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.

BEYOND THE IMAGE:
THE ROLE OF TEXT IN SHAPING MEME ENGAGEMENT ON REDDIT

MARTA DINIS

Work project carried out under the supervision of:

Michail Batikas

22/01/2025

Abstract

The study explores how textual elements in memes and titles influence engagement on Reddit. Using OCR, NLP, and sentiment analysis, it examines text presence, length, and sentiment to assess their impact on score, comments, and awards. Posts from r/memes are analyzed through regression models and machine learning techniques to identify key textual drivers of meme virality. Findings offer insights into how linguistic features shape engagement, providing a framework for understanding text's role in meme success. This research aids content creators and marketers in optimizing engagement strategies by leveraging textual components in meme design.

Reddit, Memes, NLP, OCR, Regression

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

Memes have become a significant medium for spreading cultural, social, and political ideas, effectively engaging online audiences through humour, relatability, and adaptability. Understanding what makes a meme successful—often measured by views, likes, and shares—has become essential knowledge, particularly in advertising, political communication, and online content moderation (Barnes et al. 2021). With its topic-focused subreddits and interactive community dynamics, Reddit offers a unique platform for studying meme trends. Its voting and ranking systems amplify posts based on user preferences, enabling memes to spread rapidly and evolve. This dynamic environment provides valuable insights into user engagement across diverse subcultures and contexts (Medvedev, Lambiotte, and Delvenne 2019). As of December 2023, Reddit boasted an average of 73.1 million daily active users and 267.5 million weekly active users, nearly half of whom reside outside the United States. This broad and diverse user base significantly enriches data on meme trends and user interactions with various topics and communities (Statista 2024).

Advances in artificial intelligence, particularly in computer vision and generative AI, have opened new pathways for analysing meme success. Computer vision techniques can identify visual features such as colour patterns, facial expressions, and object presence, which are critical to assessing memes visual appeal and impact (Barnes et al., 2021). Generative AI models like DALL-E contribute to this research by generating images from textual descriptions, enabling a more profound exploration of the synergy between visual and textual meme components (OpenAI 2021).

Previous research highlights how sentiment analysis and machine learning approaches enhance the evaluation and prediction of meme engagement on Reddit. By identifying the emotional tone—such as humour, sarcasm, or motivation—present in memes and comments, sentiment analysis provides valuable insights into user reactions. For instance, one study categorised

memes based on their overall sentiment, showing how this approach helps understand user responses (Hossain et al. 2022). Furthermore, machine learning methods, particularly those integrating multiple data sources, have proven useful for predicting meme virality. For example, a framework combined visual and textual features to determine sentiment across different content types, demonstrating how these technologies collectively forecast meme engagement (Pranesh and Shekhar 2020). Studies reveal that factors like image dimensions, brightness, and text density significantly influence meme popularity on platforms like Reddit (Barnes et al. 2021). These findings underscore the necessity of holistic analysis, as meme success depends on its message, visual design, and contextual relevance.

Predicting meme and comment virality remains challenging, mainly due to the complexity of aligning text and image data with their distinct structures and meanings. Recent advancements in deep learning and computational power have enabled sophisticated analyses that consider both data types, advancing the field of meme research (Ford, Krohn, and Weninger 2023). The highly skewed distribution of meme success further complicates machine learning tasks, necessitating specialised data preprocessing and evaluation techniques to address minority trends without overrepresenting majority classes (Shyalika, Wickramarachchi, and Sheth, 2024). Incorporating rare event prediction methodologies into multimodal AI frameworks holds promise for enhancing the accuracy of meme engagement classification.

Understanding meme dynamics has applications beyond academia, benefiting brands that use memes in advertisements to connect authentically with younger audiences and political campaigns that leverage meme culture to shape opinions and disseminate impactful messages (Ford, Krohn, and Weninger, 2023). Platforms like Reddit have become vital arenas for political and cultural discourse. For example, subreddits like r/PoliticalCompassMemes use humour to critique and discuss ideological perspectives (Colacrai et al. 2024). Additionally, insights into meme success can inform more effective digital communication strategies, as memes are

compelling tools for community building and cultural expression on social media (Nissenbaum and Shifman 2018).

This research adopts a multifaceted approach, integrating computer vision, generative AI, sentiment analysis, and machine learning to examine engagement with memes and comments on Reddit. Using historical data from r/memes, including actual meme images, this study aims to identify key drivers of engagement, enhancing the understanding of meme virality. By analysing the unique features of digital memes—adaptability, emotional resonance, and community-driven evolution—this paper seeks to quantify and model their success based on visual content and contextual data, offering novel insights into the dynamics of meme culture in the digital age.

2. Literature Review

2.1. The Theoretical Foundation of Memes

The idea of a meme was first proposed by evolutionary biologist Richard Dawkins in his work *The Selfish Gene*. Dawkins (1976, 248-249) suggested that evolution occurs because of the survival of self-replicating entities. In the context of biological evolution, the gene is the leading replicator, passing from one generation to the next and driving natural selection. However, the existence of another type of replicator was proposed, the meme, which he defined as a "unit of cultural transmission or a unit of imitation". Etymologically, it is derived from the Greek word *mimeme*, meaning “imitation", and was deliberately shortened to resemble "gene." Dawkins also intended the term to evoke memory, emphasising the role of replication and retention in cultural transmission. This analogy underscores the parallel between genetic and cultural evolution: just as genes propagate through biological organisms, memes spread through human minds, influencing cultural development through imitation.

Although Dawkins' analogy between genes and memes laid the foundation for the study of cultural transmission, it has also faced significant critique for being, according to some, overly

simplistic. For instance, cognitive anthropologist Dan Sperber challenged the notion of memes as replicators like genes in the early 1990s. The author introduced the metaphor epidemiology of representations to compare the spread of ideas and cultural elements to the transmission of diseases within a population (Ananth 2001). Much like diseases are passed between individuals and often mutate, memes are shared and modified during the process of communication. For Sperber, Dawkins's definition failed to account for how individual interpretation and social contexts actively transform these cultural elements during transmission rather than simply replicating them (Caton 1997, 319).

