as timing or content structure. Among the least important features are the predicted categories and low-impact emotions such as *disgust*, *surprise*, and *anger*, revealing that their inclusion is less impactful for these specific models.

8. Additional Research

8.1 Clickbait Detection and Its Implications Using a Machine Learning Approach Applied to the Upworthy Archive

According to the Cambridge Dictionary, clickbait is "an internet story, title, image, etc., that is intended to attract attention and encourage people to click on a link." Cambridge Dictionary (n.d.). This suggests that when used effectively and targeted to the appropriate audience, clickbait can serve as a valuable tool for several platforms, including Upworthy, which is renowned for mastering this technique, as stated in Thompson (2013). This research focus was on identifying and analyzing the impact of clickbait on the Upworthy platform, thus answering the research questions: How can predictive models be developed and applied to accurately detect clickbait in the Upworthy Research Archive? What is the impact of clickbait engagement metrics, and how can these insights inform content strategies?

Clickbait detection is a fascinating field, with a variety of approaches and techniques used to tackle this challenge, Agrawal (2016) showed how the use of deep learning methods can achieve impressive results. Simultaneously, Chen, Conroy, and Rubin (2015) demonstrated the complexity and diversity of clickbait cues. Furthermore, Zuhroh and Rakhmawati (2020) provided an excellent review of the wide range of predictive models, including support vector machines, random forests, and other classifiers. The authors highlighted the upsides and downsides of each technique, offering a view of their applicability. In this research, these techniques were used to find the most simple and effective solution to identify clickbait in the Upworthy Research Archive.

In this study, clickbait detection began with additional data collection that included headlines and a binary indicator of clickbait. This process involved cleaning, preparing, and building new features such as *sentiment_polarity*, *readability_score*, and others, followed by applying several machine learning models, such as Naïve Bayes, Support Vector Machine, and

Logistic Regression. The top three were comprehensively evaluated using traditional metrics and assessed for their predictions on the Upworthy dataset. To ensure alignment with the Upworthy Research Archive, a spot-check was performed on the chosen model using a manually annotated sample of 200 records named ground truth, for comparison. This whole process achieved high metrics and allowed the identification of clickbait in the Upworthy Research Archive.

The analysis of clickbait, on its own, is a fantastic field where many great studies have delved into the consequences of using this technique. According to Mukherjee, Dutta, and De Bruyn (2022), the use of this technique can often be seen by the reader as a manipulative technique leading to resistance to engage in future posts, harming performance in the long term. Other studies, such as Zannettou et al. (2018) delved into how clickbait is a problem, like spam, and harms users. The present study aimed to understand how the Upworthy audience reacted to clickbait and developing strategies to use it more effectively.

The second part of this research focused on a comprehensive analysis of clickbait and its impact on the performance of headlines to generate actionable insights for digital content platforms. The process to reach this included statistical methods, hypothesis testing, and an evaluation of records, among other techniques. These approaches culminated in well-justified conclusions that provided meaningful insights to drive impactful content strategies, such as applying clickbait to specific categories, creating a hybrid with characteristics of clickbait and non-clickbait headlines, and monitoring *CTR* performance over time to better adapt to the natural changes in the audience.

Assessing the prevalence of clickbait in the Upworthy Research Archive and its effect on engagement was successfully achieved. It uncovered significant new insights such as the extreme use of this technique by the Upworthy platform, the way different categories respond to clickbait, and how to leverage clickbait to improve metrics in a sustainable way using the

audience's response as a guideline. These conclusions can enhance the application of these techniques following the objectives of the platforms and similar publishers and reaching the goal of the present research to create an impact.

8.1.1 Identifying Clickbait

To identify clickbait in the Upworthy Research Archive dataset, several methods can be utilized, as observed in significant works by Chen, Conroy, and Rubin (2015) and Pujahari and Sisodia (2019). As a first step, it was essential to gather external data that included headlines and a binary indicator assessing the presence or absence of clickbait, to then build several predictive models until acceptable metrics were achieved. Once these steps were completed, the best model was applied to the Upworthy Research Archive, paving the way for new analysis and insights into this intriguing topic.

8.1.1.1 Collection of Data

The external clickbait data collected was from Kaggle, an online community of data practitioners and enthusiast. This dataset was also used in significant works such as Bhoj et al. (2021), it contains 32,000 records collected from various news sources. The clickbait headlines originate from websites such as BuzzFeed, Upworthy, ViralNova, ThatScoop, Scoopwhoop, and ViralStories, while the non-clickbait headlines derive from WikiNews, The New York Times, The Guardian, and The Hindu, as shown by Anand (n.d.). The distribution of both classes is nearly perfectly balanced, with 15,999 records being clickbait and 16,001 records being non-clickbait. There is no missing data or repeated titles, and although some of the clickbait headlines come from the Upworthy platform, there is no overlap in headlines between this dataset and the Upworthy Research Archive dataset.

8.1.1.2 Feature Engineering

Before conducting statistical analysis or predictive modeling, it was essential to create meaningful features that capture the variability of the target feature, thus having higher

explanatory power. Feature engineering played a vital role in improving model performance by providing additional information for the algorithms to learn from. The initial set of features created included *word_count*, *character_count*, *punctuation_count*, binary features indicating the presence of punctuation (*has_punctuation*), question (*has_question*), and exclamation marks (*has_exclamation*). These features drew inspiration from Li, Goh, and Jin (2018) that highlighted the importance of lexical and structural analysis in understanding textual data. To add more value to this preparation, the level of readability of a headline was also implemented, using the Flesch-Kincaid Grade Level, a readability test designed to understand how difficult a passage in English was to understand, as seen in Rooein et al. (2024), and it follows the following formula:

$$(3) \quad FKGL = 0.39 \left(\frac{\text{Total Words}}{\text{Total Sentences}} \right) + 11.8 \left(\frac{\text{Total Syllables}}{\text{Total Words}} \right) - 15.59$$

Other features, such as text vectorization, were implemented using a term frequency-inverse document frequency (TF-IDF) approach as inspired by Liu et al. (2018). As a last indicator, the sentiment score was calculated using the VADER sentiment technique, described in Hutto and Gilbert (2014), enabling the creation of the *sentiment_polarity* feature that indicates whether a headline has a negative, neutral, or positive sentiment.

8.1.1.3 Correlation and Multicollinearity Analysis

To validate the features and refine the final dataset for modeling, it was essential to assess correlations, address multicollinearity, ensure statistical significance, and ensure a solid dataset for prediction. The first correlation analysis, as shown in Figure 21, demonstrated that many features such as *word_count* (0.34), *readability_score* (-0.25), *punctuation_count* (-0.27), and *sentiment_negative* (-0.21), were the most correlated with the target, clickbait, suggesting that headlines that had more words, were easier to read, had fewer punctuation marks and less negativity were more associated with clickbait. Unfortunately, the features *word_count* and *character_count* exhibited a strong positive correlation (0.86), as expected,

since longer headlines naturally have more characters. This could indicate redundancy between these two features, which might lead to multicollinearity, to solve this, the features were combined into a new one, *avg_characters_per_word*, which was formulated by the division of the character count by the word count. Additionally, the sentiment features were merged into a single categorical one *sentiment_polarity*, reducing dimensionality while keeping relevant information about the emotional charge of a headline.

This new feature, *avg_characters_per_word*, exhibited a moderate negative correlation with the dependent feature (-0.40). However, it revealed signs of multicollinearity with the *readability_category*, with a strong positive correlation (0.75). Multicollinearity was confirmed by the variance inflation factor (VIF) as used in Salmerón, García, and García (2020), both features revealed values over 5, to solve this, the *readability_category* was removed. Finally, a new correlation analysis was performed, as shown in Figure 22:

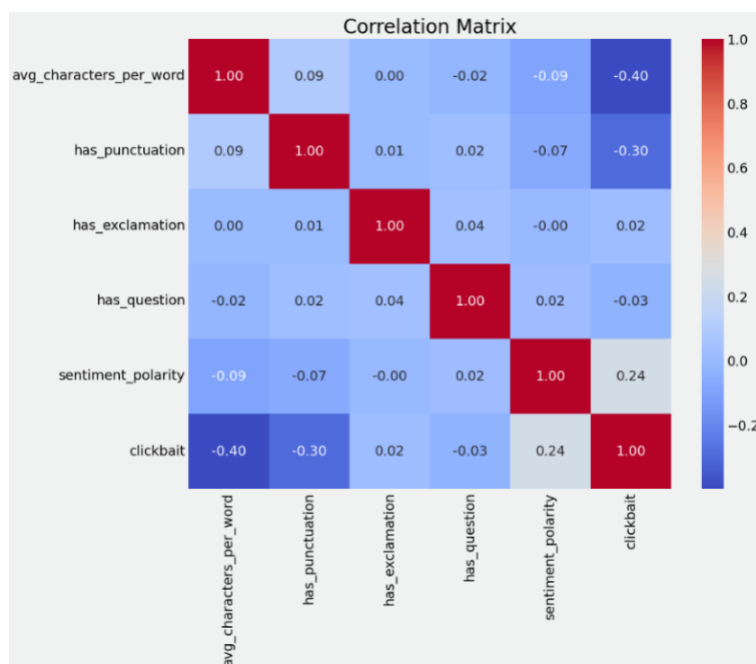


Figure 22: Final Dataset Correlation Matrix

The final dataset correlation matrix included five features, excluding the target variable. Of these, three exhibited a moderate correlation with the clickbait variable, these were *avg_characters_per_word*, *has_punctuation*, and *sentiment_polarity*, suggesting that clickbait headlines tended to use simpler and shorter words, avoided punctuation marks, and had a

positive tone. Simultaneously, there were no signs of multicollinearity, to ensure this, a follow-up VIF test was done, and no feature demonstrated multicollinearity, as shown in Table 12.

8.1.1.4 Data Description

The dataset comprised 4,103 features, with 4,096 representing the vectorized text, as shown in Table 13. The analysis highlighted the balanced distribution between clickbait and non-clickbait headlines, as shown in Figure 23, and a predominance of neutral headlines, as shown in Figure 24. The average number of characters per word was around 5-6 characters, suggesting that a neutral and simple tone was common. In terms of punctuation use, it was minimal, as only 0.5% of headlines contained a question mark, 0.1% an exclamation mark, and 18.8% had any punctuation at all, as displayed in Table 14. Collectively, these features provided a solid basis for the statistical analysis and modeling.

8.1.1.5 Statistical and Regression Analysis

To ensure all the five independent features, excluding the vectorized text, were significant predictors of clickbait and were worth retaining in the dataset, it was performed a t-test on the only numerical feature, *avg_characters_per_word*, and chi-squared tests on the categorical ones, *has_punctuation*, *has_exclamation*, *has_question*, and *sentiment_polarity*, following the approach outlined by Ratul et al. (2022). The tests produced the results shown in Table 15.

