Group Part

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

Content-based Recommendation System for Engaging Headlines from the Upworthy Research Archive

Guilherme Macedo Soares Gonzalez Bernardo

Work project carried out under the supervision of:

Professor Michael Kummer

03/03/2025

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 impacts 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 similar news recommendations, clickbait impact, headline category impact, and time trends impacting Upworthy's engagement.

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, Recommendation Systems, Natural Language Processing, Text Embeddings, Text Analysis, 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 & 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 & 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. For example, Kumar et al. (2020) utilized XGBoost to analyze user engagement on Twitter, demonstrating its effectiveness in classifying tweets based on their likelihood of being retweeted. For predicting engagement, studies demonstrate the effectiveness of these models regarding online user content. Since their findings indicated an enhancement in predictive accuracy.

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. Manning, Raghavan, and Schütze (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 adress
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 Hillary C. 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.

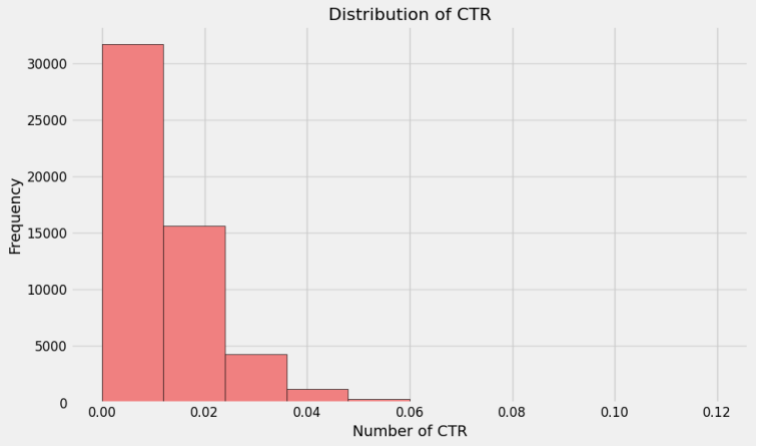

Figure 1 – Distribution of CTR

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, entertainment, weather, health, and autos.

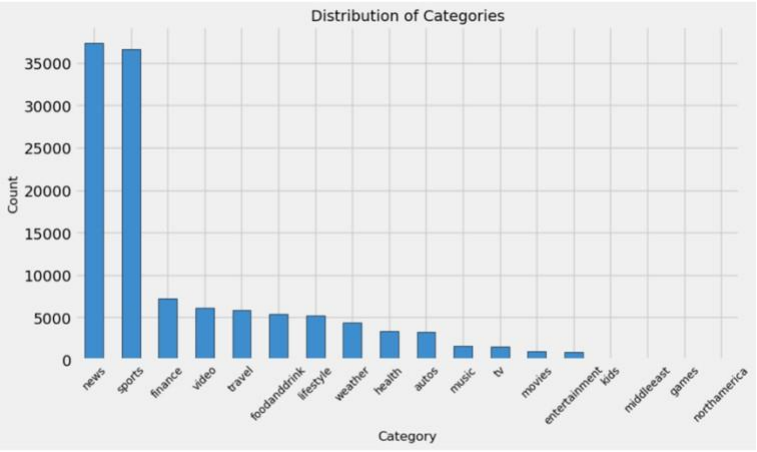

Figure 10: Distribution Categories MIND Dataset

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.

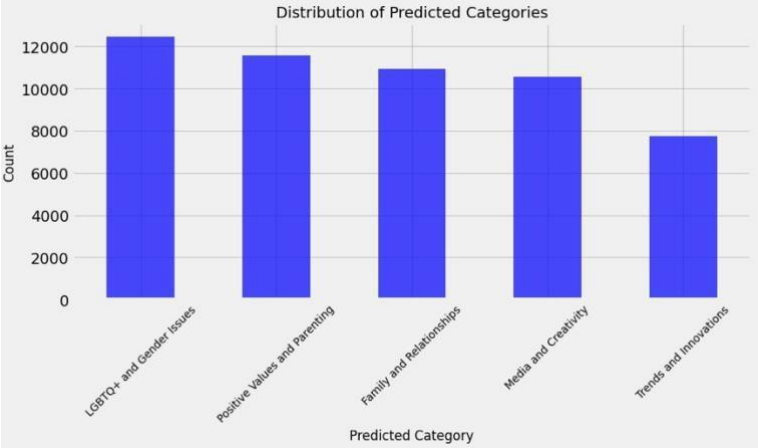

Figure 14: Distribution Categories Topic Modelling

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 *sckit-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 percentual 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.

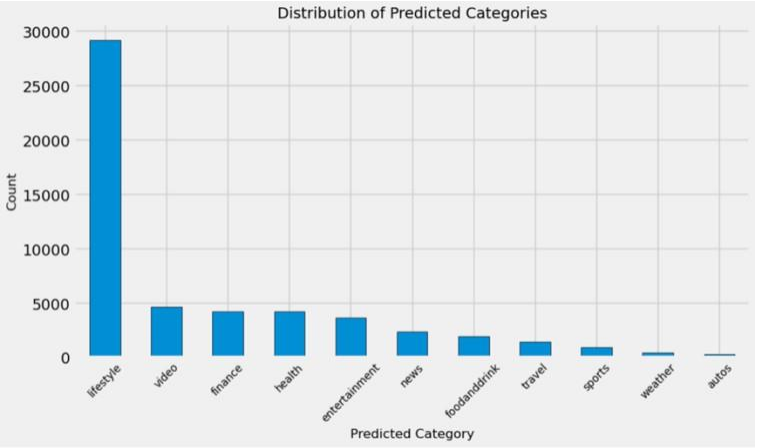

Figure 15 - Distribution Categories Upworthy Dataset

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.

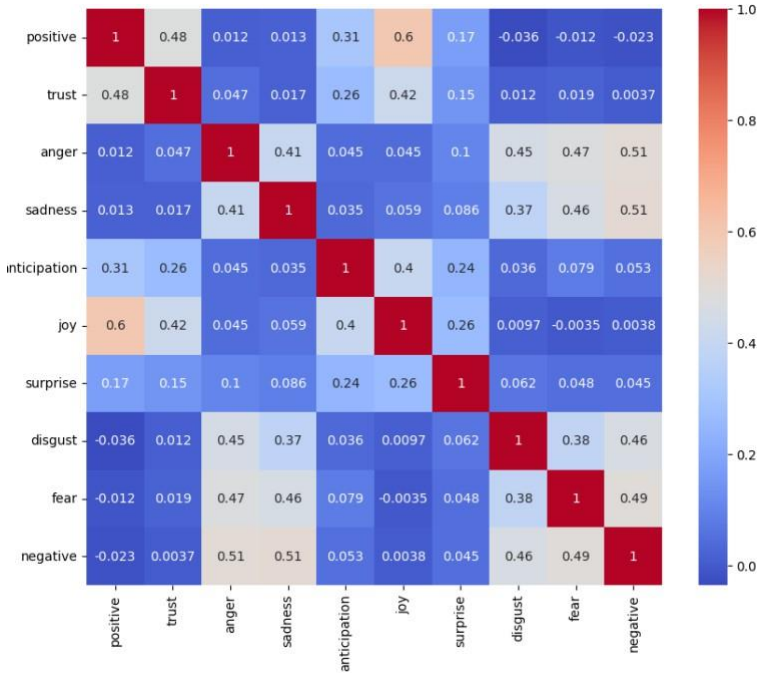

Figure 16 – Correlation Matrix between Emotions

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) $$High\ Engagement = Top\ 20\%\ CTR$$

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:

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

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 Content-based Recommendation System for Engaging Headlines from the Upworthy Research Archive

Making accurate suggestions provides value for management and operational insights, contributing to increased user engagement with the platform. In an article repository, is invaluable to have a recommendation system that provides relevant information regarding similar articles the reader might be interested in. As the repository's goal is to make engaging content, providing relevant suggestions can lead to an increase in the number of clicks, as well as captivate a larger audience to engage with the archive repository.

While traditional systems rely on user preferences, item-based, or collaborative-filtering approaches, this method focuses on the potential of embedding-based approaches to enhance recommendations. The features used have been mainly retrieved from the *headline* column, imposing a high dependency on the quality of this feature and on the methods previously used to retrieve insights from it.