Sperber's critique is just one of the responses to Dawkins' theory. Other scholars, such as Susan Blackmore and Daniel Dennett, chose to expand on Dawkins' ideas rather than challenge them. In *The Meme Machine*, Blackmore supported the concept of memes as true replicators but delved deeper into their influence on human evolution. The author elaborated on the idea of *meme-gene co-evolution*, which suggests that the human brain evolved, at least in part, to become more efficient at imitating and spreading memes (Blackmore 2001, 244). This is because individuals who were better at sharing knowledge and adapting quickly through cultural learning would have gained a survival advantage.

Additionally, the philosopher and scientist Daniel Dennett argued that memes are critical in shaping human consciousness and its sense of self. As individuals encounter memes throughout life, these ideas and behaviours are adopted, either consciously or unconsciously, influencing actions, interactions with others, and the definition of personal identity. Dennett believed that, just as genes shape physical traits, memes shape one's identity (Block 1993).

2.2. Memes in the Digital Era

While the foundational ideas proposed by these authors have guided early discussions around cultural transmission, many of these perspectives were developed before the rise of the modern internet. As a result, there are core ideas that still hold, but they do not fully capture the nuances

of digital memes. Therefore, it is important to explore more recent definitions and frameworks that adapt and add to previous concepts by contextualising them within today's online culture. Limor Shifman, a key figure in defining the concept of internet memes, recontextualises Dawkins' original ideas of longevity, fecundity, and copy fidelity for the digital age (Shifman 2014, 17). The author explains that the internet increases longevity by allowing memes to be archived and rediscovered over time, enhances fecundity by enabling the widespread sharing of memes across platforms, and boosts copy fidelity by making it easier to replicate memes with precision while still allowing for remixes. Shifman also adds to her definition of a meme as a "group of digital items sharing common characteristics", which include content (the ideas conveyed), form (the visual format), and stance (the tone expressed) (Shifman 2014, 41-47). The "Distracted Boyfriend" meme is a well-known example. The form remains the same - a man looking away from his girlfriend - but the content changes based on the labels applied to the characters, such as "Me," "My Responsibilities," and "Procrastination." The stance is usually humorous or sarcastic, although it can vary depending on the message. Shifman also reinforces the idea that memes are not isolated entities by explaining that they are "aware of each other", meaning they remix and reference previous memes. For Shifman, this meme characteristic is fundamental as it differentiates them from viral content, which spreads as identical copies without significant alteration (Shifman 2014, 56-58). In contrast, memes are inherently collaborative and iterative, constantly evolving through user participation, a process greatly facilitated by the internet - something that will be discussed further in the next paragraph.

The emergence of *Web 2.0* transformed the web from a static to an interactive ecosystem, where users became active participants rather than passive consumers. This shift was largely driven by the rise of user-generated content, which empowered individuals to create and share their content, such as memes, across various platforms (Walther and Jang 2012). These platforms

have low barriers to entry, requiring only an internet connection and basic digital literacy, and are often free or low-cost to use (Elkin-Koren 2010). This combination of affordability and accessibility enables users to easily engage with content, thus fostering a highly participatory and interactive online culture that allows memes to evolve dynamically. Memes inherently rely on a dual process of replication and modification, with users copying but also remixing core elements such as images, text, or humour to adapt them to various cultural and social contexts. Therefore, the participatory culture of digital platforms plays a fundamental role in shaping how memes evolve and spread as they serve as the primary environment for meme proliferation.

Among these platforms, Reddit stands out as a central hub for meme creation and circulation, making it an ideal platform for analysing meme dynamics in this research. Reddit is a heterogeneous, crowd-sourced news aggregator and online social platform, originally self-declared as "the front page of the internet" (Eldridge 2024). Launched in 2005, it has consistently attracted over 1 billion monthly visits worldwide since 2023 (Bianchi 2024). Reddit's core structure is built around subreddits, which are interest-based communities that operate independently with their own rules and moderators, where users can post, comment, and engage with content specific to those topics. Posts are subject to a voting system, where upvotes and downvotes determine the visibility and ranking of content, with the most engaging posts rising to the top of the feed. Comments follow a tree-like structure, encouraging in-depth discussions within each post (Medvedev, Lambiotte, and Delvenne 2019). Reddit's large user base and open access make it a rich resource for research (Cauteruccio et al. 2022) across diverse fields, including computer science, health, and social sciences (Proferes et al. 2021). Furthermore, its subreddit structure enables researchers to conduct detailed, topic-specific studies within focused communities, making it easier to analyse niche areas of interest.

2.3. The Role of Memes

With Reddit established as a valuable platform for research, it is equally important to consider

why memes themselves warrant academic attention and, thus, explore their broader significance in the digital age, namely as a dominant form of communication in modern culture. Memes are highly effective in conveying messages because they combine visuals with short, impactful text, making them easy to consume (Holland 2020). This characteristic is particularly valuable in an era where attention spans are shrinking, as memes offer a fast and accessible way to communicate ideas or emotions in a format that is easily absorbed (Fillmore 2025).

Moreover, memes evoke universal emotions, ranging from joy to frustration, that resonate with a wide range of audiences. Their emotional relatability strengthens the connection between the content and the user, making memes an effective channel for conveying emotions. When a meme captures emotions that resonate with personal or collective experiences, it becomes more likely to be shared (Wagener 2021). An example of this emotional connection can be seen during the COVID-19 pandemic, where memes were widely used as coping mechanisms. According to a study on coronavirus-related memes from a dedicated subreddit, these memes allowed individuals to process and share their emotions, helping them cope with the anxiety and uncertainty of the global crisis through humour and creativity (Glăveanu and De Saint Laurent 2021).

Memes also play a significant role in community building and identity formation. Online communities, such as the r/DankMemes subreddit, use memes as a form of in-group communication, where shared cultural references and humour create a sense of belonging and mutual understanding among members (Newton et al. 2022). This shows that users not only relate to the content of memes through the emotions they evoke but also to the community of other users who share the same cultural references, reinforcing a collective identity.

Beyond forming niche communities, memes also act as a global communication medium, transcending cultural barriers as they are easily adapted to many contexts and, thus, can be understood by people from different backgrounds (McGrady and Hamm 2019). Their versatility

allows them to be reshaped, making them a powerful tool for cross-cultural dialogue and global communication.