Table 15: Statistical Test Results for Numerical and Categorical Features

| Tests | Features | p-value | Significance |
|-------------|--------------------------------|---------|--------------------------------|
| T-Test | <i>avg_characters_per_word</i> | 0.0000 | ***($p < 0.01$; significant) |
| Chi-Squared | <i>has_punctuation</i> | 0.0000 | ***($p < 0.01$; significant) |
| Chi-Squared | <i>has_exclamation</i> | 0.0007 | ***($p < 0.01$; significant) |
| Chi-Squared | <i>has_question</i> | 0.0000 | ***($p < 0.01$; significant) |
| Chi-Squared | <i>sentiment_polarity</i> | 0.0000 | ***($p < 0.01$; significant) |

Note:

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Table 15 results demonstrate that all tested features were statistically significant predictors of the target variable, clickbait. Both tests show significant associations, with p-values lower

than 0.01, indicating that the tested features are strongly linked to clickbait classification, confirming the relevance in identifying clickbait on Upworthy’s headlines. To understand how each independent feature impacted the dependent one, it was performed a regression analysis, as shown in Table 16. While the explained variance is moderate, 22.38%, the model provides valuable insights into how the features influence the outcome.

Table 16: Logistic Regression Coefficients for Clickbait Prediction

| Logistic Regressions | |
|--------------------------------------|---|
| Dependent variable: <i>clickbait</i> | |
| constant | 6.0045*** (0.103) |
| <i>sentiment_polarity</i> | 0.6585*** (0.017) |
| <i>avg_characters_per_word</i> | -1.0616*** (0.017) |
| <i>has_punctuation</i> | -1.7784*** (0.038) |
| <i>has_exclamation</i> | 2.3323*** (0.457) |
| <i>has_question</i> | -1.5714*** (0.201) |
| Observations: | 32,000 |
| Note: | *p<0.1; **p<0.05; ***p<0.01 Standard errors in parentheses |

Considering the results, the corresponding model equation was defined as follows:

$$\text{logit}(P_{\text{clickbait}}) = 6.0045 + 0.6585 \times \textit{sentiment_polarity} + (-1.0616) \times \textit{avg_characters_per_word} + (-1.7784) \times \textit{has_punctuation} + 2.3323 \times \textit{has_exclamation} + (-1.5714) \times \textit{has_question} \quad (4)$$

Overall, the findings indicated that all predictors tested in the regression analysis were statistically significant, with a p-value lower than 0.01, with *sentiment_polarity* and *has_exclamation* positively influencing the likelihood of clickbait classification, while *avg_characters_per_word*, *has_punctuation* and *has_question* decreased it.

8.1.1.6 Modeling and Evaluation

The first step in the modeling process involved splitting the dataset into three subsets: 60% for training, 20% for validation, and 20% for testing. The second step before prediction

was to scale numerical features, in this case, *avg_characters_per_word*, using a Min-Max scaler. This approach was adopted based on the rationale provided in Ozsahin et al. (2022), which highlighted the importance of feature scaling in various machine learning models. Three different predictive models were applied these were Logistic Regression, Multinomial Naïve Bayes, and Support Vector Machine. The choice of these specific models came from the understanding of prior research, by Cao, Le, and Zhang (2017), Potthast et al. (2016) and Chakraborty et al. (2016) that suggested these machine learning algorithms deal well with textual data for classification. The goal was to get the best possible metrics with minimal dimensionality possible. To this end, the models were trained and evaluated in three different sets of data, the first being all features, the second one only the ones with a moderate correlation to the target feature and vectorized text, and the third with only the vectorized text features.

Thankfully, most models obtained great metrics, as shown in Table 17, the three best ones being the SVM with all the features, the Logistic Regression with all features, and the SVM with only the vectorized text. The Logistic Regression with all features posed itself as an excellent solution to this problem, it had a great performance, slightly worse than the other two, however, it was simpler, more interpretable, and robust. The SVM with all the features had the best metrics out of all the predictive models, however, that came at the price of higher dimensionality, and the marginal gain in metrics came with the cost of higher complexity. The SVM with only the vectorized text achieved the best loss and great metrics across all other evaluation categories, it was less interpretable than the Logistic Regression but generalized better and had a lower complexity than the other SVM model.

Following this conclusion, the model was applied to the Upworthy Research Archive, where it classified 9.2% of records as non-clickbait and 90.8% as clickbait. To assess the ability of the model to generalize, the predictions were spot-checked against a manually curated dataset of 200 records, with 160 records identified as clickbait and 40 as non-clickbait. The

model achieved great metrics, with all macro averages above 80%, unfortunately, class 0 (non-clickbait) exhibits a worse performance, but still, a very good one for such a minority class in the Upworthy population overall, the model had a great performance, as shown in Table 18. The confusion matrix showed the preference of the model on prediction class 1 (clickbait), it also displayed the great performance of the model, as shown in Figure 25.

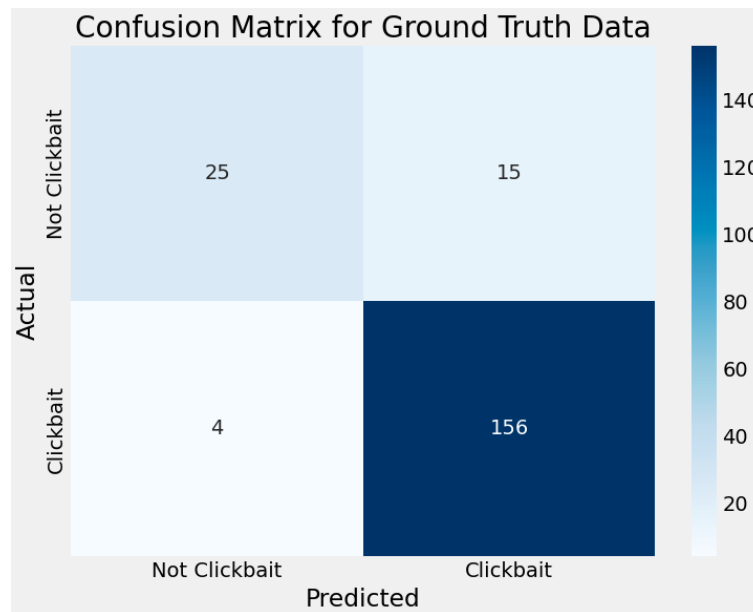


Figure 25: Confusion Matrix for Ground Truth Data

Overall, the clickbait identification was a successful and crucial first part of the research that enabled the comprehensive analysis of clickbait in the Upworthy Research Archive.

8.1.2 Clickbait Analysis

To start, it was of the utmost importance to understand the ratio between clickbait and non-clickbait records to better grasp the scale of each class. The ratio was 9.8 clickbait to not clickbait headlines, meaning that, as expected, the Upworthy platform and the data retrieved from the Upworthy Research Archive used this technique in abundance.

One of the current research questions of the present research is how clickbait impacted performance metrics and how can the insights from this impact be transformed into actionable, informative insights. To measure if this feature impacted performance, a T-test for *CTR*

differences was performed, hoping to find a statistically significant result and throw out the null hypothesis. Thankfully, the test provided statistical evidence, as shown in Table 19.

Table 19: Statistical Test Results for Clickbait

| Tests | Features | p-value | Significance | T-Statistic |
|--------|-----------|---------|--------------------------------|-------------|
| T-Test | clickbait | 0.0001 | ***($p < 0.01$; significant) | 3.762 |

Note: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

This test revealed a statically significant difference in *CTR* between clickbait and non-clickbait content, and the positive t-stat indicates that clickbait content likely has a higher mean *CTR* than non-clickbait content. This was expected as many articles suggested that clickbait can improve digital content performance, like Jung et al. (2022). To understand if this insight that clickbait improves performance in the short-term, harming long-term goals, an analysis was performed where the *CTR* mean was calculated per week of creation throughout the whole dataset, and separated by classes of clickbait, as shown in Figure 26.

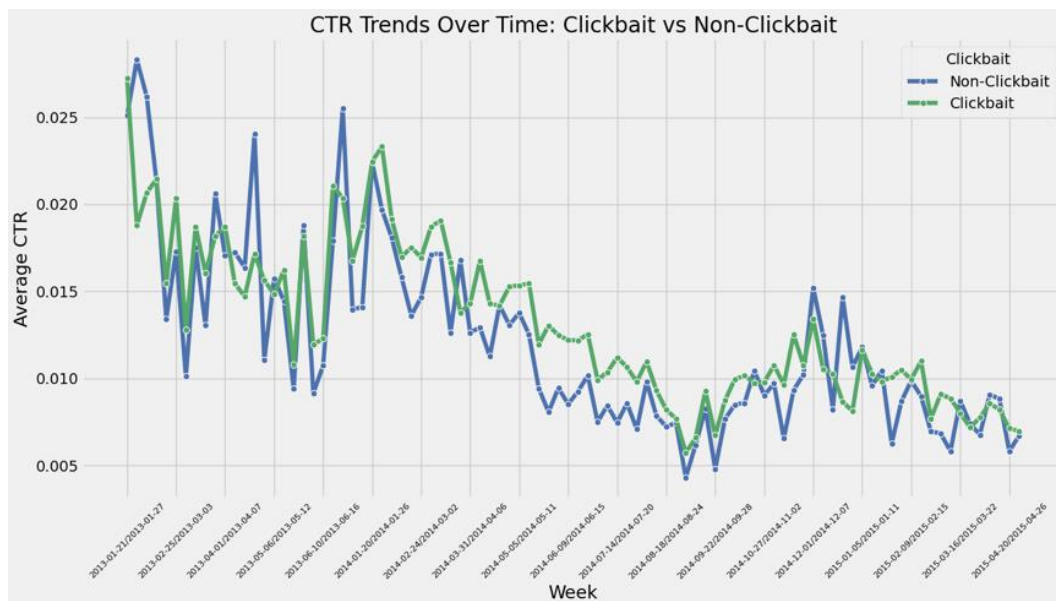


Figure 26: CTR Trends Over Time: Clickbait vs Non-Clickbait

The visualization showed that the overall trend demonstrated a decline over time, this might reflect audience fatigue, platform changes, or broader shifts in engagement. At most points in time, clickbait content had a slightly higher *CTR*, simultaneously, earlier weeks show more volatility for both classes and an overall trend where the gap between classes also decreased over time, perhaps indicating a declining effectiveness of clickbait strategies.

Another very relevant analysis is how the performance of clickbait or non-clickbait headlines, measured by the *CTR* changes within different categories of data. Was clickbait always effective in boosting performance, or were there cases where audiences of a specific niche demonstrated a preference for more serious, straightforward content? To analyze this, the *CTR* distribution by *category* previously identified in chapter 4, and *clickbait* was calculated and plotted, as shown in Figure 27.