By leveraging natural language processing techniques, this research uses methods previously tested by papers to focus on using clustering and similarity measurement techniques while addressing the challenge of defining relevance through statistically justifiable thresholds.

The system shows effectiveness by outperforming target articles more than half the time in CTR, identifying content with higher engagement potential. Emotional sentiment analysis reveals a slight increase in negative sentiments among recommended articles, suggesting a preference for emotionally intense content while recommended headlines show a small increase in word count. The results prove the system's capacity to enhance engagement through personalized, semantically relevant recommendations tailored to readers' emotional and semantic preferences, improving content discoverability, loyalty, and repeat visits, which will lead to an overall increase in clicks.

8.1.1 Data Cleaning

In the context of the Upworthy Research Archive dataset, A/B testing has been used to compare different approaches to presenting the same article. To mitigate bias in the recommendations made by the model, the dataset has been filtered, this time, to only contain the news that won the test and was published.

After having data only with "winning" articles, duplicates within the same test group have been removed, keeping the article with a higher CTR.

Moreover, a function is created that is used as preprocessing to clean unnecessary characters from the *headline* and *excerpt* features, standardizing and tokenizing them by splitting the content of the variables into words for further analysis and model input.

8.1.2 Feature Engineering

To create a model that makes recommendations based on semantic similarity, sentiment similarity and category, it is important to choose the right embedding method. As different embedding methods have different scales, the StandasrdScaler python library has been used to standardize results across the different methods tested. This way, it is easier to compare the performance of different embedding approaches.

Choosing the right embedding method can be a challenging task, as each approach can outperform the other according to a specific scenario. Term frequency-inverse document frequency (TD-IDF), Word2Vec, Global vectors for word representation (GloVe), Sentence Bidirectional Encoder Representations from Transformers (SBERT) and text-to-text transformers (T5) models have been compared to assess which one is the most adequate for the task proposed.

Alodadi and Janeja (2015) discuss TF-IDF's utility in extracting key features from textual data and measuring document similarity effectively using cosine similarity. This method is used as a baseline as it is computationally inexpensive and does not require

pretraining as the other models used ahead, making it ideal for experimentation. The paper highlights that coupled with cosine similarity, the model can achieve meaningful results when comparing the similarity between textual features, setting a foundational standard against more complex models.

Unlike TD-IDF, the computationally efficient Word2Vec generates dense vector representations of words based on their semantic relationships, handling synonyms. Despite this, provides static embeddings, meaning that each word has the same vector representation regardless of its context and doesn't capture word order or syntactic structure, which can be crucial for understanding sentence similarity. GloVe is a robust unsupervised word embedding model introduced for word representation that incorporates both global matrix factorization and local context windowing methods. Mohammed, Jacksi, and Zeebaree (2021) compare this model to Word2Vec, mentioning that usually outperforms it due to including co-occurrence statistics.

Reimers and Gurevych (2019) explain how Sentence-BERT focuses on adapting the BERT model to create high-quality sentence embeddings, being ideal for comparing semantic textual similarity. SBERT embeddings usually outperform raw BERT embeddings with a reduced computational cost without sacrificing accuracy. Ni et al. (2021) explore the use of T5 for generating sentence embeddings and demonstrating its effectiveness. The paper shows that it is a well-suited task for text similarity tasks, providing robust representation, even without fine-tuning.

To assess the performance of each model on computing text similarity through cosine similarity, five simple recommender system functions have been created, one for each embedding method mentioned. The formula of cosine similarity is:

$$Cosine\ Similarity = \frac{A \cdot B}{\| A \| \| B \|} \tag{3}$$

With A being the target vector and B representing the possible recommendation, the embedding vectors of any the other articles in the dataset. The function outputs the three articles with larger cosine similarity score for the target article as recommendations.

To compare the overall performance of each embedding method the metrics average cosine similarity and category match ratio have been calculated. These metrics have been chosen because there is no ground truth of what to consider as relevant recommendations.

It is also important to highlight that the embeddings values were standardized, making the average cosine similarity scores more comparable despite not being a perfect solution. While this measure becomes less influenced by the scale or distribution of the embeddings, other factors like cosine similarity focus on angles between vectors rather than magnitudes, and different approaches capture different kinds of relationships, like TF-IDF that focuses on term frequency and SBERT or T5 that capture semantic similarity on a deeper level, must also be considered and weighed when choosing the right option.

The average cosine similarity is being calculated by the mean value of the average cosine similarity of recommended articles for every target article.

$$(4) \quad Average\ Cosine\ Similarity = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{k}\sum_{j=1}^{k} Cosine\ Similarity(A_i, B_j)\right)$$

With k being the number of articles recommended per target, that is usually k = 3, and n being the amount of news in the dataset.

The category match ratio has also been assessed as an additional measure of recommendation quality. It evaluates how many times the recommended articles are part of the same article category as the target, serving as reference of semantic similarity.

$$(5) \quad Category\ Match\ Ratio = \frac{Total\ Number\ of\ Matches}{n * k}$$

With k being the number of articles recommended per target, that is usually k = 3, and n being the amount of news in the dataset.

Comparing the results of the models using the evaluation metrics mentioned above, GloVe is the worst-performing metric, followed by Word2Vec. The one that presents the largest cosine similarity between targets and recommendations is TF-IDF, closely followed by the two transformer-based methods, despite SBERT being the one with the highest category match ratio.

Table 12 – Embeddings performance

Model	Average Cosine Similarity	Category Match Ratio
TF-IDF	0.345452	0.396554
Word2Vec	0.254409	0.370370
GloVe	0.238841	0.289649
SBERT	0.331974	0.544160
T5	0.321998	0.413105

SBERT embeddings have been chosen to be added to the dataset as it presents the most balanced option between both metrics. It is also important to mention that TF-IDF is based purely on term frequency and may miss context, while the transformer that was selected is specifically designed to capture deeper semantic meaning. Consequently, when the difference in similarity scores is marginal (as it is here), SBERT might provide more meaningful semantic relationships due to its transformer-based embedding capabilities.

The embeddings column has been prepared in a way that the column was stacked into a single NumPy array. This has been done since methods like Principal Component Analysis (PCA), clustering algorithms, or similarity computation require input data in the form of a 2D array. Without stacking the embeddings into this structure, these operations would not work as intended.

The vectors generated by models like SBERT are often high-dimensional, leading to increased computational cost for clustering or similarity computations and noise in the embeddings space, where some dimensions carry minimal useful information. Using PCA on the embeddings feature, the dimensionality was reduced while retaining 95% of the variance in the original data.

For tasks like recommendations through clustering, having uniformization between variables is essential. The *predicted_category* has been converted to numeric data type using a label encoder, ensuring each unique category is assigned a numeric value. Then, both the category and the sentiment columns (positive, negative, joy, disgust, sadness, anticipation, anger, surprise, trust, fear) were standardized.

As the embeddings had already been normalized previously, all the features that the system is using are on a similar scale. A *combined_features* matrix is obtained when merging the category, embeddings, and sentiments.

8.1.3 Recommender System

The methodology for implementing the recommendation system has drawn inspiration from the approach presented by Ahuja, Solanki, and Nayyar (2019), where clustering, using k-means, and nearest neighbor techniques were employed to enhance recommendation accuracy.

The data has been split into train and test using the *train_test_split*, while distributing the categories in a stratified way due to the skewness of the dataset towards some categories, as it can be seen in Table 12.

To decide how many clusters to choose for the k-means clustering Elbow Method, Silhouette Score, Calinski-Harabasz Score and Davies-Bouldin index have been used for validation of the choice. Despite the Elbow Method, seen in Figure 21, suggesting 5 clusters, it does not directly account for inter-cluster separation and may overestimate clusters. The other three methods, on the other hand, all suggested choosing 3 clusters as the optimal choice. The Silhouette Score shown in Figure 22 has had the best score, of 0.26, with 3 clusters. The Calinski-Harabasz Score observed in Figure 23 has its best results with 2 and 3 clusters, with scores between 470 and 500. And the Davies-Bouldin Index shown in Figure 24 has its lowest score, 1.60, using 3 clusters. Therefore, the optimal number chosen has been 3 clusters.