The focus now shifts to exploring the underlying intentions of meme creators. The dominant function of memes is to entertain, often using humour, satire, or irony to capture attention and provoke laughter or amusement (Milosavljevic 2020). However, their role extends far beyond this, as memes can convey deeper messages. Examining these other intentions provides valuable insights into how memes function not only as sources of entertainment but also as drivers of public discourse and societal change.

Regarding marketing, memes have become a powerful tool in advertising, allowing brands to communicate in a way that is relatable, adaptable, and easily shareable at a lower cost when compared to traditional methods (Ngo 2021). Due to their viral nature, memes can enhance a company's visibility and engagement, especially among younger audiences who spend significant time on social media (Rathi and Jain 2023). Brands often use memes to advertise products by blending humour with cultural references, making their marketing efforts feel more authentic and less intrusive than traditional ads. While meme-based marketing offers brands the potential for significant visibility and engagement, it has risks. The viral and unpredictable nature of memes means that they can sometimes misfire or be misinterpreted, leading to backlash or controversy, as seen in cases like SunnyD's controversial tweet (Morabito 2019). Despite these risks, success stories such as Netflix's creative use of memes in its social media strategy (Beer 2019) prove that, when done correctly, memes can positively impact brand perception.

Just as brands have harnessed the power of memes to promote products, political campaigns, and movements have also embraced meme culture to communicate and engage with the public while promoting their points of view (Milosavljevic 2020). Nevertheless, most often, memes in politics are used to mock or criticise political figures, policies, or ideologies. Memes provide

an accessible and often humorous way to challenge authority and express dissatisfaction, making them a popular tool for political satire and critique (Moody-Ramirez and Church 2019). For instance, during the 2016 U.S. Presidential election, memes were employed as tools for political participation, often aiming to delegitimise political figures like Donald Trump and Hillary Clinton to sway public opinion and achieve a desired political outcome, such as the election of one candidate over another (Ross and Rivers 2017). Political actors themselves are aware of the power of memes and actively use them to shape public opinion. In fact, campaigns frequently hire paid bloggers and commenters to generate meme content, crafting images and narratives that influence how the public perceives political figures or ideologies (Kulkarni 2017). This calculated use of memes in political discourse underscores the recognition of their cultural and persuasive power.

However, that influence can be exploited negatively as well. The same viral qualities that make memes engaging also turn them into effective vehicles for spreading misinformation (Lynch 2022). Memes often oversimplify complex issues, distorting facts or promoting harmful narratives. During events like elections or the COVID-19 pandemic, memes were used to spread conspiracy theories (Basch et al. 2021). Their rapid, anonymous spread makes tracing or correcting false narratives difficult, further fuelling disinformation online.

Beyond the spread of misinformation, memes are being weaponised for harmful purposes, such as cyberbullying and hate speech (Greene 2019, 36), with offensive content often concealed behind humour or visuals. While social media platforms work to detect and moderate such content, traditional automated systems and human moderators frequently struggle to identify nuanced or disguised offensive material. This challenge has led to the development of advanced detection methods, including a recent study using multimodal deep learning models (Ahmed, Bhadani, and Chakraborty 2021) that combine text and image analysis to more effectively identify hateful memes, enabling them to be flagged and removed.

Therefore, although memes are often perceived as light-hearted and humorous, their influence can be profound, with both positive and negative impacts on society.

2.4. Meme Success as Object of Study

As memes continue to play a prominent role in the digital age, researchers have also turned their attention to understanding the factors that drive their success. Some studies emphasise external influences, such as platform algorithms and the amplifying role of social networks, while others focus on the internal attributes of memes, like their content, emotional resonance, and visual design. These varied approaches underscore the complexity of meme success, suggesting that both external dissemination mechanisms and the intrinsic qualities of memes play vital roles in their impact. Rather than being mutually exclusive, these perspectives often intersect, offering a more holistic understanding of meme dynamics.

Beginning with studies that mainly underscore the surrounding environment of memes, Weng, Menczer, and Ahn (2013) explored the role of community structure in meme virality, drawing an analogy to contagions. The authors found that viral memes behave like simple contagions, quickly crossing community boundaries, much like infectious diseases spreading through populations. In contrast, memes confined to a single community act like complex contagions, requiring reinforcement from within the group to spread. Their research emphasised that memes permeating multiple communities are far more likely to go viral, highlighting the importance of network diversity for meme virality. Building on this, their 2014 study expanded the analysis to include additional factors, such as the position of early adopters in the network and the early growth rate of meme adoption, which was found to be the weaker predictor. While the position of early adopters in the social network provides insights into the potential size of its audience, influencing its reach, it was still concluded that community diversity remained the most robust predictor of meme popularity (Weng, Menczer, and Ahn 2014).

Similarly, Ford, Krohn, and Weninger (2023) emphasise the continued importance of

community structure in meme virality but shift focus to the competitive dynamics among memes. The authors view memes as entities vying for limited user attention within social networks. As the number of memes being created and shared increases, the diversity and longevity of individual memes are influenced by the finite attention available from users. This perspective provides a more nuanced understanding, highlighting that meme virality relies not only on reaching diverse communities but also on successfully navigating the competitive pressures of attention economics.

Barnes et al. (2021) also found that neutral memes performed better than extreme sentiments overall, but among extreme sentiments, negative sentiment outperformed positive sentiment in driving upvotes. Building on this focus on sentiment, other studies have demonstrated how tools such as VADER and more advanced machine learning models enable the classification of social media content into sentiment categories (positive, neutral, negative), offering valuable insights into user attitudes and behavioural patterns. For instance, Srinidhi et al. (2024) emphasized that VADER's lexicon-based approach efficiently captures nuanced emotional expressions and is particularly well-suited for analysing short text, such as social media comments.