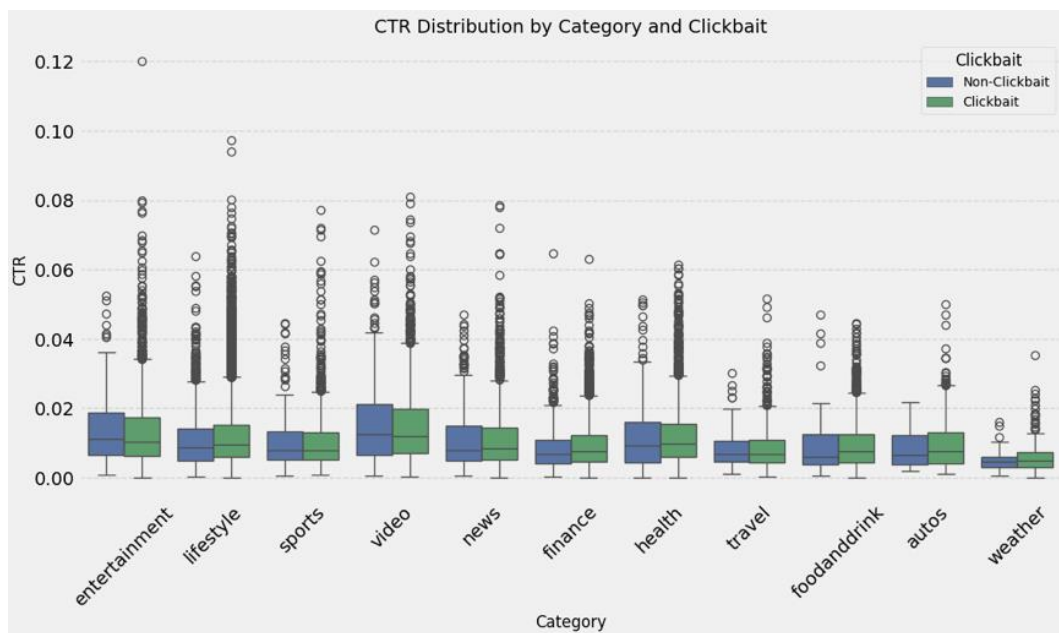


Figure 27: CTR Distribution by Category and Clickbait

From this visualization, it was possible to understand that across most categories, clickbait headlines tend to have a slightly higher median CTR. However, the difference was not very pronounced in some categories, indicating that the effectiveness of clickbait might vary by category. Categories such as “entertainment” and “video” demonstrated a higher CTR overall, suggesting that these audiences were more engaged regardless of whether the headline was clickbait or not. In some categories, such as “lifestyle” and “finance,” clickbait showed a noticeable boost in CTR, whereas in others, such as “health” and “news,” the impact was minimal, implying that the effectiveness of clickbait is context dependent. The maximum *CTR* values for clickbait were noticeably higher than for non-clickbait in most categories, suggesting

that the effectiveness of reaching exceptional engagement was, in most cases, associated with clickbait techniques.

To better understand in which categories clickbait had a positive or negative effect, the performance impact was calculated by subtracting the mean CTR of the clickbait headlines by the non-headlines, as shown in Figure 28.

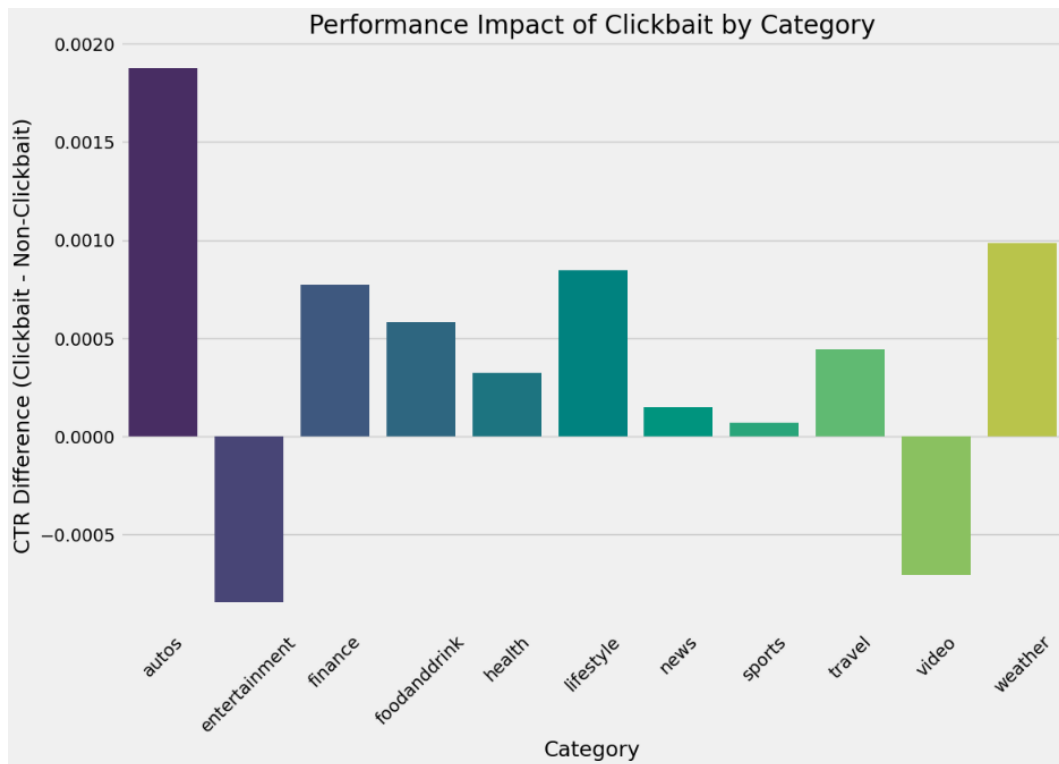


Figure 28: Performance Impact of Clickbait by Category

The visualization revealed that different audience categories respond differently to the clickbait techniques of Upworthy. Categories such as "autos," "lifestyle," and "weather" demonstrated the highest positive impact, whereas "entertainment" and "video" were the only categories with a negative impact, possibly indicating audience fatigue regarding clickbait techniques. Others, such as "finance", "foodanddrink", "health", "news" and "sports" showed relatively neutral impact, indicating that clickbait had a limited effect on engagement.

As a final analysis, a comparison of the top five headlines of each class was made, as shown in Table 20, revealing different strategies. The top clickbait headlines employed techniques such as Juxtaposition to evoke curiosity and emotional hooks contrasting emotions

and hint at conflict resolution. The *CTR* varied from 0.08 to 0.12, demonstrating strong engagement levels for these strategies. The top non-clickbait headlines, on the other hand, focused on delivering informative or substantive content without relying on hooks, with a focus on social issues, direct descriptions, and highlighting achievements, these headlines underperformed compared to the clickbait ones with a *CTR* ranging from 0.058 to 0.072.

8.1.3 Business Insights and Impact

After all the previous analysis, it was now necessary to transform the insights drawn from the data into real, actionable business insights that can create an impact on how to use these clickbait tools more effectively, driving up engagement.

The first recommendation is to do category-specific clickbait, it was observed in the analysis, categories like “autos”, “weather” and “lifestyle” were niches where clickbait was more effective, whereas in “sports”, “news”, “health” and others did not have that much of an improvement. The strategy should be to focus efforts on high-performing categories to maximize short-term engagement while limiting to niches where it was less impactful because those specific niches might already be in a phase of saturation of clickbait. This strategy might lead to an instant improvement in overall engagement metrics, however, it must be used carefully to not be overused and harm the niches that still respond. This can be done by performing A/B testing, by applying it in top-performing categories, fine-tuning the clickbait strategies, and limiting efforts on low-impacted categories, more time will be available to use in other projects.

The second recommendation is to monitor and adapt to *CTR* trends, as seen in Figure 26, the gap between clickbait and non-clickbait headlines over time suggests diminishing returns on performance for clickbait strategies. Audience behaviors change naturally over time, what works in one period might not be ideal in another, requiring continuous monitoring and adaptation. This use of data and analytics will allow digital content publishers to be more data-

driven and will be able to ensure that strategies remain effective and relevant. Said strategy can be applied by setting up automated analytics tools such as dashboards and by identifying periods of high *CTR* and the contributing factors.

The third and final recommendation is to develop a hybrid headline strategy, as seen prior in the analysis, clickbait boosts short-term performance metrics. However, it seems to decrease its effect in the long term. Simultaneously, it was noted the different techniques and elements of headlines classified as clickbait and non-clickbait, developing a hybrid strategy that incorporates curiosity-driven elements while ensuring the content still aligns with audience expectations without creating audience fatigue can have a great impact. This can be achieved by testing different combinations of headlines, and performance can be assessed through A/B testing or by utilizing the predictive model in previous chapters.

8.1.4 Conclusion, Future Research, and Limitations

This research faced a lot of challenges from the get-go, from the scarcity of relevant features and overall lack of high-quality data from the Upworthy Research Archive. The archive contained multiple features that appeared to have a lot of potential to improve this study, such as the images presented next to the headline. Unfortunately, the archive only had an identifier and not the image file, in future research, visual elements could be studied and explored, and how these affect performance metrics and how could an image be clickbait.

Another element that was not optimal was the lack of validation data to understand if a headline in the Upworthy Research Archive was or was not clickbait, even if, according to the modeling made, the record was classified as such, it would be very interesting to get the input from the publishers of said article, understand what the Upworthy platforms considers as clickbait. In future studies, performing this kind of analysis directly with a publisher could draw even more interesting results.

Simultaneously, it would have been great to be able to get information about how users reacted to clickbait articles. The performance metric used in this research was the CTR as it allowed the understanding of which articles were performing better, however, other metrics such as time spent reading the article, time spent in the platform, and if that user came back to the website on other occasions could have increased the detail and improved the overall analysis. Having this kind of data in future research might allow for even better business insights, enabling any digital content publisher to further improve performance metrics.

Despite all these limitations, this research did meet its proposed objectives. Clickbait was identified in the Upworthy platform, and the impact of this technique on its titles was analyzed. All of this allowed for the identification of insights that led to the creation of several business insights that aim to improve performance and experience for all audiences of the platform. All three business insights have the potential to significantly impact digital content publishers way of looking and applying clickbait. These achievements highlight the practical applicability of the findings in shaping content strategies and optimizing engagement.

9. Limitations and Future Research

Over the last few years, as technology has become increasingly embedded in information channels, multiple studies have been conducted on how to achieve popularity. News articles are written to become as engaging as possible, but publishing an article that will likely be popular is extremely challenging. Defining content as highly engaging is particularly subjective and hard to quantify, with several variables that might not be present in the data.

During the discussion on identifying patterns that predict content success and derive actionable insights for creating high-engagement content, the lack of features and large amounts of entries with null or irrelevant content related to headline or excerpt imposed several limitations when performing the feature engineering and predictive tasks. Despite having taken relevant and insightful conclusions from the Upworthy Research Archive, a larger dataset with more features could have improved the results, as more data typically enhances machine learning performance. Nonetheless, with the results of this research, future studies could focus on creating a model that could give a performance score for each headline and excerpt for a certain article, making the process of a data-driven decision more dynamic and useful for the creator.

Moreover, while deploying NLP models, computational power limitations were reflected in several hours to perform tasks and frequent kernel disconnections despite the usage of external cloud-based services to implement the models. The lack of resources available reflected the impossibility of customizing and testing parameters of the NLP models and created the need to use transfer learning methods imported from Hugging Face. With higher computational power, future research could achieve better results and utilize more complex techniques. In future studies, the retrieving of image data associated with each headline could

also lead to the use of models like Convolutional Neural Networks. This would lead to interesting results regarding the reader's behaviors to different aspects of these images.

Additionally, most feature engineering tasks were carried out by extracting information from the headline and excerpt columns. The lack of textual features created a strong reliance on the quality of these two variables, which may result in missing context and limit model complexity. More information regarding each article's content could have significantly enhanced the results, as a more comprehensive dataset with various variables typically leads to richer, more accurate outcomes analyzes.