To check if the clusters are properly separated, a visualization with PCA has been done.

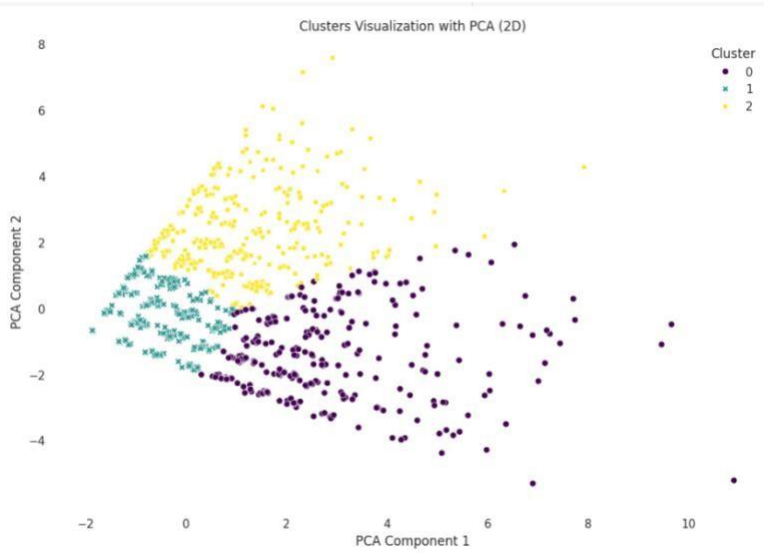

Figure 25 – PCA with 3 clusters

Cluster 0 appeared to be well-defined and compact. It had a clear boundary between Cluster 1 and Cluster 2. Cluster 1 had a partial overlap with Cluster 2 but still relatively distinct. The visualization suggests Cluster 0 as the most cohesive, while Clusters 1 and 2 distributions could indicate that these two groups might share similarities in some features.

To have a better understanding of how clusters have been distributed, the t-SNE allowed visualization in a 3D-generated space, as can be seen in Figure 26. The plots indicate a significant separation, although some groups are not fundamentally different. Some data points seem to be closer to other clusters' centroids, yet k-means recomputes proximity several times, and metrics indicate 3 clusters as the optimal number to ensure maximum separation. It is also important to be aware that when reducing from a high-dimensional space to 3D, t-SNE compresses information, which may distort how the data is represented.

The recommendations are then generated by finding similar articles within the same cluster using k-Nearest Neighbors. Firstly, the target article is chosen, and the model locates it and identifies the target cluster. Then, using the Euclidean distance metric, the model finds the three closest neighbors to the target article based on the metrics result.

8.1.4 Evaluate Results

To properly evaluate a recommendation system, there is either a baseline truth of what is relevant to the target item/user or an online system that can test and observe the performance. In this case, as there is no ground truth, defining relevance is done through a similarity threshold, which is essential for the performance of the system to be assessed. The most appropriate measure to set as the threshold for further evaluation is cosine similarity since it is the metric that best captures semantic similarity.

The threshold reflects the minimum level of similarity required for two articles to be considered related in terms of sentiment. To avoid bias when choosing this value, this study adapts a clustering evaluation strategy inspired by Harakawa et al. (2019), which has shown effectiveness in multimedia content analysis. The paper highlights that sentiment similarity thresholds are context-dependent and require empirical testing.

As in the paper, thresholds are iteratively tested, ranging until 1 in 0.05 intervals. In this study, the threshold determines whether a recommended article is similar enough to the target based on embedding similarity. Unlike the paper, the ground truth is not predefined, so it is generated dynamically using the clustering results and threshold similarity.

Performance has been evaluated through the F1-score, balancing precision (relevance of recommendation) and recall (coverage of ground truth). Precision measures the fraction of relevant recommendations among all recommendations made by the system, which means it tells how many suggestions are relevant. Recall focuses on completeness and reflects how the system performs when trying to find all the relevant suggestions. The highest F1-score is used as the optimal threshold as it assesses both relevance and coverage, ensuring none is prioritized over the other and avoiding bias.

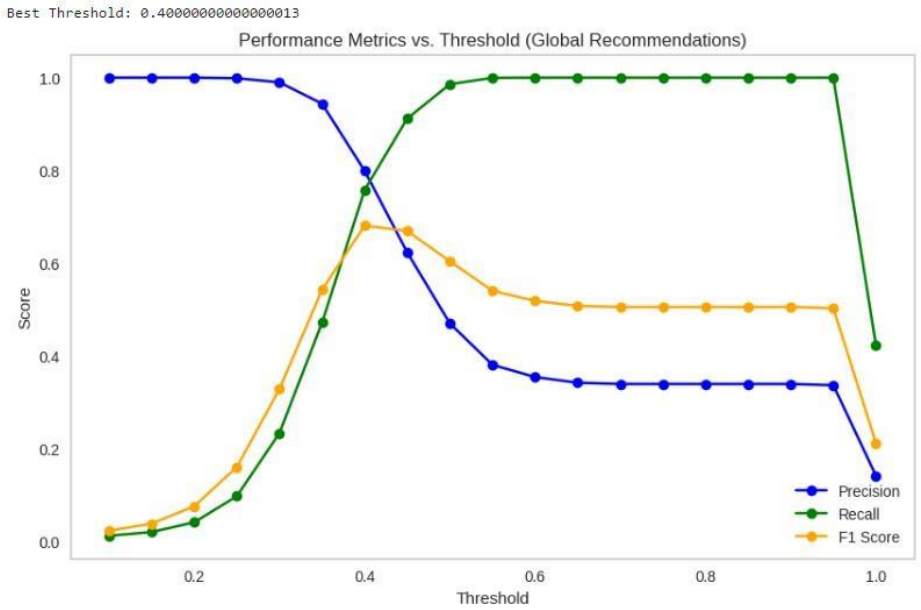

Figure 27 - Testing thresholds for optimal F1-score

The threshold of relevance has been set to a minimum cosine similarity of 0.4, having the highest F1-score of the overall recommendations of 0.6807, precision of 0.7988, and recall of 0.7576.

Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (nDCG) metrics have been used to evaluate the quality of the recommendations at a threshold of 0.4. MAP measures the quality of the recommendations by averaging the precision values at different cutoff points for all relevant items. The MAP score obtained of 0.9814 indicates an extremely high value, suggesting that the system is performing well in identifying relevant items. The nDCG is indicative of the quality in ranking the recommendations. A score of 0.9251 means that the system ranks the most relevant articles near the top with a high degree of accuracy.

Moreover, the same baseline truth of a cosine similarity set on 0.4 has been used in a cluster-restricted scope that can be observed in Figure 28. The F1-score value decreased to 0.6010 as precision also decreased to 0.5705, while recall stayed roughly the same, 0.8167, compared to the global model performance. This has happened because, in this scope, the ground truth only included entries of the same cluster, making the pool of relevance to a subset

of the dataset while making the number of false positives increase. This is normal since the cluster-restricted scope has a narrower definition of relevance.

8.1.5 Limitations and Future Research

The main limitation found has been the lack of user interaction information on the data that was provided, relying heavily on content attributes to make recommendations, limiting the possibility of personalized recommendations, and validating emotional triggers.

Moreover, in this analysis, there was a need to artificially create relevance of recommendations between different articles. Having a predefined ground truth that indicated how different articles relate to each other would significantly increase the confidence in the results gathered.

Furthermore, creating clusters of articles imposes an assumption of uniform audience behavior that may oversimplify what each user prefers. This can be related to the static nature of the recommendations that, without feedback, cannot adapt to the fast-changing domain of news articles.

Finally, the features used to cluster the articles are all dependent on the quality of the headline and excerpt since they have been extracted from it. If these features do not represent the articles properly, the recommendations could become irrelevant as the data does not have information about what the article mentions besides these two columns.

In a future scenario, not accounting for all the stated limitations, moving from a cluster-based approach to a personalized user recommendation with lick behavior, time spent, and scrolling pattern while incorporating reinforcement learning techniques to make real-time changes according to changing preferences could significantly increase the quality of the recommendations. Creating feedback on the quality of the recommendation through ratings or A/B testing to assess how different suggestions impact similar user profiles could be highly

beneficial to understanding how the target audience reacts. This could have also been a metric that might be worth exploring to increase engagement.