In practical terms, the conclusions of studies that aim to predict meme success and virality have various applications, as they provide insights into emerging topics that can be useful across multiple areas (Colbaugh and Glass 2011). For instance, in marketing, brands can use these insights to craft meme-based campaigns that resonate with audiences and increase engagement. Political campaigns can also benefit by identifying and leveraging memes that align with their messaging to shape public opinion. Additionally, content moderators can leverage predictive models to anticipate which memes might go viral, enabling them to manage content more effectively and address potential issues before they escalate. On a more relaxed note, communities such as r/MemeEconomy have turned meme success prediction into a form of entertainment, where users speculate and "trade" memes as if they were stocks, betting on which

will rise or fall in popularity (Literat and Van Den Berg 2019).

2.5. Meme Engagement as a Rare Event

All these studies share a common core challenge – they aim to predict a rare event. A rare event is an infrequent occurrence within a dataset that, while uncommon, is expected (Illowsky and Dean 2023, 526). Examples include extreme weather conditions, such as hurricanes or tornadoes, sudden stock market crashes, rare disease outbreaks, or equipment breakdowns in manufacturing environments. It is important not to mistake rare events for anomalies or novelties. While rare events are infrequent yet expected occurrences, anomalies are unexpected deviations from the norm, representing irregularities that break established data patterns, like a sudden spike in network activity indicating potential fraud. On the other hand, novelties refer to entirely new and previously unobserved phenomena that are not present in the training dataset (Carreño, Inza, and Lozano 2020). Substantial meme engagement falls under the rare event category because high levels of engagement, though infrequent, are anticipated within meme datasets. Engagement on social media typically follows a Pareto distribution, where a small number of memes capture most interactions. Therefore, while most memes receive low to moderate engagement, it is expected that, on rare occasions, a select few posts will attract significant attention.

Due to their nature, datasets with rare events lack sufficient data within the minority class. This scarcity can manifest as either absolute rarity, where the minority class has very few instances, or relative rarity, where the majority class outnumbers the minority class. As a result, identifying rare events often requires analysing large volumes of data, which are both time-consuming and computationally expensive, to capture enough minority instances for meaningful learning (Shyalika, Wickramarachchi, and Sheth 2024). Furthermore, the imbalance in the dataset may lead the model to develop a maximum-generality bias, favouring broad patterns that fit the majority class while overlooking the features of the minority class.

Another issue posed by imbalanced datasets is that traditional evaluation metrics, such as classification accuracy, are often unsuitable, as they tend to favour the majority class. For instance, in a dataset with a 100:5 imbalance, a model that misclassifies all minority class instances could still report a high accuracy of 95% despite failing to detect any rare events (Maalouf and Trafalis 2011).

These hurdles have been referred to as the curse of rarity (Liu and Feng 2024). In a study focused on autonomous vehicles, this concept was introduced to illustrate how the scarcity of safety-critical events, such as crashes caused by unexpected obstacles or sudden changes in driving conditions, constrain the model's capacity to learn effectively from these crucial occurrences. Numerous studies have focused on ways to mitigate the curse of rarity.

In terms of data preprocessing, it is essential to use tailored techniques to retain and emphasise critical features that might otherwise be overlooked in imbalanced data, as they may appear redundant or lack sufficient expression in the majority class (Shyalika, Wickramarachchi and Sheth 2024). Feature selection techniques include supervised methods like wrapper-based approaches, such as Recursive Feature Elimination (RFE), which remove less relevant features while preserving patterns linked to rare events. Filter-based methods use statistical measures to detect subtle and low-frequency signals. Intrinsic methods like attention mechanisms and adapted decision trees focus on rare, dispersed features, ensuring critical elements are retained during training. Additionally, feature engineering techniques like data augmentation create synthetic samples to increase the representation of rare patterns, while dimensionality reduction and feature scaling refine the dataset, increasing the detectability of rare event signals and boosting overall model performance.

Recent advancements in predicting rare events focus on approaches to handle imbalanced data effectively (Chen et al. 2024). Data-level methods, like oversampling (e.g., SMOTE) and under-sampling, modify datasets to rebalance classes, while algorithm-level techniques, such as cost-

sensitive learning, enhance models' sensitivity to minority classes by adjusting learning goals to prioritize underrepresented groups. Ensemble learning (e.g., boosting, bagging) merges multiple classifiers to capture minority class patterns, with boosting emphasizing difficult-to-classify instances. Deep learning's long-tail learning addresses deep imbalances, redistributing focus across rare classes.

Finally, performance metrics vary depending on the specific tasks in rare event prediction, including classification, clustering, regression, and simulation (Shyalika, Wickramarachchi, and Sheth 2024). In classification, metrics like G-Mean and Matthews's correlation coefficient (MCC) are preferred as they balance performance across minority and majority classes, providing a more accurate picture in imbalanced scenarios. Clustering tasks use metrics like the Elbow method and Silhouette coefficient to maintain distinct clusters for rare events. For regression, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are commonly used to measure prediction precision. Simulation tasks rely on probability evaluation, statistical robustness, and confidence intervals to assess model reliability. Selecting the appropriate metric for each task enhances the accuracy and reliability of rare event predictions.

3. Methodology

3.1. Data Collection

3.1.1 Data Sources

Data plays a crucial role in this study, forming the backbone of the analysis and conclusions. The strength and dependability of the data are key to ensuring accurate predictions and great results. Various sources were explored to gather the necessary historical data, primarily focusing on meme-related content on Reddit, specifically within the r/memes subreddit. The Arctic Shift GitHub Repository, the Pushshift Reddit Dataset, and the dataset stored at Regensburg University from the work *A Corpus of Memes from Reddit* were among the valuable resources consulted for this purpose.

Regensburg University Dataset (Schmidt et al. 2023) was helpful for initial testing. However, it only contained the top 1,000 memes (measured by the highest number of upvotes) for each half-year for the 2013-2021 period. By excluding memes with lower or moderate engagement, this dataset fails to capture the full spectrum of meme performance, limiting the ability to generalize findings to the broader meme population. This selection bias could also distort the analysis by inflating the perceived relationships between independent variables and engagement metrics, as it only examines cases where engagement is already high.

On the other hand, Arctic Shift (Heitmann 2023) provided a comprehensive dataset with several advantages for this study. It contains functional URLs for downloading images, which was crucial as some independent variables include visual characteristics of memes. Additionally, the dataset's organization and flexibility allowed for efficient data retrieval using “The Arctic Shift Download Tool”, enabling the selection of specific time frames, subreddits, and distinctions between posts and comments. These features made the Arctic Shift a practical source for data collection, offering a manageable and representative foundation for analysing meme dynamics on Reddit.