Finally, the lack of features related to the engagement of the article proved to be a high barrier to this study. Metrics like time spent reading and shares could significantly improve the results of the models developed across this study. The fact that the only metric that resembled engagement was CTR restrained the study from more insightful results, in future studies, having different perspectives on success can lead to other valuable results.

10. Conclusion

As the nature of digital content creation continues to evolve, the findings extracted from the present study emphasize the capability to predict engagement in the Upworthy dataset using a different set of features. This discovery paves the way for several actionable insights that enable digital content creators to drive engagement and improve overall performance.

Throughout the whole study, different features were tested and analyzed in their performance of predicting. The research concluded the most important features for prediction was the timing, revealing that the moment a piece of digital content is published plays a big role in dictating its success. The second most important characteristic was the structure of the headline, indicating that the number of words and characters are also very important in the Upworthy context. The emotional charge delivered in the articles was, in some cases, moderately impactful, emotions like *trust* and *positivity* did contribute, while others were no good.

In the present research, the focus was always on using several different techniques to reach a common goal, to be able to create meaningful insight that can create an impact on Upworthy and other digital content publishers. The result of this study enables Upworthy and other creators to test without the need for A/B testing, minimizing the risk of underperformance while reducing the use of resources. Simultaneously, this tool can be used to optimize headline recommendations, making sure a headline is tailored to the specific audience of the publisher. Finally, this predictive model can also be used to make real-time engagement predictions, by integrating it into the publication workflow, teams will be able to prioritize and publish headlines in the best possible moment.

To further expand the understanding of how to drive engagement in the Upworthy context, additional studies were conducted. These contributed a lot to reach new perspectives and even more insightful and actionable insights into how to improve performance. For starters, a deeper dive was made into the relationships between different features and engagement, enabling the understanding that by tailoring headline lengths, reducing excessive punctuation, and strategically employing sentiments and content categories, engagement in the Upworthy context will rise. Following that, a recommendation system was developed, and with that, the possibility to increase engagement by building loyalty with the audience through tailored recommendations, driving engagement even more. In a different analysis, time was thoroughly studied, leading to the understanding of how to refine content strategies so that they engage audiences over the long term and enhance CTR, enabling digital content publishers to know when the right time is. Finally, clickbait was identified and analyzed in the Upworthy Research Archive context, enabling multiple business insights such as using category-specific clickbait, monitoring and adapting clickbait to *CTR* trends, and developing a hybrid headline strategy, all leading to one goal, how to use clickbait to improve performance in the short and long-term.

Overall, this study not only achieved its proposed objectives but also went beyond by transforming straightforward conclusions into actionable insights. All these insights enable digital content creators to optimize and enhance their content by leveraging data analysis, machine learning, and other techniques, this study was able to not only provide a conclusion but also create a meaningful impact.

11. References

Agrawal, Amol. 2016. "Clickbait Detection Using Deep Learning." In *Proceedings of the 2nd International Conference on Next Generation Computing Technologies (NGCT-2016*, Dehradun, India, October 14–16), 268–272. <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7877426>.

Akuma, Stephen, Tyosar Lubem, and Isaac Terngu Adom. 2022. "Comparing Bag of Words and TF-IDF with Different Models for Hate Speech Detection from Live Tweets." *International Journal of Information Technology* 14 (7): 3629–3635. <https://doi.org/10.1007/s41870-022-01096-4>

Ali, Rasikh, Tayyaba Farhat, Sanya Abdullah, Sheeraz Akram, Mousa Alhajlah, Awais Mahmood, and Muhammad Amjad Iqbal. 2023. "Deep Learning for Sarcasm Identification in News Headlines." *Applied Sciences* 13 (9): 5586. <https://doi.org/10.3390/app13095586>.

Anand, Aman. n.d. Clickbait Dataset. Kaggle. <https://www.kaggle.com/datasets/amananandrai/clickbait-dataset>.

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

Bhoj, N., A. R. Dwivedi, A. Tripathi, and B. Pandey. 2021. "LSTM Powered Identification of Clickbait Content on Entertainment and News Websites." In *Proceedings of the 13th International Conference on Computational Intelligence and Communication Networks (CICN)*, Lima, Peru, 2021, 29–33. <https://doi.org/10.1109/CICN51697.2021.9574677>.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022.

Blom, J. N., and K. R. Hansen. 2015. "Click Bait: Forward-Reference as Lure in Online News Headlines." *Journal of Pragmatics* 76: 87–100. <https://doi.org/10.1016/j.pragma.2014.11.010>.

Cambridge Dictionary. n.d. "Clickbait." Retrieved December 3, 2024, from <https://dictionary.cambridge.org/dictionary/english/clickbait>.

Cao, Xinyue, Thai Le, and Jason Zhang. 2017. "Machine Learning Based Detection of Clickbait Posts in Social Media." arXiv. <https://doi.org/10.48550/arXiv.1710.01977>.

Chakraborty, A., B. Paranjape, S. Kakarla, and N. Ganguly. 2016. "Stop Clickbait: Detecting and Preventing Clickbaits in Online News Media." In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, San Francisco, CA, USA, 2016, 9–16. <https://doi.org/10.1109/ASONAM.2016.7752207>.

Chen, Yimin, Niall J. Conroy, and Victoria L. Rubin. 2015. "Misleading Online Content: Recognizing Clickbait as 'False News'." In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection (WMDD '15)*, New York, NY, USA, 15–19. <https://doi.org/10.1145/2823465.2823467>.

De Choudhury, Munmun, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. "Predicting Depression via Social Media." In *Proceedings of the International AAAI Conference on Web and Social Media*, 7 (1): 128–137.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv. <https://doi.org/10.48550/arXiv.1810.04805>.

Ekman, P. 1992. "An Argument for Basic Emotions." *Cognition and Emotion* 6 (3): 169–200.

Evidently AI. "Accuracy vs. Precision vs. Recall in Machine Learning: What's the Difference?" Evidently AI. Accessed November 25, 2024. <https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall>.

Farias, Felipe C., Teresa B. Ludermir, and Carmelo J. A. Bastos-Filho. 2020. "Similarity-Based Stratified Splitting: An Approach to Train Better Classifiers." Preprint. <https://doi.org/10.48550/arXiv.2010.06099>.

Fathima, Afrah, G. Shree Devi, Mohd Faizaanuddin, et al. 2023. "Exploring the Potency of Machine Learning Approaches in Enhancing Spam Detection Accuracy." PREPRINT (Version 1), August 28, 2023. Research Square. <https://doi.org/10.21203/rs.3.rs-3270960/v1>.

Foody, G. M. 2023. "Challenges in the Real World Use of Classification Accuracy Metrics: From Recall and Precision to the Matthews Correlation Coefficient." PLOS ONE 18 (10): e0291908. <https://doi.org/10.1371/journal.pone.0291908>.

Gligorić, K., G. Lifchits, R. West, and A. Anderson. 2023. "Linguistic Effects on News Headline Success: Evidence from Thousands of Online Field Experiments (Registered Report)." PLOS ONE 18 (3): e0281682. <https://doi.org/10.1371/journal.pone.0281682>.

Hardt, D., D. Hovy, and S. Lamprinidis. 2018. "Predicting News Headline Popularity with Syntactic and Semantic Knowledge Using Multi-task Learning." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, edited by E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, 659–664. EMNLP 2018. Association for Computational Linguistics.

Heath, Chip, Chris Bell, and Emily Sternberg. 2001. "Emotional Selection in Memes: The Case of Urban Legends." **Journal of Personality and Social Psychology** 81: 1028–1041. <https://doi.org/10.1037/0022-3514.81.6.1028>.

Heimbach, Irina, and Oliver Hinz. 2016. "The Impact of Content Sentiment and Emotionality on Content Virality." **International Journal of Research in Marketing** 33 (3): 695–701. <https://doi.org/10.1016/j.ijresmar.2016.02.004>.

Shulman, Hillary C., et al. 2024. "Reading Dies in Complexity: Online News Consumers Prefer Simple Writing." **Science Advances** 10: eadn2555. <https://doi.org/10.1126/sciadv.adn2555>.

Hutto, C., and E. Gilbert. 2014. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." **Proceedings of the International AAAI Conference on Web and Social Media** 8 (1): 216–225. <https://doi.org/10.1609/icwsm.v8i1.14550>.

Iarovici, Edith, and Rodica Amel. 1989. "The Strategy of the Headline." **Semiotica** 77 (4): 441–460. <https://doi.org/10.1515/semi.1989.77.4.441>.

Jhaveri, S., I. Khedkar, Y. Kantharia, and S. Jaswal. 2019. "Success Prediction Using Random Forest, CatBoost, XGBoost and AdaBoost for Kickstarter Campaigns." In **2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)**, 1170–1173. Erode, India. <https://doi.org/10.1109/ICCMC.2019.8819828>.

Jung, A. K., S. Stieglitz, T. Kissmer, M. Mirbabaie, and T. Kroll. 2022. "Click Me...! The Influence of Clickbait on User Engagement in Social Media and the Role of Digital Nudging." *PLOS ONE* 17(6): e0266743. <https://doi.org/10.1371/journal.pone.0266743>.

Kundu, Rohit. 2022. "F1 Score in Machine Learning: Intro & Calculation." *V7 Labs*. Published December 16, 2022. <https://www.v7labs.com/blog/f1-score-guide>.

Li, L., Goh, T.T., & Jin, D. 2020. How textual quality of online reviews affect classification performance: a case of deep learning sentiment analysis. *Neural Comput & Applic* 32: 4387–4415. <https://doi.org/10.1007/s00521-018-3865-7>.

Liu, C.-Z., Sheng, Y.-X., Wei, Z.-Q., & Yang, Y.-Q. 2018. "Research of Text Classification Based on Improved TF-IDF Algorithm." In 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE), Lanzhou, China, 218-222. <https://doi.org/10.1109/IRCE.2018.8492945>.

Schütze, H., Manning, C. D., & Raghavan, P. 2008. *Introduction to Information Retrieval* (Vol. 39, pp. 234-265). Cambridge: Cambridge University Press.

Matias, J.N., Munger, K., Le Quere, M.A. et al. 2021. The Upworthy Research Archive, a time series of 32,487 experiments in U.S. media. *Sci Data* 8, 195. <https://doi.org/10.1038/s41597-021-00934-7>.

Mohammad, Saif M., and Peter D. Turney. 2013. "Crowdsourcing a Word-Emotion Association Lexicon." *Computational Intelligence* 29(3): 436-465. <https://doi.org/10.48550/arXiv.1308.6297>.

Mukherjee, P., Dutta, S., and De Bruyn, A. 2022. "Did Clickbait Crack the Code on Virality?" *Journal of the Academy of Marketing Science* 50: 482–502. <https://doi.org/10.1007/s11747-021-00830-x>.

Ozsahin, D. U., Mustapha, M. T., Mubarak, A. S., Ameen, Z. S., and Uzun, B. 2022. "Impact of Feature Scaling on Machine Learning Models for the Diagnosis of Diabetes." In *Proceedings of the 2022 International Conference on Artificial Intelligence in Everything (AIE)*, Lefkosa, Cyprus, 87–94. <https://doi.org/10.1109/AIE57029.2022.00024>.

Petty, Richard E., and John T. Cacioppo. 1986. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. New York: Springer-Verlag.