8.1.6 Conclusion and Business Impact

The findings validate the system's capacity to make semantically relevant recommendations. Suggested articles outperform in CTR the target they are being compared to 55.08% of the time, indicating that the system is effective in identifying entries with a higher engagement potential. There is also a slight improvement between the absolute difference in the average target CTR, 0.0220, vs the Mean recommended CTR, 0.0223, indicating a better performance of the suggested news. In terms of sentiment performance, the positive articles marginally decrease in the recommended articles while the negative sentiments increase, indicating that the recommendations usually lean towards emotionally intense content. The word count of recommended headlines has a small increase across clusters.

These results support the overall goal of this paper of identifying emotional and psychological triggers that drive engagement. The improvement in CTR validates the system's capacity to suggest highly engaged news articles. The emotional pattern observed reveals the triggers that are most likely to resonate with Upworthy's users, providing evidence-backed guidance for content creation. Variations in headline length and emotional preferences are examples of cluster-specific insights that underscore the significance of customizing content for audience segments and the role of sophisticated analytics in engagement tactics.

This system creates the possibility to increase engagement with the news platform. Having tailored recommendations supported by personalization from semantics and sentiment while improving content discoverability by making suggestions of articles that the reader could not have found for itself and consistently recommending content that aligns with a reader's emotional and semantic preferences at that moment, the system can build loyalty and repeat visits.

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 analysis.

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 that 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 insights 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 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

Ahuja, Rishabh, Arun Solanki, and Anand Nayyar. 2019. "Movie Recommender System Using K-Means Clustering and K-Nearest Neighbor." 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Greater Noida, India, 263- 268.

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.

Alodadi, Mohammad, and Vandana P. Janeja. 2015. "Similarity in Patient Support Forums Using TF-IDF and Cosine Similarity Metrics." In *2015 InternationalConference on Healthcare Informatics*. Baltimore: University of Maryland Baltimore County.

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.

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. 2024. "Accuracy vs. Precision vs. Recall in Machine Learning: What's the Difference?" *Evidently AI.* Accessed November 25. 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.

Harakawa, Ryosuke, Shoji Takimura, Takahiro Ogawa, Miki Haseyama, and Masahiro Iwahashi. 2019. "Consensus Clustering of Tweet Networks via Semantic and Sentiment Similarity Estimation." IEEE Access 7: 116207–116217. https://doi.org/10.1109/ACCESS.2019.2936404.

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.

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

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.

Kumar, A., A. Singh, and R. Singh. 2020. "Predicting User Engagement in Social Media Using Machine Learning Algorithms." *Journal of King Saud University - Computer and Information Sciences*.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval.* 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.

Mohammed, Shapol M., Karwan Jacksi, and Subhi R. M. Zeebaree. 2021."A State- of-the-Art Survey on Semantic Similarity for Document Clustering Using GloVe and Density-Based Algorithms." *Indonesian Journal of Electrical Engineering and Computer Science* 22, no. 1: 552–562. https://doi.org/10.11591/ijeecs.v22.i1.pp552-562.

Ni, Jianmo, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, DanielCer, and Yinfei Yang. 2021. "Sentence-T5: Scalable Sentence Encoders from Pre-

trained Text-to-Text Models." arXiv. December 14, 2021. https://arxiv.org/abs/2108.08877.

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

Reimers, Nils, and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. 2019. " *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3982–3992. Hong Kong, China: Association for Computational Linguistics, https://aclanthology.org/D19-1410/.

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.

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.

Sabia, Séverine, Michael Marmot, Carole Dufouil, 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, no. 3: 228–35. https://doi.org/10.1016/S2213-8587(13)70192-X.

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.

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.

Appendix:

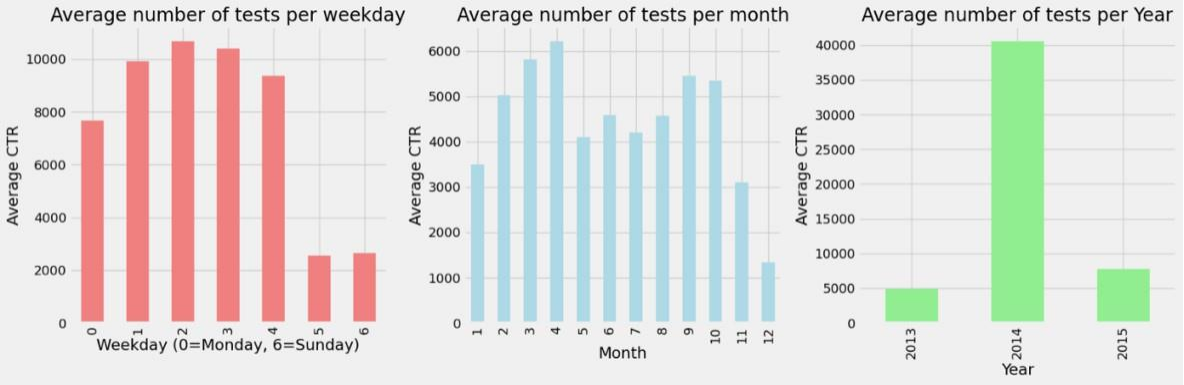

Figure 2 – Average CTR per Average Number of Tests per Weekday, Month and Year (Created_at)

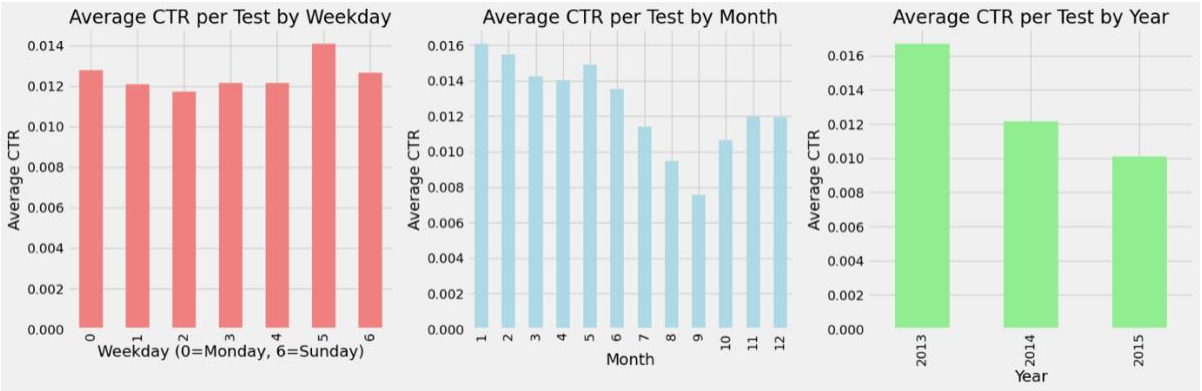

Figure 3 – Average CTR per Test by Weekday, Month and Year (Created_at)

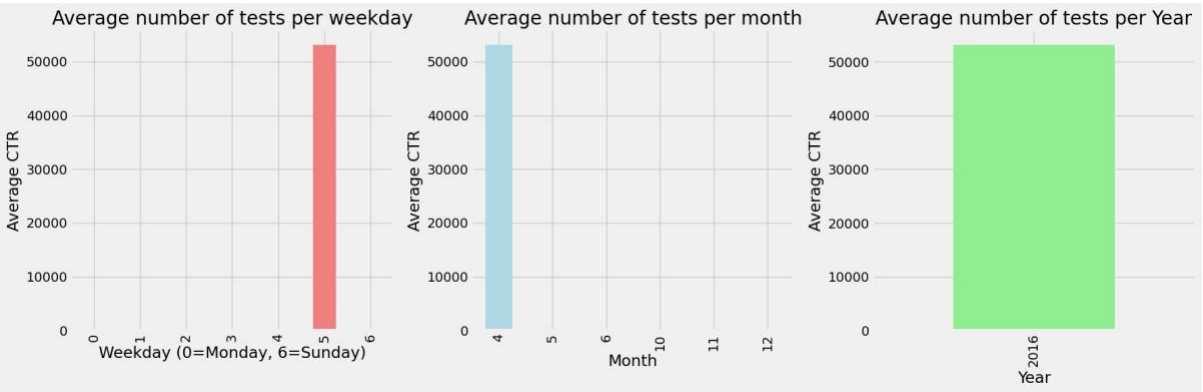

Figure 4 – Average CTR per Average Number of Tests per Weekday, Month and Year (Updated_at)