It is essential to mention that the data available in the Arctic Shift database reflects the state of Reddit posts and comments when Arctic Shift collected them, not when the dataset was downloaded. Arctic Shift collects Reddit data through periodic snapshots. In April 2023, a retrieved timestamp was added to all posts. This timestamp indicates the specific moment when the content was fetched from Reddit. Fields such as score, or number of comments represent the values at the time of collection rather than their real-time values. Moreover, starting in November 2023, Arctic Shift introduced a secondary retrieval mechanism. A process that revisits collected data 36 hours after the initial retrieval to capture changes, such as post edits, engagement metrics updates, or deletions.

So, the Arctic Shift database is not designed to provide real-time data. When researchers

download a dataset, a historical snapshot of Reddit content is obtained, which may not accurately reflect the present state of the post's score and comments. However, this limitation is acceptable for this study, as engagement on posts typically declines significantly after the first few days of publication (Vassio et al. 2021). This trend occurs mainly because Reddit's algorithm prioritizes newer content, making it more visible to users shortly after posting. As time passes, newer posts take precedence, leading to less attention on older posts.

3.1.2. Subreddit

Reddit has been established as the platform for this study. Given its multitude of subreddits, it is necessary to determine which specific communities will be included in the analysis. Since many subreddits are dedicated to memes, each with distinct user behaviours and approaches to content management, attempting to incorporate data from multiple subreddits could introduce a range of challenges. For instance, one subreddit may enforce strict guidelines against certain memes, while another might actively encourage them. Community size also plays a crucial role, as larger subreddits offer greater post visibility, impacting engagement metrics. To address these challenges and ensure consistency, this research focuses on a single subreddit: r/memes.

Focusing on a single subreddit places greater importance on selecting the right community, as the entire analysis hinges on its characteristics. When multiple subreddits are used, any limitations or biases of one can be balanced out by others, reducing their overall impact. However, with only one subreddit, a poor choice could significantly affect the reliability of the findings. r/memes, regarded as a central hub for meme culture on Reddit, was deemed an appropriate choice based on its characteristics.

Firstly, r/memes was established in 2008, making it one of the oldest meme-focused subreddits on the platform. This longevity has allowed the community to mature and refine its rules, moderation practices, and overall structure, fostering a stable and consistent environment for analysis. Moreover, it has accumulated a substantial subscriber base of over 35 million

members. A large audience reflects broader user behaviours and ensures a significant volume of engagement metrics, providing robust data. Additionally, unlike many meme-focused subreddits catering to specific themes, r/memes encompass various meme content. This generality makes it particularly suitable for analysing overarching trends in meme culture without being constrained by the biases of a specific niche.

3.1.3. Observation Period

The year 2023 was chosen for this analysis because it represents the most recent complete year at the time of the study. Focusing on the latest full year ensures the dataset reflects current trends within meme culture, making the findings more relevant to contemporary digital environments. Another advantage is that 2023 is more comprehensive in data than past years. Older years on the Arctic Shift dataset, 2018 and before, lack details such as whether an author is a premium user, the awards received by the post, or the post's removal status, either because such features did not exist at the time or because the data on them was not retrievable. Moreover, if an older year had been chosen, the analysis would have captured the subreddit during its earlier stages, missing out on the advantages of its current maturity, such as fine-tuned rules and a more extensive subscriber base.

In addition, 2023 offered a manageable dataset size compared to the peak activity years of 2020, and 2021, when Reddit experienced significant surges in traffic, likely driven by global lockdowns that increased online activity during the pandemic (Veselovsky and Anderson 2023). These years represent a unique and atypical context, with user behaviour and engagement heavily influenced by the extraordinary circumstances of the COVID-19 pandemic. The unprecedented spikes in activity during this time may not accurately reflect typical patterns, introducing potential biases into the analysis. By focusing on 2023, the study avoids these irregularities, offering a recent and more representative dataset of standard online behaviour while remaining manageable for analysis.

3.2. Data Preprocessing

The datasets comprising 462,274 posts and approximately 6 million comments from r/memes for the year 2023 were sourced from Arctic Shift. Figure 1 provides an overview of the steps taken to achieve the final datasets used in this research, which are further detailed in this section.

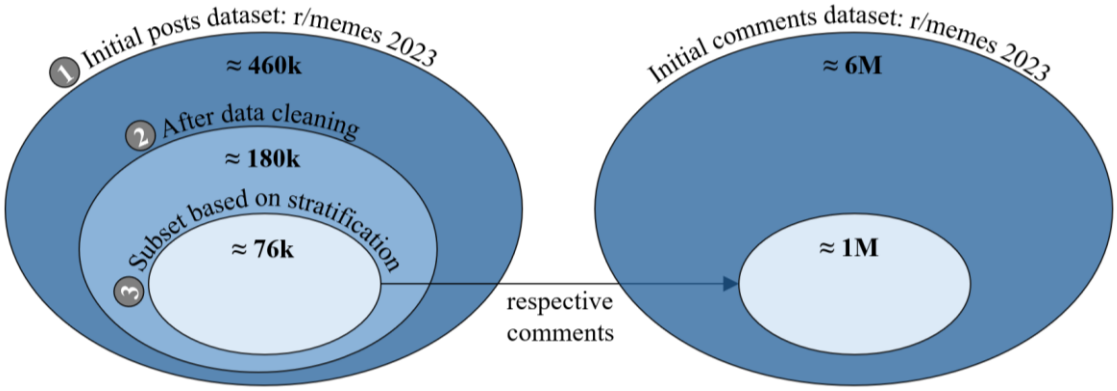


Figure 1: Data Preprocessing for r/memes 2023 posts and comments datasets

3.2.1. Data Cleaning

The preprocessing phase begins with fundamental data cleaning. Video posts were removed, as the study does not focus on video content. Posts submitted on the author's "cake day", the anniversary of their account creation, were also excluded. This decision is based on the understanding that such posts may attract comments unrelated to the meme's content and receive a higher upvote ratio due to the associated event, introducing noise and potential bias into the dataset.