Potthast, M., Köpsel, S., Stein, B., and Hagen, M. 2016. "Clickbait Detection." In **Advances in Information Retrieval**, edited by N. Ferro et al., 9626: Lecture Notes in Computer Science. Springer, Cham. https://doi.org/10.1007/978-3-319-30671-1_72.

Pujahari, A., and Sisodia, D. S. 2021. "Clickbait Detection Using Multiple Categorisation Techniques." *Journal of Information Science* 47(1): 118-128. <https://doi.org/10.1177/0165551519871822>.

Ratul, Ishrak Jahan, Ummay Habiba Wani, Mirza Muntasir Nishat, Abdullah Al-Monsur, Abrar Mohammad Ar-Rafi, Fahim Faisal, and Mohammad Ridwan Kabir. 2022. "Survival Prediction of Children Undergoing Hematopoietic Stem Cell Transplantation Using Different Machine Learning Classifiers by Performing Chi-squared Test and Hyper-parameter Optimization: A Retrospective Analysis." arXiv:2201.08987 [cs.LG]. <https://doi.org/10.48550/arXiv.2201.08987>.

Robertson, C.E., Pröllochs, N., Schwarzenegger, K., et al. 2023. "Negativity Drives Online News Consumption." **Nature Human Behaviour** 7: 812–822. <https://doi.org/10.1038/s41562-023-01538-4>.

Rooein, Donya, Paul Rottger, Anastassia Shaitarova, and Dirk Hovy. 2024. "Beyond Flesch-Kincaid: Prompt-based Metrics Improve Difficulty Classification of Educational Texts." **arXiv** (May 15, 2024). <https://doi.org/10.48550/arXiv.2405.09482>.

Roy, Arpita, and Shimei Pan. 2021. "Incorporating Medical Knowledge in BERT for Clinical Relation Extraction." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5357–5366. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Salmerón, Román, Catalina B. García, and José García. 2020. "Overcoming the Inconsistencies of the Variance Inflation Factor: A Redefined VIF and a Test to Detect Statistical Troubling Multicollinearity." arXiv, May 5, 2020. <https://arxiv.org/abs/2005.02245>.

Schmitt, Josephine B., Christina A. Debbelt, and Frank M. Schneider. 2017. "Too Much Information? Predictors of Information Overload in the Context of Online News Exposure." *Information, Communication & Society* 21 (8): 1151–67. <https://doi.org/10.1080/1369118X.2017.1305427>.

Szabo, Gabor, and Bernardo A. Huberman. 2010. "Predicting the Popularity of Online Content." *Commun. ACM* 53 (8): 80–88. <https://doi.org/10.1145/1787234.1787254>.

Thompson, Derek. 2013. "Upworthy: I Thought This Website Was Crazy, but What Happened Next Changed Everything." *The Atlantic*, November 14, 2013. <https://www.theatlantic.com/business/archive/2013/11/upworthy-i-thought-this-website-was-crazy-but-what-happened-next-changed-everything/281472/>.

Tuligenga, Richard H., Aline Dugravot, Adam G. Tabák, Alexis Elbaz, Eric J. Brunner, Mika Kivimäki, and Archana Singh-Manoux. 2014. "Midlife Type 2 Diabetes and Poor Glycaemic Control as Risk Factors for Cognitive Decline in Early Old Age: A Post-Hoc Analysis of the Whitehall II Cohort Study." *The Lancet Diabetes & Endocrinology* 2 (3): 228–235. [https://doi.org/10.1016/S2213-8587\(13\)70192-X](https://doi.org/10.1016/S2213-8587(13)70192-X).

Wallach, Hanna M., David Mimno, and Andrew McCallum. 2009. "Rethinking LDA: Why Priors Matter." In *Proceedings of the 23rd International Conference on Neural Information Processing Systems (NIPS'09)*, 1973–1981. Curran Associates Inc., Red Hook, NY, USA.

Wei, Jason, and Kai Zou. 2019. "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks." In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 6382–6388. Hong Kong, China: Association for Computational Linguistics. <https://doi.org/10.18653/v1/D19-1670>.

Weiss, Howard M., and Russell Cropanzano. 1996. "Affective Events Theory." *Research in Organizational Behavior* 18 (1): 1-74.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. "Transformers: State-of-the-Art Natural Language Processing." In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45. Online: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-demos.6>.

Wu, Fangzhao, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. "MIND: A Large-Scale Dataset for News Recommendation." In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 3597–3606. Online: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.331>.

Yang, L., Zheng, W., and Xiao, Y. 2022. "Exploring Different Interaction Among Features for CTR Prediction." *Soft Computing* 26: 6233–6243. <https://doi.org/10.1007/s00500-022-07149-x>.

Zannettou, S., Chatzis, S., Papadamou, K., and Sirivianos, M. 2018. "The Good, the Bad and the Bait: Detecting and Characterizing Clickbait on YouTube." In 2018 IEEE Security

and Privacy Workshops (SPW), 63-69. San Francisco, CA, USA: IEEE.
<https://doi.org/10.1109/SPW.2018.00018>.

Zuhroh, Nurrida Aini, and Nur Aini Rakhmawati. 2019. "Clickbait Detection: A Literature Review of the Methods Used." Register: Jurnal Ilmiah Teknologi Sistem Informasi.
<https://doi.org/10.26594/register.v6i1.1561>.

Appendix:

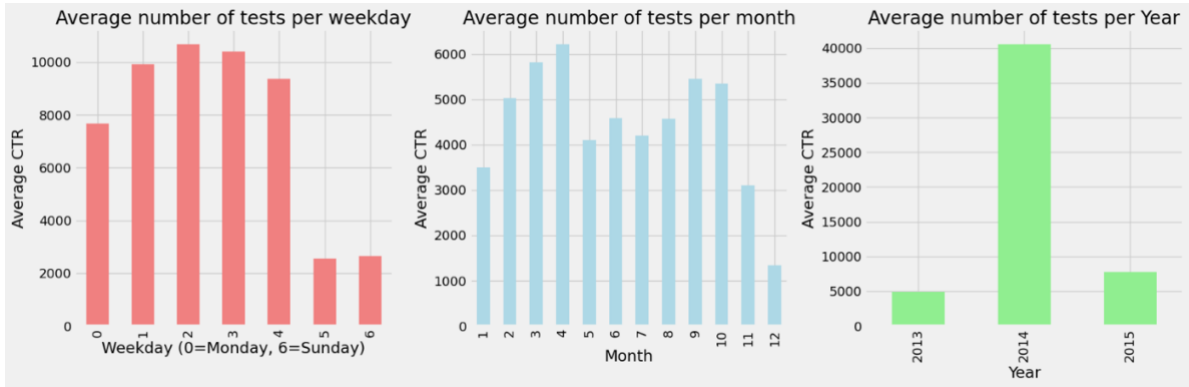


Figure 2 – Average CTR per Average Number of Tests per Weekday, Month and Year (Created_at)

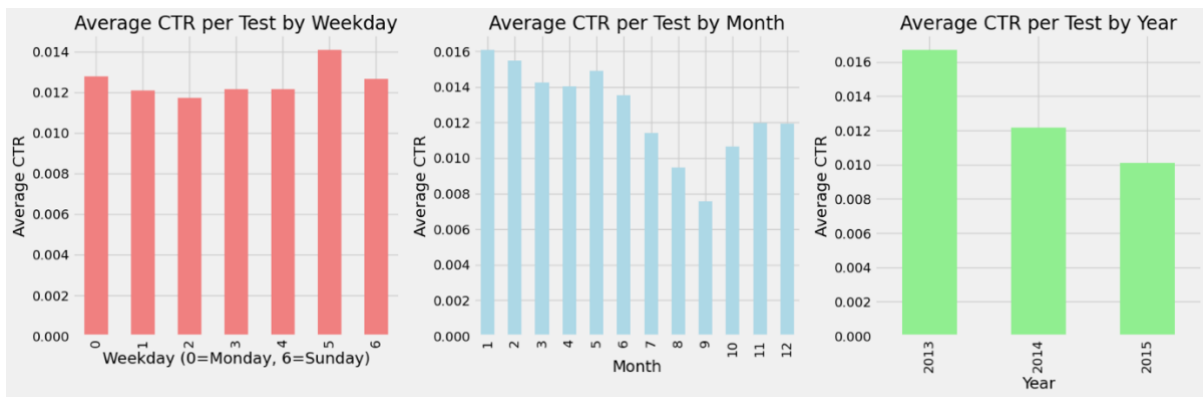


Figure 3 – Average CTR per Test by Weekday, Month and Year (Created_at)

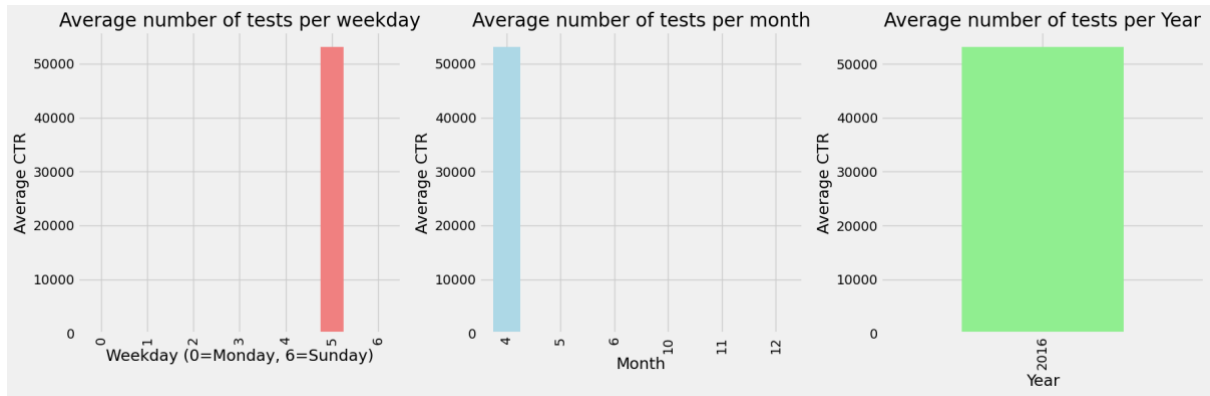


Figure 4 – Average CTR per Average Number of Tests per Weekday, Month and Year (Updated_at)