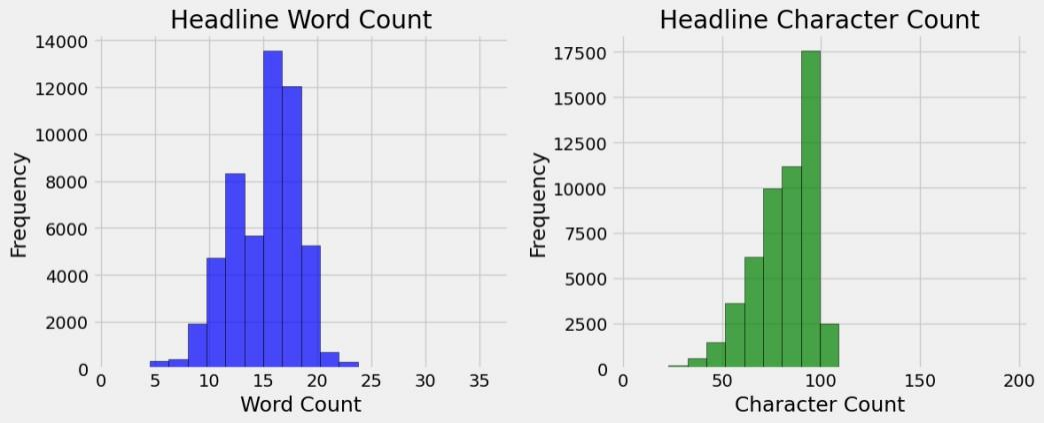

Figure 5 – Distribution of Headlines Word's and Character's

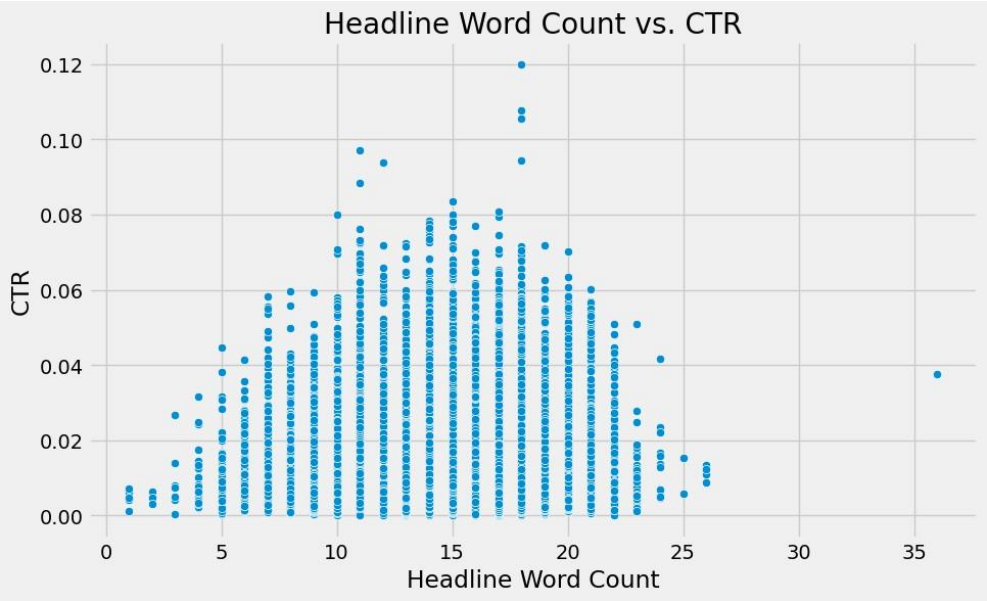

Figure 6 – Headline Word Count versus CTR

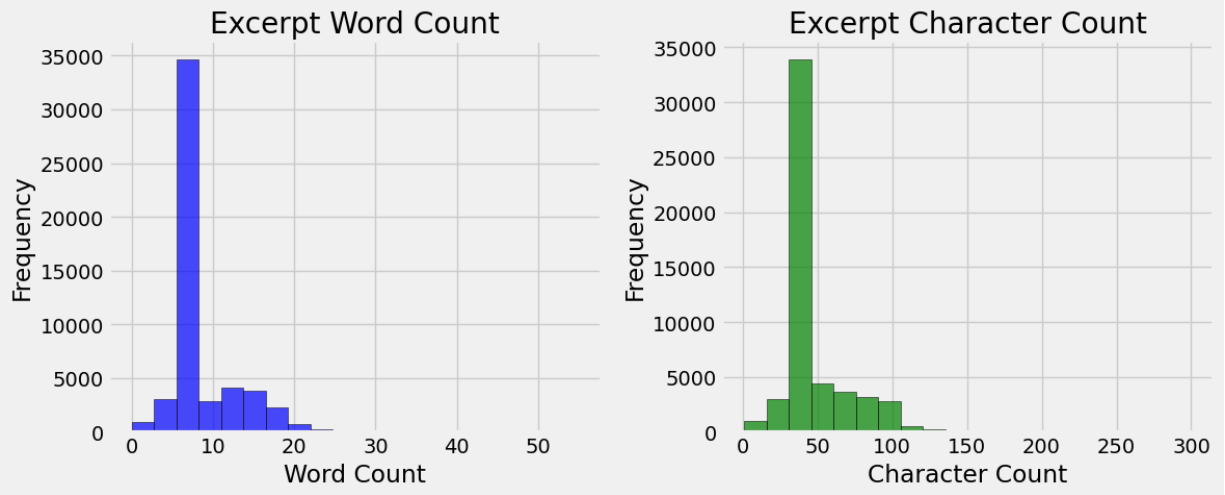

Figure 7 - Distribution of Excerpt's Word and Characters

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
Null	38
The next time someone tells you gay marriage w...	37