After this initial step, the dataset was reduced to 447,136 posts. However, a problem was identified: not all URLs were valid, which meant the visual content of the memes could not be used due to the absence of accessible images. To address this, a dedicated Python function was developed to verify the URL data for each post and identify those with images available for download. The function checks whether a URL is reachable and points to a valid image file, excluding GIFs, as the analysis focuses on static images. To handle network errors, timeouts, or other issues, the function employs a retry mechanism with exponential backoff, allowing multiple attempts to access a URL while progressively increasing the delay between retries.

This approach ensures that transient issues, such as server overload or network instability, are handled effectively.

To enhance the speed and efficiency of the process, the function employs multithreading, allowing for the concurrent evaluation of multiple URLs. Additionally, it utilises custom HTTP headers, such as a "User-Agent", to emulate browser behaviour and reduce the likelihood of being blocked by websites during the evaluation. For users wishing to monitor progress, the function includes an optional feature to display the status of each URL as it is processed.

Once all posts have been assessed, the results of the status of their image URLs are stored in a new variable called "status", together with the rest of the dataset in a new CSV file.

Posts with URL error status codes of 404 (not found) or 403 (forbidden), with URL missing data, or with other unresolved issues were removed, resulting in 179,161 posts with valid URLs.

3.2.2. Stratification

Although necessary, the cleaning process introduced unintended biases by disproportionately removing posts with specific characteristics. To address this and ensure the dataset represented the entire year, a subset was created based on three key stratification categories, allowing for the selection of a representative sample for analysis.

Comment Category: Posts were categorised based on the number of comments they received, using different levels, such as Low [0-5], Medium]5, 15], High]15+].

Score Category: Posts were also divided into score categories, such as Low [0,10], Medium]10,25], High]25,50], and Very High [50+].

To choose the appropriate numerical intervals for these categories, the percentile distribution of posts for the number of comments and scores had to be examined in the original 2023 dataset (Appendix 1). For the number of comments, 85% of the posts have 5 or fewer comments, 90% have 9 or fewer comments, and 95% have 25 or fewer. For the score, the intervals had to reflect a higher magnitude of values since 85% of the posts have a 32 score or less, 90% have a 73 or

less, and 95% of posts have 325 or less.

Created Month: The posts were also divided into categories of the month they were created, i.e., January to December. From the created UTC data available, the hour of the creation of the post was also extracted. However, this variable proved unusable because of the diversity of users' time zones. This made the month variable the only reliable one regarding posting time.

Using the stratification categories, a representative subset was selected to closely match the distribution of three key variables—Score, Number of Comments, and Month Created—from the original 2023 dataset. The selection process prioritized maximizing the subset's size while ensuring it accurately reflected the original dataset's distribution. This approach resulted in a refined dataset comprising 76,281 posts, which served as the basis for extracting all associated comments and downloading the corresponding meme images.

3.2.3. Images Download

The download_images.py module was developed to manage image downloads for the project efficiently. This module processes pairs of post IDs and previously validated URLs. In the initial stage, images are fetched using the requests library. Next, the format of each downloaded image is verified, and if the format is invalid, the image is processed through a format converter to ensure compatibility. Finally, the validated images are saved in a user-specified folder, with file names corresponding to their respective post IDs.

To optimize performance, these stages are orchestrated by a central function and executed in parallel using the *concurrent.futures* library with the *ThreadPoolExecutor* function. This approach transforms Python's standard single-threaded execution into multi-threaded execution, significantly speeding up the download process by allowing multiple images to be processed simultaneously. In practice, downloading images involves waiting for a network response, during which the program is idle. By using multi-threading, the program can continue downloading and processing other images simultaneously, significantly speeding up the overall

process. This methodology is highly recommended for I/O-bound tasks, as it allows multiple operations to proceed concurrently, reducing idle time and yielding significant improvements in download times.

The preprocessing stages were implemented as functions, enabling easy replication for additional years or different subreddit datasets. By standardizing the process, all researchers involved in the study, whether working with images or other aspects, were provided with a consistent dataset. This alignment ensured greater comparability and coherence across the individual research components.

3.3. Variables

3.3.1 Dependent Variables

In the scope of this analysis, the dependent variables used aim to define meme engagement. However, it is essential to recognise that there is no universally agreed-upon or clear set of measures that should always be used to define it. Researchers may choose different variables to assess it, resulting in various interpretations and approaches. For instance, standard measures often include likes, shares, comments, click-through rates, time spent interacting with the content, frequency of reposts, etc. Furthermore, there is a thin line between studying engagement, success, popularity, and virality, with researchers often using different terms while employing the same dependent variables. The fact that these concepts are frequently used interchangeably underscores the importance of clearly defining how engagement is quantified in this study. Thus, considering the variables provided by the extracted dataset, the following were identified as the most relevant:

Score: The score reflects the cumulative upvotes minus downvotes that a meme receives, providing a measure of its approval among Reddit users. The score serves as a net indicator of community sentiment, where positive engagement (upvotes) and negative engagement (downvotes) combine to reflect overall reception. Higher-scoring memes are more likely to gain

visibility on the platform, increasing their reach to wider audiences (Medvedev, Lambiotte, and Delvenne 2019).

Number of Comments: The volume of comments on a meme captures the level of conversation it stimulates, indicating user interest and depth of engagement. Compared to the score, commenting requires active participation and reflects a more direct interaction with the content, suggesting higher user involvement. Research highlights that content generating significant discourse is often emotionally resonant, controversial, or particularly relevant to its community (Barnes et al. 2021; Weng, Menczer, and Ahn 2014).

Awards: The presence of awards on a meme reflects its recognition and appreciation by the community, indicating a higher level of approval or value placed on the content. Unlike upvotes, a relatively quick and light-hearted form of engagement, awards require users to pay using Reddit Coins purchased with real money. This means that awarding a meme is often a more intentional decision, indicating a stronger endorsement of the content's quality.

3.3.2. Independent Variables

For this study, the variables or attributes used to explain or predict the dependent variables are referred to as independent variables. The predictors are changeable and so, have been carefully developed with the specific aims of the analysis in mind. These variables were selected and developed to find trends in the usage of memes and comments, combining a wide range of attributes on posts and comments while excluding those clearly defined as dependent.