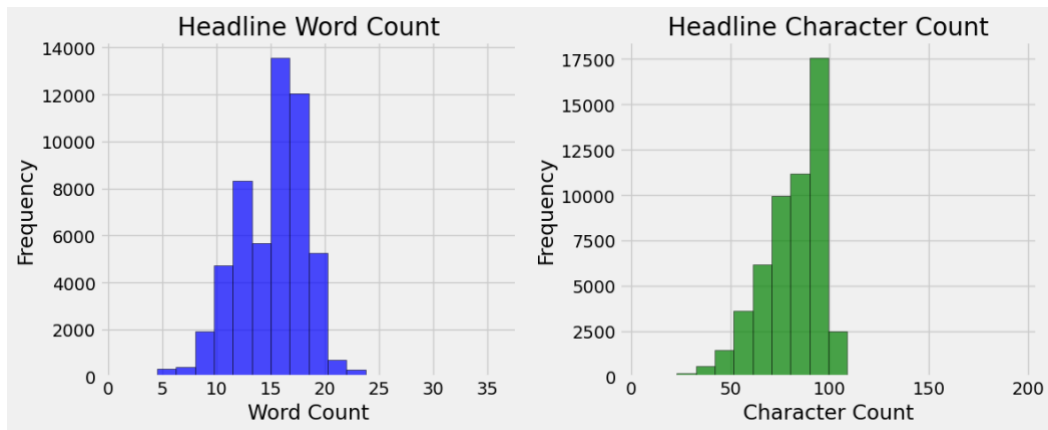


Figure 5 – Distribution of Headlines Word's and Character's

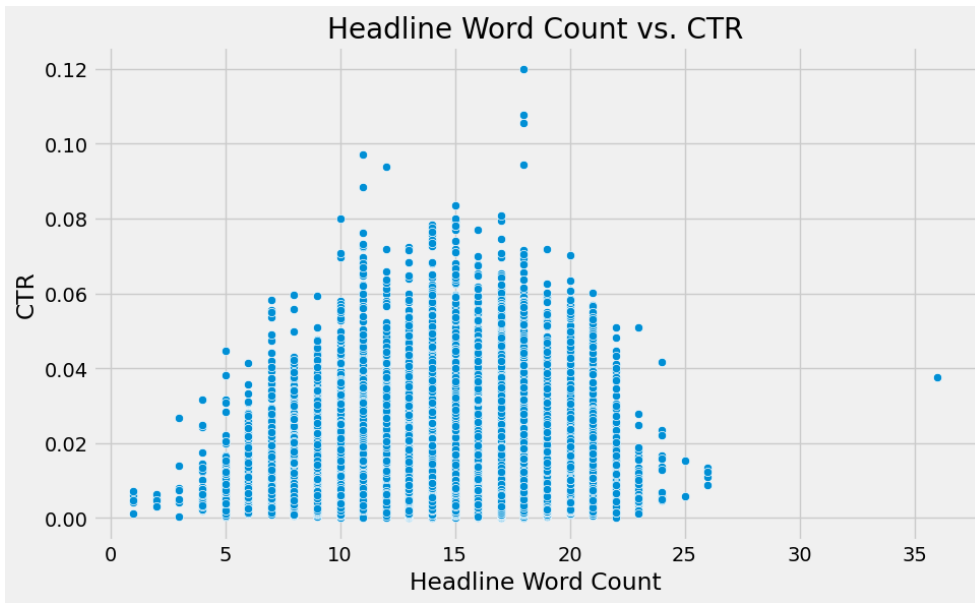


Figure 6 – Headline Word Count versus CTR

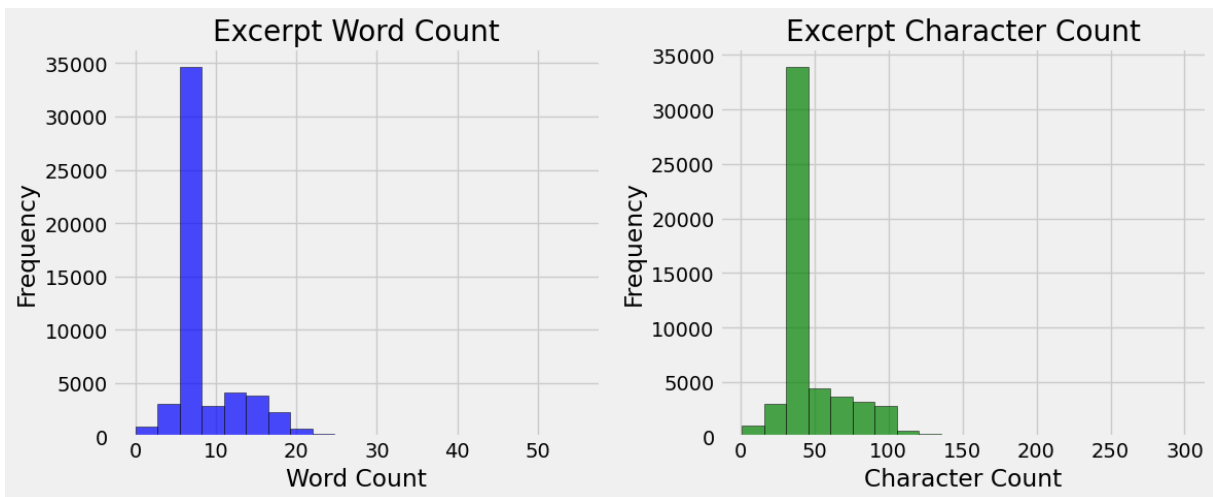


Figure 7 - Distribution of Excerpt's Word and Characters

Table 3 – Excerpt unique values frequency

| Excerpt | Count |
|---|-------|
| Things that matter. Pass 'em on. | 29264 |
| Things that matter. Pass 'em on. #PromotedPost | 239 |
| Things that matter. Pass 'em on. | 188 |
| #PromotedPost Things that matter. Pass 'em on. | 84 |
| Fascinating. | 66 |
| Stories that matter. Pass them on and on. | 54 |
| The goal here isn't to pick sides. It's to hav... | 40 |
| Things that matter. Pass 'em on. #promotedpost | 39 |
| <i>Null</i> | 38 |
| The next time someone tells you gay marriage w... | 37 |

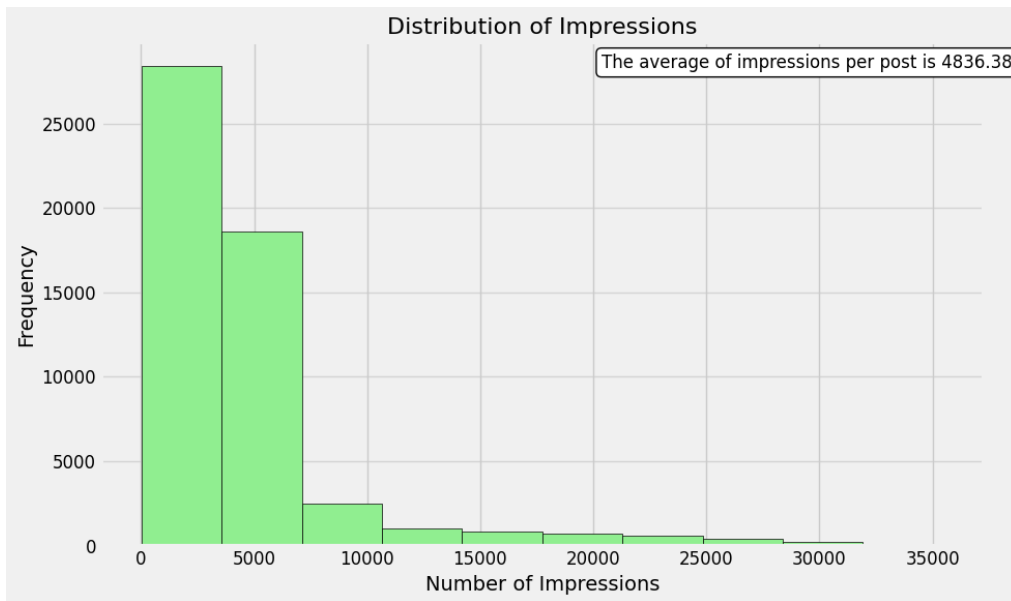


Figure 8 – Distribution of Impressions

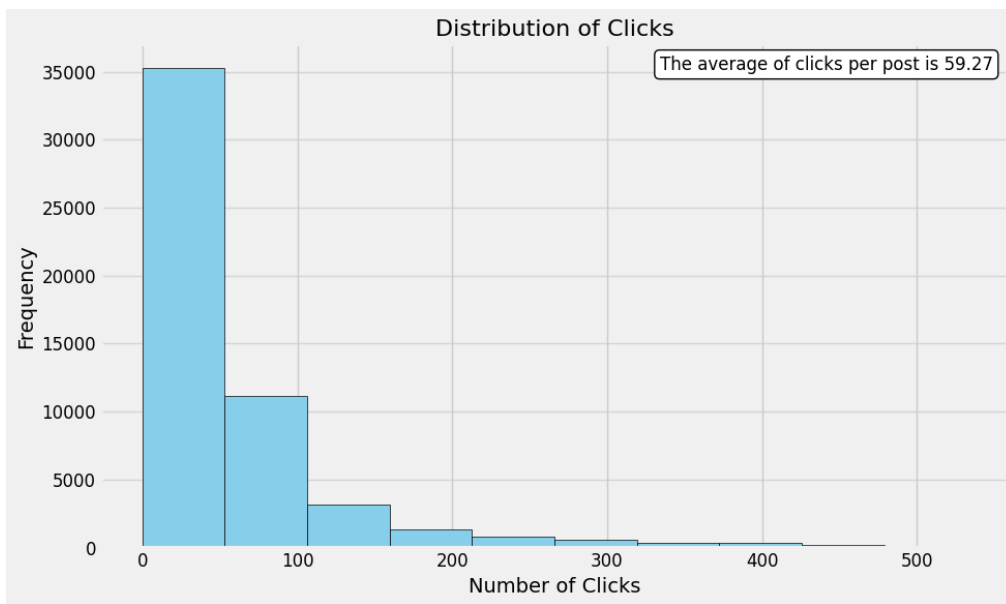


Figure 9 – Distribution of clicks

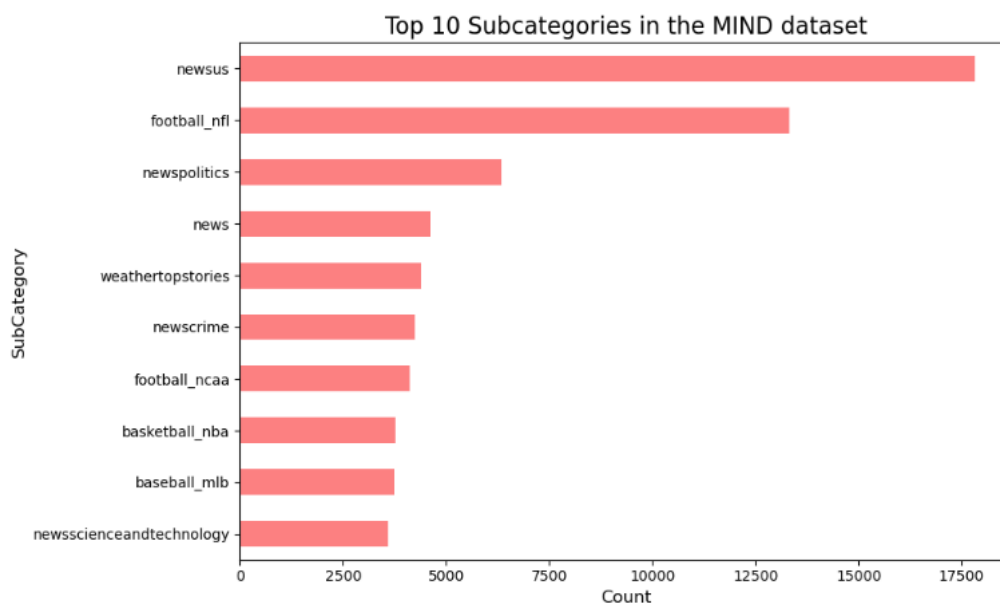


Figure 11: Distribution SubCategories MIND Dataset

Table 4: Metrics Comparison Between All Models

| Model | Test Accuracy | Test Precision | Test Recall | Test F1-score | Test Log Loss |
|---------------------|---------------|----------------|-------------|---------------|---------------|
| Logistic Regression | 73% | 72% | 73% | 73% | 1.02 |
| Naïve Bayes | 72% | 72% | 72% | 72% | 0.85 |
| SVM | 77% | 76% | 77% | 73% | 0.73 |
| Voting Classifier | 77% | 76% | 77% | 76% | 0.83 |
| BERT | 86% | 86% | 86% | 86% | 0.48 |