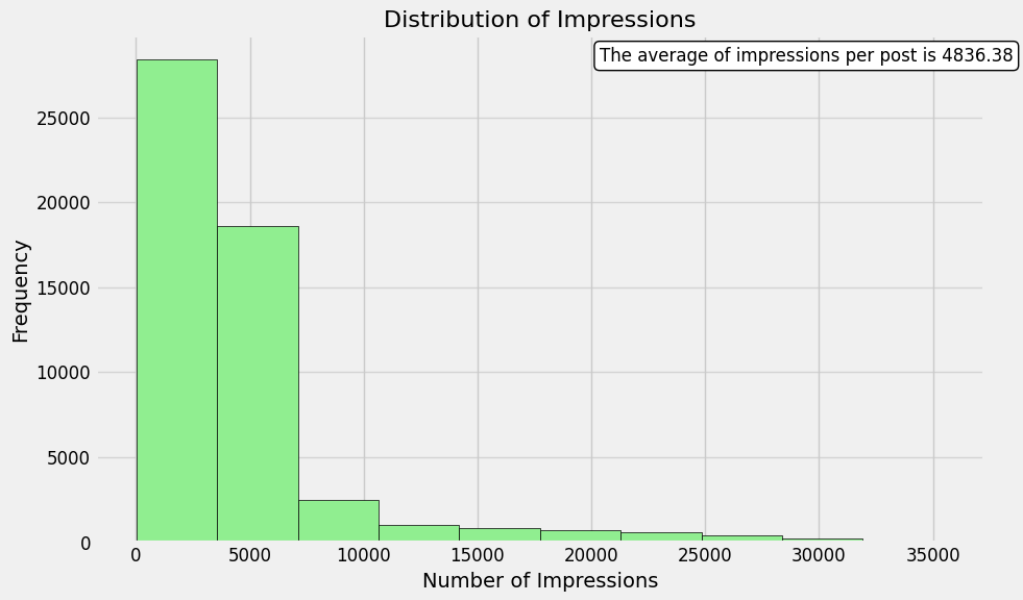

Figure 8 – Distribution of Impressions

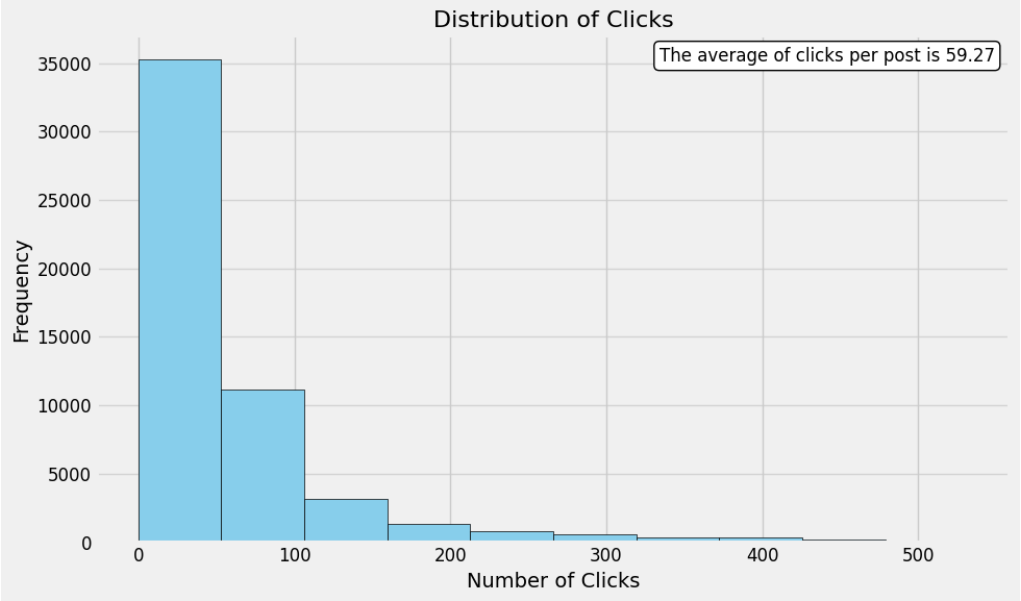

Figure 9 – Distribution of clicks

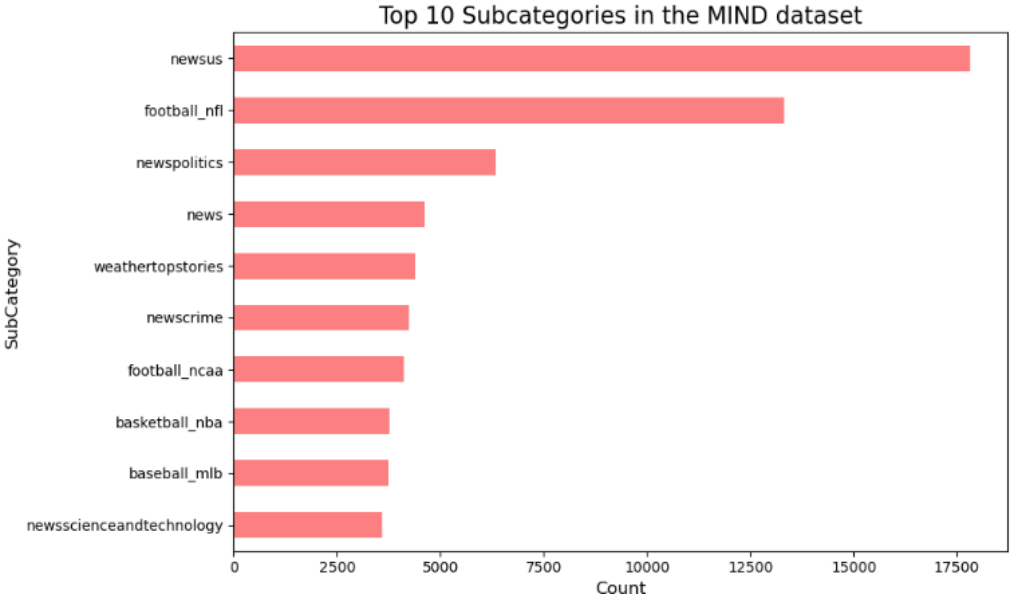

Figure 11: Distribution SubCategories MIND Dataset

Figure 12: - Word Cloud for Titles

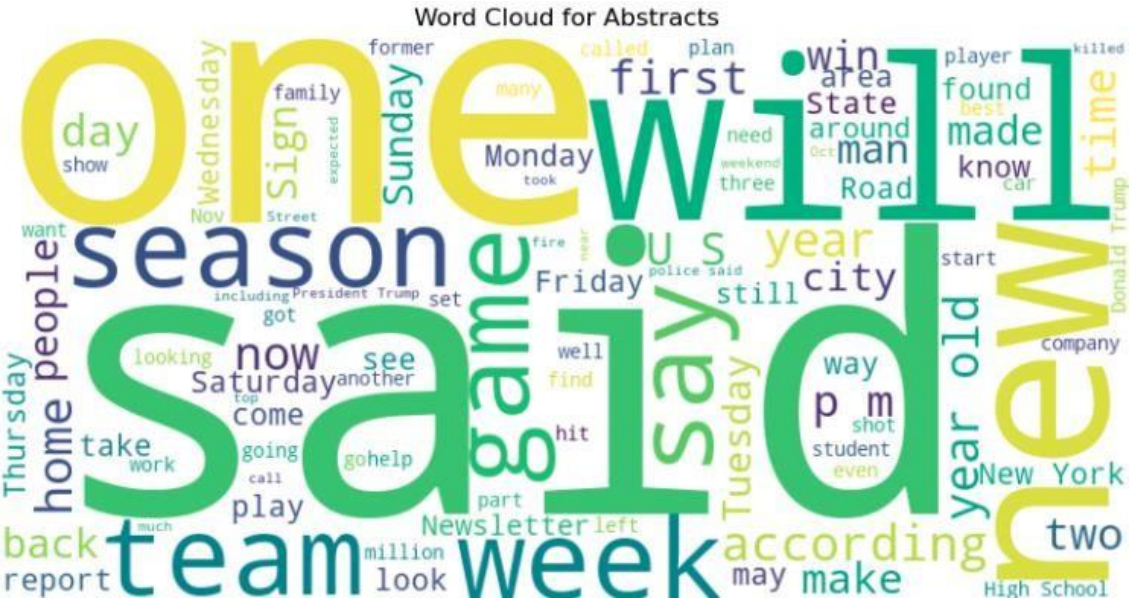

Figure 13: - Word Cloud for Titles

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

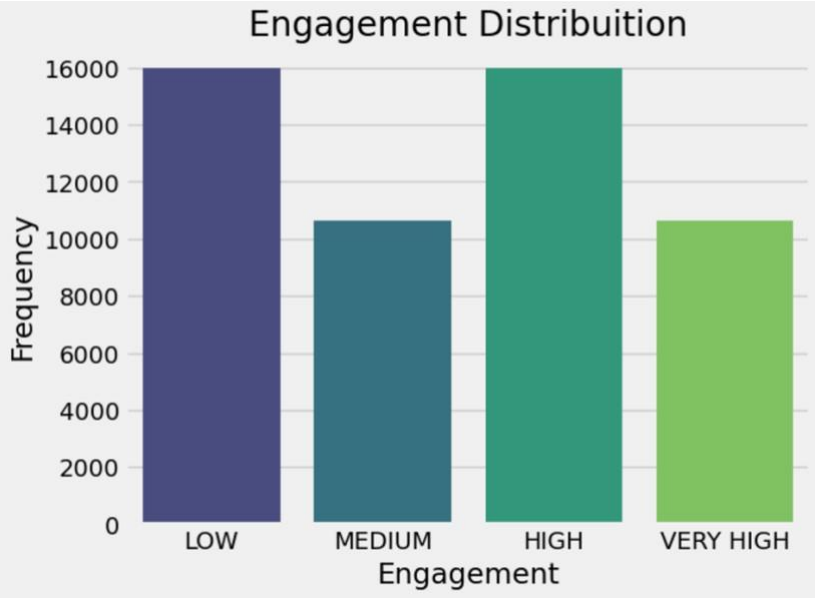

Figure 17 – Distribution of engagement categories

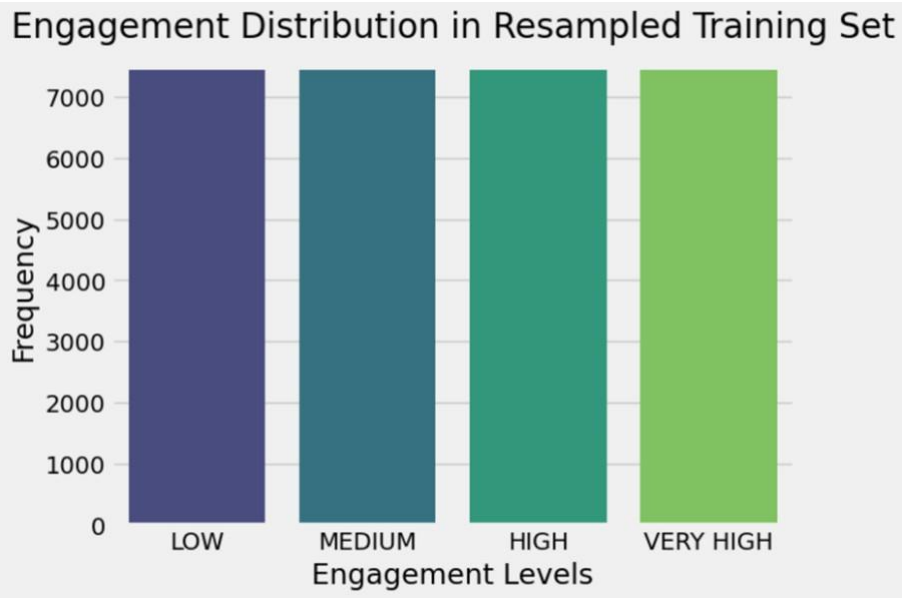

Figure 18 – Distribution of Engagement Levels in Resampled Training Set

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_entertainement	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
Iierations	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_strenght	10	Controls the randomization strength to improve model robustness and generalization