Research in meme engagement seeks to understand the diversity seen in these prediction factors. Common examples include visual features, such as colour palettes and image resolution (Ford, Krohn, and Weninger 2023); textual features, such as sentiment analysis and textual content (Barnes et al. 2021); and contextual features, such as post timing and subreddit-specific features (Shifman 2014). These variables help to establish a baseline for testing patterns of meme engagement and will be explored in greater depth later in this study.

3.3.3. Control Variables

Control variables are factors included in models that, while not the primary focus, account for external influences that might confound the relationship between the independent and dependent variables. By holding these variables constant, the analysis isolates the effects of the primary independent variables and minimizes the risk of results being skewed by external factors (Memon 2024). Since this study focuses on measuring engagement, it is crucial to account for variables that could independently influence user interactions with posts. To select the variables, features that could impact engagement were selected and descriptive statistics were calculated to verify whether those features impacted the dependent variables.

Quarter: Quarters were included as a control variable to account for potential seasonal variations and time-based trends in user activity and engagement on Reddit. Different times of the year can influence the volume and type of content being posted, as well as how users interact with it since meme trends often align with current events or cultural phenomena. Indeed, the data shows notable differences in average scores and comment activity across quarters, (Appendix 2), highlighting the importance of controlling for this variable. However, quarters do not seem to have a significant effect on the proportion of awarded posts in the dataset.

To avoid the dummy variable trap when one-hot encoding, Q4 was excluded from the analysis, becoming the baseline reference category. This aligns with standard practices in regression analysis, allowing the effects of Q1, Q2, and Q3 to be interpreted relative to Q4. By treating Q4 as the reference, we account for the temporal variation in engagement patterns and isolate the unique contributions of each quarter to the dependent variables.

Author premium status: Reddit Premium consists of a paid subscription, and while authors with premium status don’t enjoy a direct impact on the visibility of their posts, the status itself may still influence engagement. Premium users can appear more credible due to exclusive features like custom avatars and badges, which distinguish their profiles and signal their

premium membership. These visual markers may lead other users to view their content as more trustworthy or noteworthy, fostering greater interaction. Premium users' access to exclusive subreddits, such as r/lounge, may provide insights into effective posting practices and trends, potentially enhancing the engagement their posts receive. Moreover, premium users are likely to possess more experience and knowledge about meme culture, increasing the likelihood that their posts achieve high engagement. Looking at the data (Appendix 3) posts by premium authors exhibit distinct engagement patterns compared to those by non-premium authors. While the range of scores is similar for both groups, posts by premium authors receive significantly higher average upvotes, with an increase of over 5763 upvotes compared to non-premium authors. Additionally, premium authors' posts generate more comments, both in terms of range and average, with approximately 24 more comments per post. Furthermore, posts by premium authors are awarded more frequently, with an award frequency of 2.2%, compared to just 0.5% for non-premium authors. These differences highlight the need to account for premium author status as a variable, as it clearly influences engagement metrics and reflects a disparity in user interactions with content based on author type.

Removed: When a post is removed, it is no longer accessible to the public on Reddit, preventing users from interacting with it through actions such as commenting, upvoting, or awarding. This cessation of interaction directly impacts engagement metrics, as any potential future activity on the post is effectively halted. While this may not have a significant impact if the meme had already passed its peak engagement period prior to removal, it could disproportionately affect newer posts that were removed before they had the chance to accumulate interactions. As such, accounting for removal status is essential to ensure a fair analysis of engagement patterns. The data (Appendix 4) reveals a substantial decrease in the average score and number of comments for removed posts, with reductions of approximately 330 upvotes and 14 comments. Interestingly, there is a notable increase in the frequency of awards for removed posts compared

to those that were not removed.

Locked: The locked post variable was included as a control in the analysis for a similar reason. It impacts one of the dependent variables - the number of comments - by preventing further interaction. This influence on overall engagement necessitated careful consideration to avoid overcontrol, a scenario where a control variable removes variation integral to the relationship under study, potentially masking the true effect of the independent variable on the outcome. To address this concern, the analysis assessed whether the locked post variable was part of the causal pathway. For example, if the independent variable affected the number of comments indirectly by influencing whether a post was locked, controlling for the locked post variable could eliminate part of the measured effect. Conversely, omitting the locked post variable would allow both the direct and indirect effects of the independent variable on comments to be observed, including any mediated through locking. It is observable in the descriptive statistics (Appendix 5) that the average score and the frequency of awarded posts between locked and non-locked posts is very similar, while there is a difference in the average number of comments. The control variables outlined above were incorporated into regression models, with a specific set selected for each dependent variable based on their observed effects. For the dependent variable number of comments, the control variables included locked status, author premium status, removal status, and quarters. For score, the control variables were author premium status, removal status, and quarters. Lastly, for awards, the model used author premium status and removal status. The statistical significance of these relationships was evaluated using p-values, with a significance threshold of 0.05.

3.4. Descriptive Statistics

3.4.1 Posts

As Table 1 shows, the dataset contains 76,213 posts. The average number of comments per post is approximately 11.96, with a maximum of 8,383 comments on a single post. Similarly, the

average post score is 211.39, with a maximum score reaching 88,465. These figures highlight the significant disparity in post popularity and the wide range of engagement across the dataset. The large gap between the mean and the maximum values, along with the relatively low medians and high standard deviations, strongly indicates a highly positively skewed data distribution. This is further corroborated by the density plots for both the number of comments (Figure 2) and scores (Figure 3), which reveal that most posts are concentrated near the lower end of engagement metrics. To address this skewness, a logarithmic transformation was applied to the engagement metrics, allowing for a more normalized distribution to do regression analysis. Since the logarithm of zero is undefined, a value of 1 was added to all data points before the transformation. This adjustment ensures that all values are transformed consistently, preventing discrepancies where zero values would otherwise remain unchanged.