Table 5 - Emotions with Highest Average CTR

| Emotions with Highest Average CTR | |
|-----------------------------------|-------------|
| Emotion | Average CTR |
| positive | 0.008703 |
| negative | 0.007048 |
| trust | 0.006598 |
| anticipation | 0.005208 |
| fear | 0.004983 |
| joy | 0.004844 |
| sadness | 0.004035 |
| anger | 0.003966 |
| disgust | 0.003034 |
| surprise | 0.002521 |

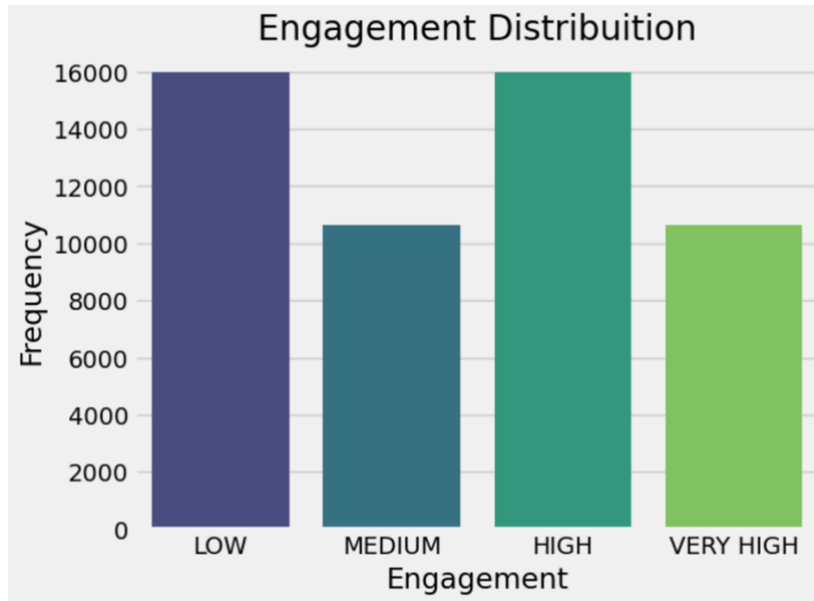


Figure 17 – Distribution of engagement categories

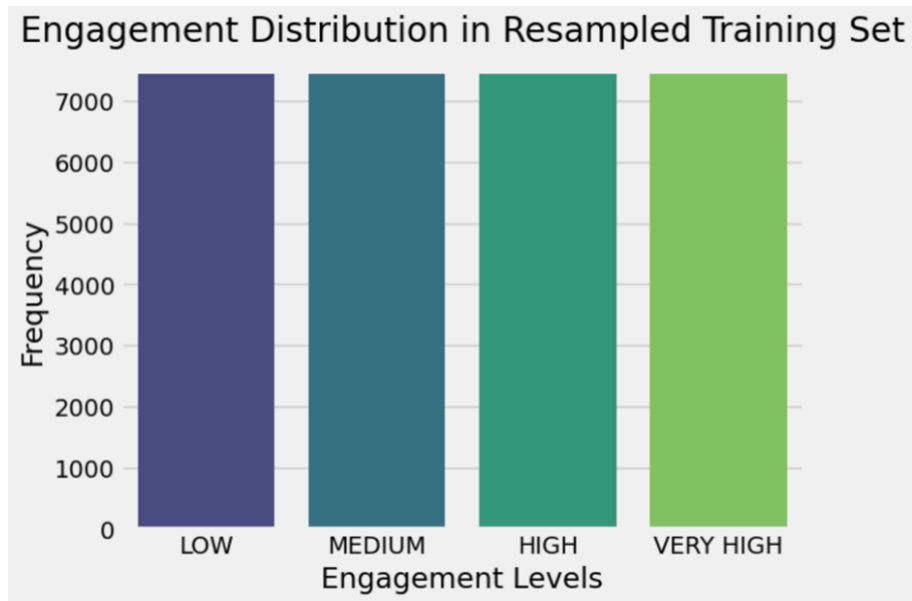


Figure 18 – Distribution of Engagement Levels in Resampled Training Set

Table 6 – Final Dataset Features Metadata for Predictions

| Feature | Description |
|----------------------------------|---|
| test_week | Encodes the year and week number |
| disgust | Emotion score representing disgust (0-5) |
| sadness | Emotion score representing sadness (0-5) |
| positive | Emotion score representing positive (0-5) |
| joy | Emotion score representing joy (0-5) |
| negative | Emotion score representing negative (0-5) |
| anger | Emotion score representing anger (0-5) |
| surprise | Emotion score representing surprise (0-5) |
| trust | Emotion score representing trust (0-5) |
| anticipation | Emotion score representing anticipation (0-5) |
| fear | Emotion score representing fear (0-5) |
| predicted_category_entertainment | Binary indicator for predicted “entertainment” category |
| predicted_category_finance | Binary indicator for predicted “finance” category |
| predicted_category_foodanddrink | Binary indicator for predicted “foodanddrink” category |
| predicted_category_health | Binary indicator for predicted “health” category |
| predicted_category_lifestyle | Binary indicator for predicted “lifestyle” category |
| predicted_category_news | Binary indicator for predicted “news” category |

Table 7 – Logistic Regression Evaluation Metrics

| Evaluation Metric | Score |
|----------------------------|--------------|
| Binary Test Accuracy | 75% |
| Precision (Other Class) | 83% |
| Recall (Other Class) | 88% |
| F1-score (Other Class) | 85% |
| Precision (Class 3) | 35% |
| Recall (Class 3) | 26% |
| F1-score (Class 3) | 30% |
| Macro Average Precision | 59% |
| Macro Average Recall | 57% |
| Macro Average F1-score | 57% |
| Weighted Average Precision | 73% |
| Weighted Average Recall | 75% |
| Weighted Average F1-score | 74% |

Table 8 – Random Forest Hyperparameters

| Hyperparameter | Value | Purpose |
|-----------------------|--------------|--|
| random_state | 42 | Ensures reproducibility of results by fixing the random seed |
| n_estimators | 100 | Specifies the number of trees in the forest, balancing performance and training time |
| max_depth | 10 | Limits the depth of each tree to prevent overfitting while maintaining sufficient complexity |
| min_samples_split | 10 | Ensures splits occur only when nodes have at least 10 samples, avoiding splits on noise |
| min_samples_leaf | 5 | Requires at least 5 samples in leaf nodes, reducing variance and preventing overfitting |
| max_features | “sqrt” | Uses the square root of total features at each split to increase tree diversity |
| class_weight | “balanced” | Adjusts weights to handle imbalanced classes, ensuring fair treatment for all |

Table 9 -LightGBM Hyperparameters

| Hyperparameter | Value | Purpose |
|-----------------------|---------------|--|
| objective | “multiclass” | Configures the model for multi-class classification tasks |
| boosting_type | “gbdt” | Specifies Gradient Boosting Decision Trees as the boosting method |
| metric | “multi_error” | Measures the multi-class error rate during training |
| num_class | 4 | Defines the number of distinct classes in the target variable |
| num_leaves | 31 | Limits the number of leaves per tree to control model complexity and prevent overfitting |
| learning_rate | 0.05 | Balances convergence speed and model performance by controlling step size |
| feature_fraction | 0.8 | Uses 80% of features at each boosting iteration to reduce overfitting |

Table 10 - CatBoost Best Hyperparameters

| Hyperparameter | Value | Purpose |
|-----------------|-----------------|--|
| bootstrap_type | “Bernoulli” | Specifies Bernoulli sampling for bootstrapping, reducing training time and improving stability |
| border_count | 128 | Defines the number of bins for numerical feature discretization to improve efficiency |
| depth | 8 | Controls the depth of trees to balance complexity and prevent overfitting |
| grow_policy | “SymmetricTree” | Ensures balanced growth of decision trees for interpretability and stability |
| Iterations | 500 | Specifies the number of boosting iterations to refine predictions incrementally |
| l2_leaf_reg | 3 | Applies L2 regularization to leaf values to reduce overfitting |
| learning_rate | 0.1 | Sets the step size for each iteration, balancing convergence speed and performance |
| od_type | “Iter” | Enables early stopping based on iterations to prevent unnecessary training |
| od_wait | 50 | Defines the number of iterations to wait before early stopping |
| random_strength | 10 | Controls the randomization strength to improve model robustness and generalization |

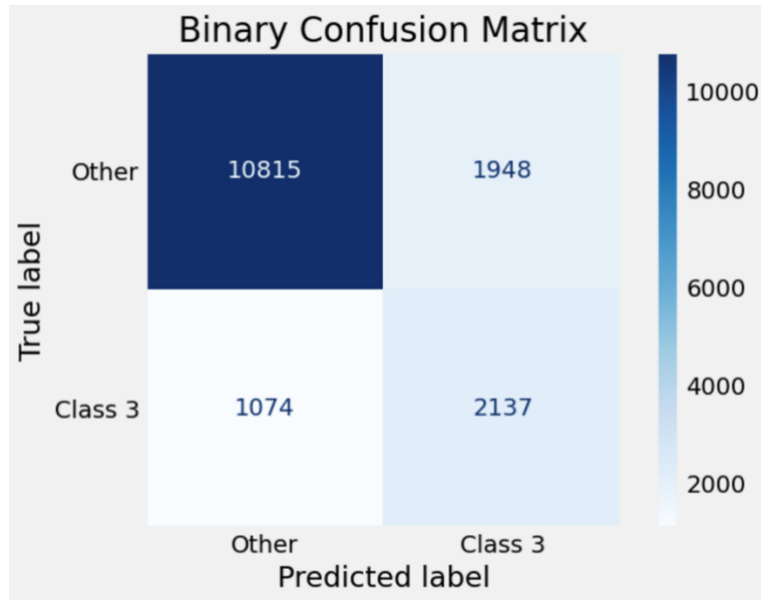


Figure 19 – CatBoost Confusion Matrix

Table 11 – Evaluation Metrics Comparison

| Evaluation Metric | Random Forest | LightGBM | CatBoost |
|----------------------------|---------------|----------|----------|
| Binary Test Accuracy | 77% | 82% | 81% |
| Precision (Other Class) | 91% | 88% | 91% |
| Recall (Other Class) | 79% | 90% | 85% |
| F1-score (Other Class) | 85% | 89% | 88% |
| Precision (Class 3) | 46% | 57% | 52% |
| Recall (Class 3) | 71% | 53% | 67% |
| F1-score (Class 3) | 55% | 55% | 59% |
| Macro Average Precision | 69% | 73% | 72% |
| Macro Average Recall | 75% | 71% | 76% |
| Macro Average F1-score | 70% | 72% | 73% |
| Weighted Average Precision | 82% | 82% | 83% |
| Weighted Average Recall | 77% | 82% | 81% |
| Weighted Average F1-score | 79% | 82% | 82% |