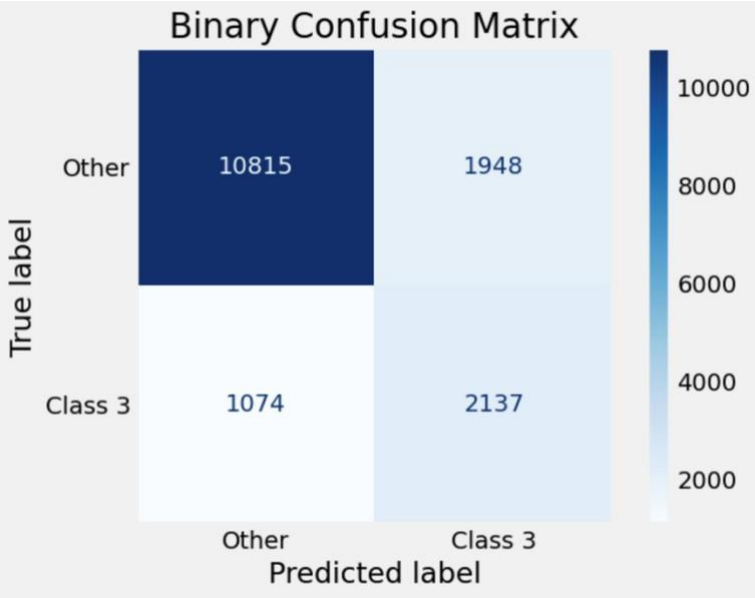

Figure 19 – CatBoost Confusion Matrix

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%

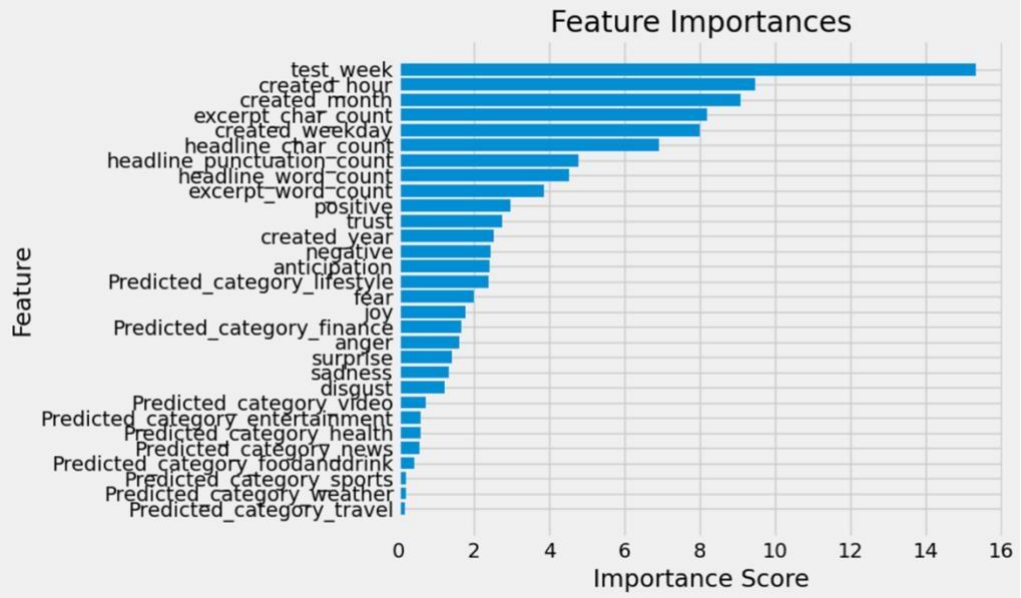

Figure 20 – Distribution of Features Importance

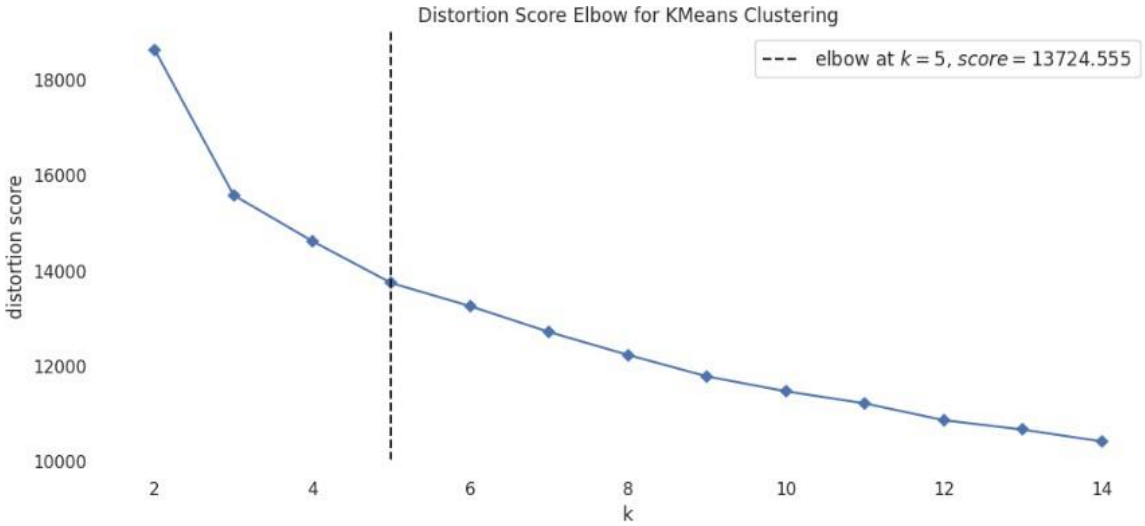

Figure 21 – Elbow method results

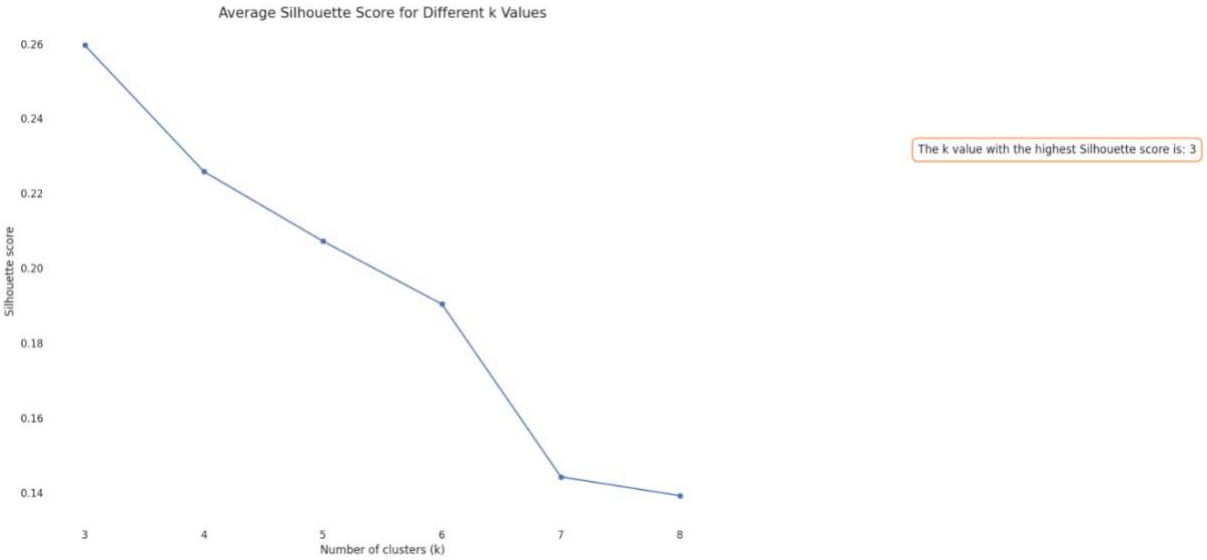

Figure 22 – Average Silhouette Score for Different k Values

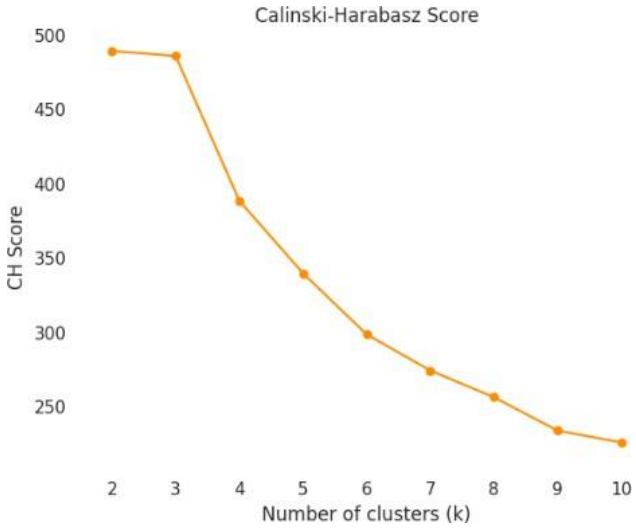

Figure 23 – Calinski-Harabasz Score results

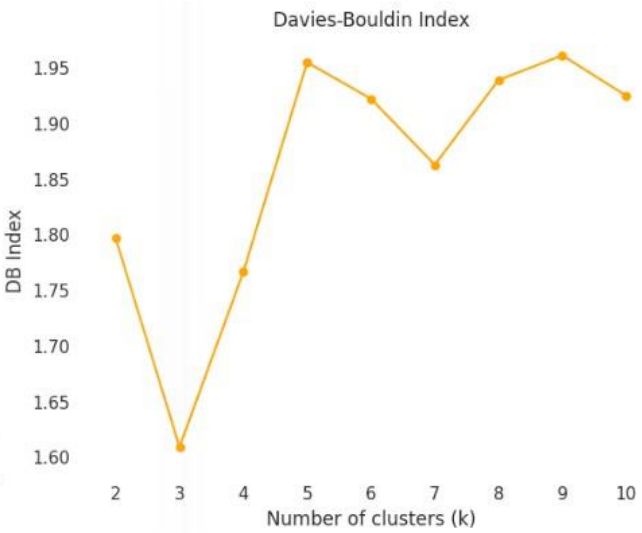

Figure 24 – Davies-Bouldin Index

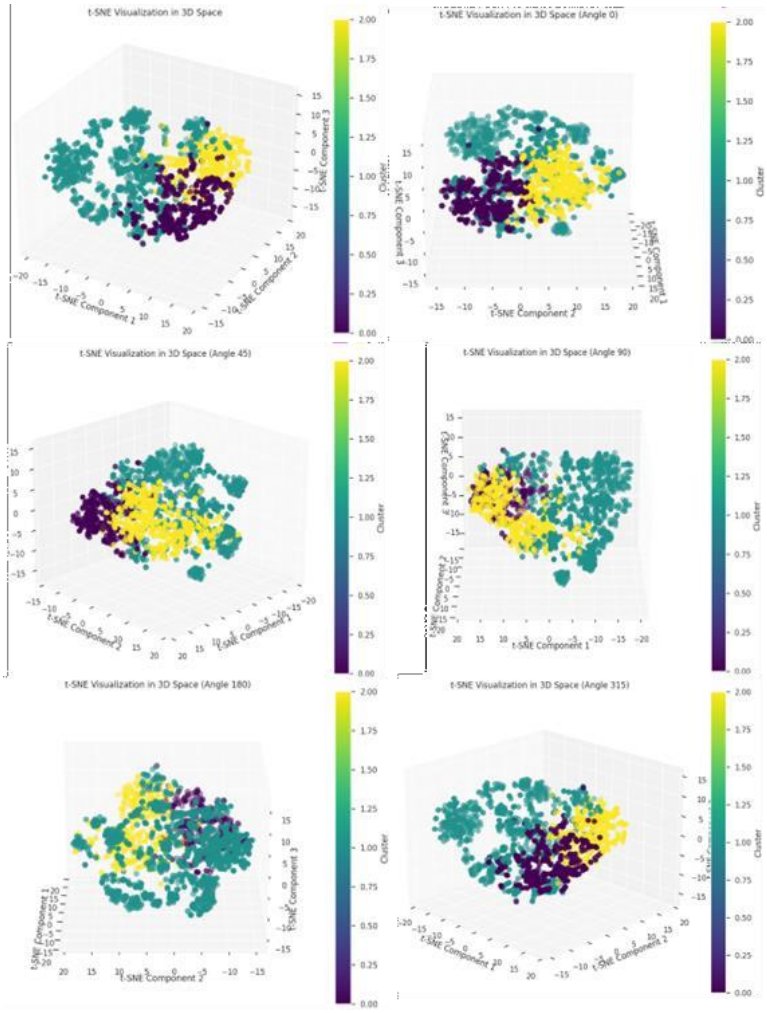

Figure 26 – Visualization of t-SNE for 3 clusters

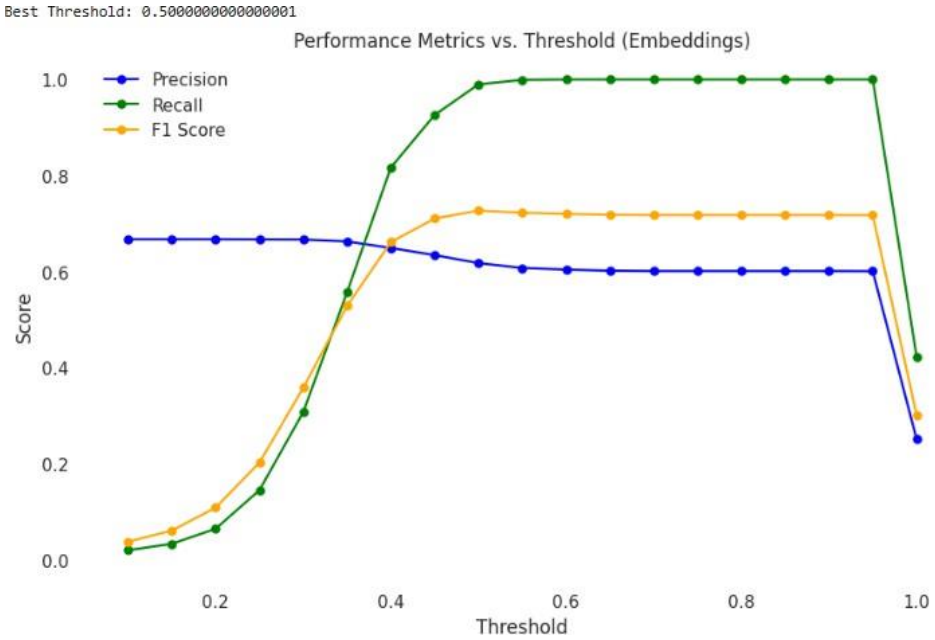

Figure 28 – Testing thresholds in a cluster-restricted scope to find optimal F1-score