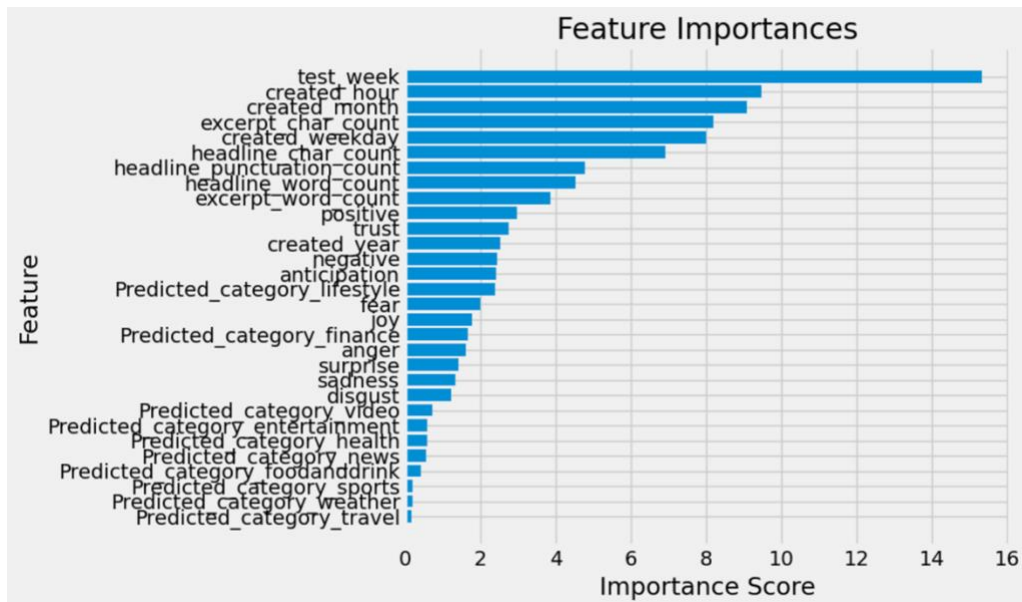


Figure 20 – Distribution of Features Importance

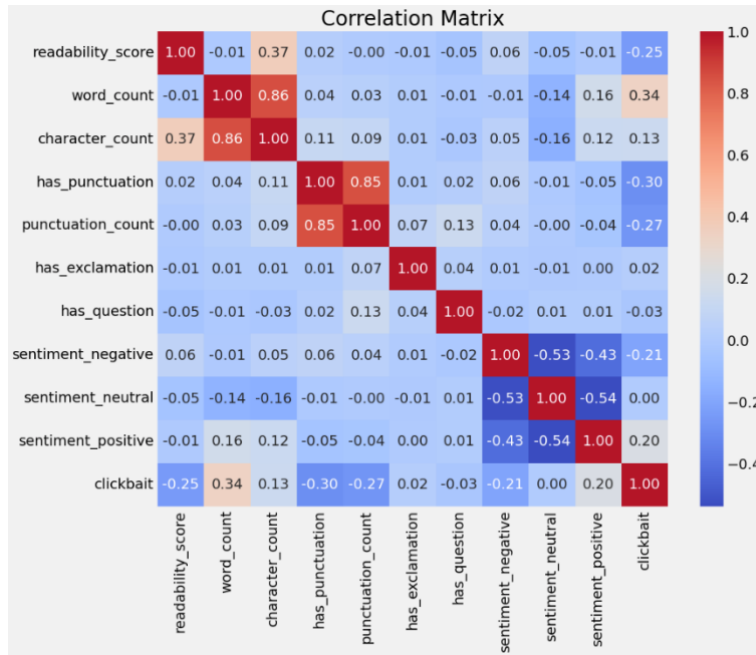


Figure 21: Correlation Analysis First Set of Features

Table 12: VIF Analysis Final Dataset

| Feature | VIF |
|-------------------------|--------|
| avg_characters_per_word | 3.2553 |
| sentiment_polarity | 2.8157 |
| clickbait | 2.0990 |
| has_punctuation | 1.3301 |
| has_question | 1.0077 |
| has_exclamation | 1.0033 |

Table 13: Dataset Description

| Feature | VIF |
|--------------------------|--|
| headline | The textual content of the headline used in the dataset |
| clickbait | Binary indicator if a headline is clickbait |
| has_punctuation | Binary indicator for the presence of any punctuation in the headline |
| has_exclamation | Binary indicator for the presence of any exclamation mark in the headline |
| has_question | Binary indicator for the presence of any question mark in the headline |
| tfidf_0, ..., tfidf_4094 | Numerical features representing the TF-IDF values for vectorized text tokens |
| avg_characters_per_word | Numerical feature representing the average number of characters per word in headline |
| sentiment_polarity | Categorical feature for sentiment polarity |

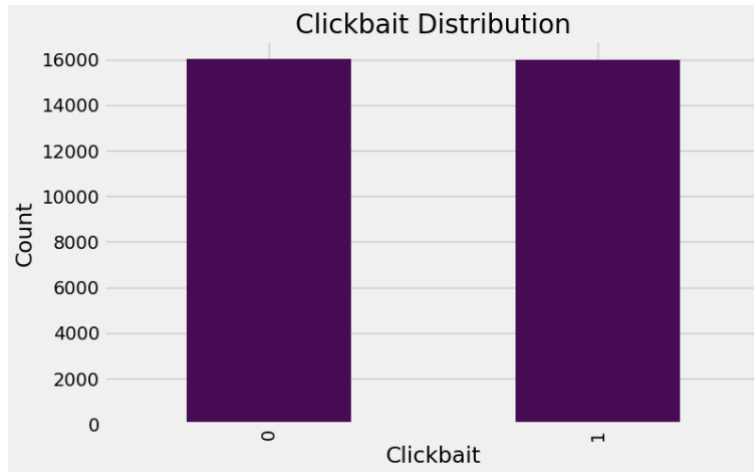


Figure 23: Clickbait Distribution

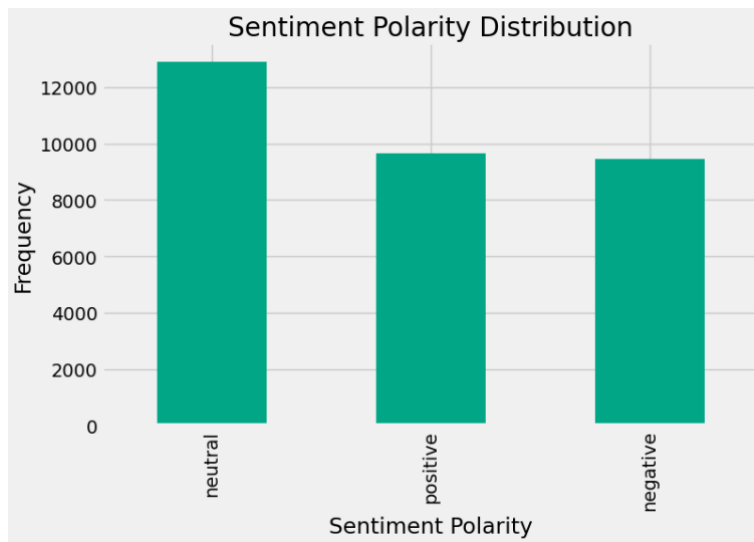


Figure 24: Sentiment_Polarity Distribution

Table 14 – Frequency of Binary Indicators (Punctuation Marks)

| Feature | Value | Frequency | Ratio |
|-----------------|-------|-----------|-------|
| has_question | 0 | 31,840 | 0.995 |
| | 1 | 160 | 0.005 |
| has_exclamation | 0 | 31,962 | 0.999 |
| | 1 | 38 | 0.001 |
| has_punctuation | 0 | 25,986 | 0.812 |
| | 1 | 6,014 | 0.187 |

Table 17: Test Metrics for Predictive Models

| Models | Accuracy | Precision | Recall | ROC-AUC | Log Loss |
|---------------------|----------|-----------|--------|---------|----------|
| Log Reg (Set 1) | 0.96 | 0.97 | 0.95 | 0.99 | 0.14 |
| Naïve Bayes (Set 1) | 0.96 | 0.96 | 0.96 | 0.99 | 0.10 |
| SVM (Set 1) | 0.97 | 0.97 | 0.97 | 0.99 | 0.08 |
| Log Reg (Set 2) | 0.96 | 0.97 | 0.95 | 0.99 | 0.13 |
| Naïve Bayes (Set 2) | 0.96 | 0.96 | 0.96 | 0.99 | 0.10 |
| SVM (Set 2) | 0.97 | 0.97 | 0.96 | 0.99 | 0.08 |
| Log Reg (Set 3) | 0.96 | 0.97 | 0.95 | 0.99 | 0.14 |
| Naïve Bayes (Set 3) | 0.96 | 0.96 | 0.97 | 0.99 | 0.10 |
| SVM (Set 3) | 0.97 | 0.97 | 0.96 | 0.99 | 0.08 |

Table 18: Test Metrics for Final Predictive Model

| Metric | Value |
|------------------------|-------|
| Accuracy | 0.91 |
| Precision (0) | 0.86 |
| Precision (1) | 0.91 |
| Recall (0) | 0.62 |
| Recall (1) | 0.97 |
| F1-score (0) | 0.72 |
| F1-score (1) | 0.94 |
| Support (0) | 40 |
| Support (1) | 160 |
| Macro Avg Precision | 0.89 |
| Macro Avg Recall | 0.80 |
| Macro Avg F1-score | 0.83 |
| Weighted Avg Precision | 0.90 |
| Weighted Avg Recall | 0.91 |
| Weighted Avg F1-score | 0.90 |

Table 20: Top 5 Clickbait and Top 5 Non-Clickbait

| Classification | Headline | CTR |
|----------------|--|------|
| Clickbait | Lorde Found A Photoshopped Photo Of Herself And Decided To Tweet The Untouched Real One Next To It | 0.12 |
| Clickbait | What A Cheery 13-Year-Old Girl Said To Matt Damon Devasted Him | 0.10 |
| Clickbait | A Cheery Exchange Between 13-Year-Old Girl And Matt Damon Left Him Devastated | 0.09 |
| Clickbait | A Bunch Of Bullies Thought It Would Be Funny To Pummel This Kid. They Aren't Laughing Now. | 0.08 |
| Clickbait | When You Sexualize Men And Women Equally, It's Amazing How Much Fun You Can Have | 0.08 |
| Non-Clickbait | A Supreme Court Justice Takes Down An Anti-Gay Marriage Argument In 1 Minute | 0.07 |
| Non-Clickbait | A Man Slams A Bigoted Question Down So Hard He Brings Down The House | 0.06 |
| Non-Clickbait | EXPERIMENT: Stick A Gay Person And An Anti-Gay Straight Person In A Room And Ask Them To Hug | 0.06 |
| Non-Clickbait | A Comedian Tears Apart A Classic Anti-Gay Argument In A Hilarious 3 Minute Rant | 0.06 |
| Non-Clickbait | Understanding The Ugly Side Of Sexual Objectification | 0.06 |