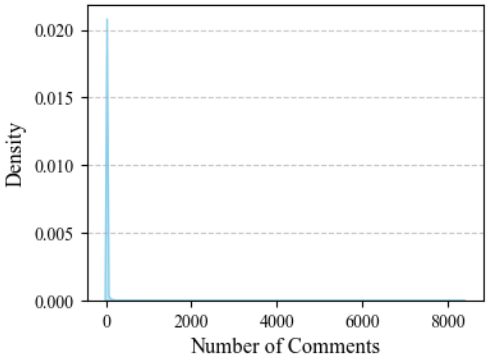


Figure 2: Density Plot for Number of Comments

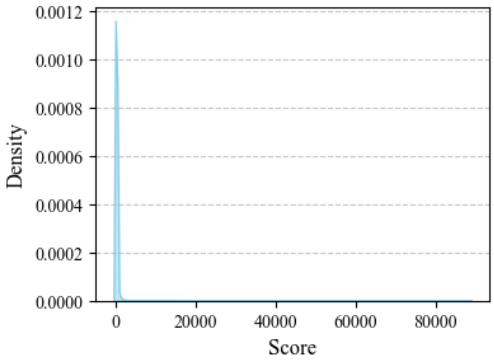


Figure 3: Density Plot for Score

The average number of awards received per post is minimal (0.008), with a maximum of 12 awards on a single post, highlighting the rarity of this phenomenon. Similarly, the proportion of posts made by premium authors is 3.6%, underscoring that most of the activity comes from non-premium users, who represent 97.54% of the user base in this dataset. Both observations are expected, as awarding posts and premium membership generally require payment, and most Reddit users prefer to use the platform for free. Moreover, over half of the posts (57.8%) were removed, suggesting a substantial rate of content moderation in this subreddit.

Table 1: Descriptive Statistics for Posts Dataset

	num_comments	score	locked	author_premium	total_awards_received	has_awards	removed_post
count	76213.000	76213.000	76213.000	76213.000	76213.000	76213.000	76213.000
mean	11.964	211.392	0.076	0.036	0.008	0.006	0.578
std	93.266	2008.409	0.264	0.186	0.148	0.075	0.494
min	0.000	0.000	0.000	0.000	0.000	0.000	0.000
25%	1.000	1.000	0.000	0.000	0.000	0.000	0.000
50%	1.000	1.000	0.000	0.000	0.000	0.000	1.000
75%	3.000	7.000	0.000	0.000	0.000	0.000	1.000
max	8383.000	88465.000	1.000	1.000	12.000	1.000	1.000

Approximately 8% of posts were locked, and around 2% were flagged as containing content suitable only for individuals over 18 years old. The fact that a large number of posts were removed yet only a small proportion were flagged as adult content suggests that r/memes enforce strict content moderation policies extending beyond filtering for explicit material.

The dataset comprises contributions from 46,667 distinct authors, with an average of 1.63 posts per author and a maximum of 2,876 posts by the most prolific contributor.

Regarding seasonality (Figure 4), there are, on average, 6,351 posts per month, with the highest number of posts occurring in January (8,140) and the lowest in May (5,191). The first peaks at the start of the year, followed by a gradual decline, while the second shows a resurgence mid-year, with activity rising again during late summer and early autumn, suggesting that seasonal trends or cultural events may play a role in driving meme creation and engagement. The peak in January might reflect engagement with "New Year, New Me" memes or viral humour about resolutions, which are popular during the start of the year. The mid-year resurgence, particularly in late summer, could align with content surrounding blockbuster summer movie releases such as the *Barbenheimer* trend, which dominated social media discussions (Hamilton 2023).

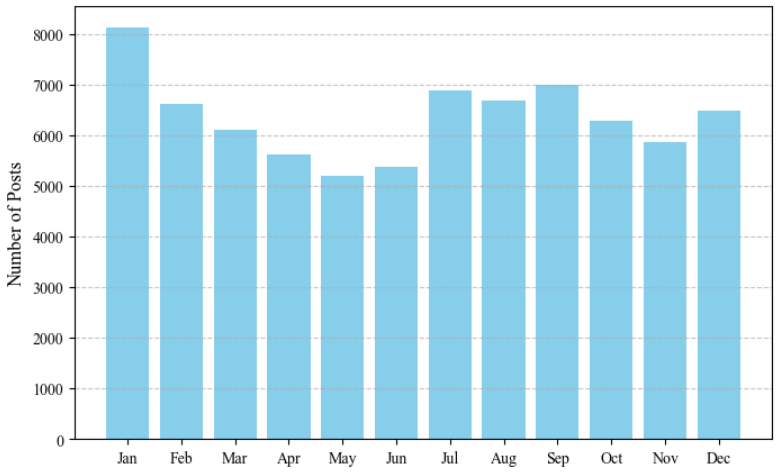


Figure 4: Monthly Distribution of Posts in the Dataset

3.4.2 Comments

As seen in Table 2, the comments dataset contains a total of 279,479 unique commenters who collectively contributed to the discussion threads. The average number of comments per user is approximately 3.45, with a median of 1 comment. However, the standard deviation is 163.41, showing a wide disparity in activity levels, with some users contributing significantly more comments. As visualized in Figure 5, 59.1% of users only commented once, while 30.5% commented between 2 and 5 times. Users contributing between 6 and 10 comments accounted for 5.9%, and users who commented more than 50 times were rare, representing less than 0.3% of the total user base. These figures underscore the skewed distribution of comment activity, with a few highly active users dominating the discussion.

Table 2: Descriptive Statistics for the Comments Dataset

	count	mean	std	min	25%	50%	75%	max
Comments per User	279479.0	3.447175	163.410102	1.0	1.0	1.0	2.0	78056.0
Token Length	963413.0	30.104948	64.336863	1.0	4.0	10.0	23.0	1778.0

Regarding comment length, the average token length of comments shows a notable pattern in Figure 6. Approximately 30.2% of comments were concise, containing 0-5 tokens. Comments with 6-10 tokens accounted for 20%, while comments in the ranges of 11-20 and 21-50 tokens comprised 21.4% and 17% of the dataset, respectively. Lengthier comments with over 100

tokens were rare, accounting for just 6.4% of the total. These statistics highlight that while the dataset captures a wide range of user interactions, most of the activity is concentrated among casual users who contribute with shorter comments.

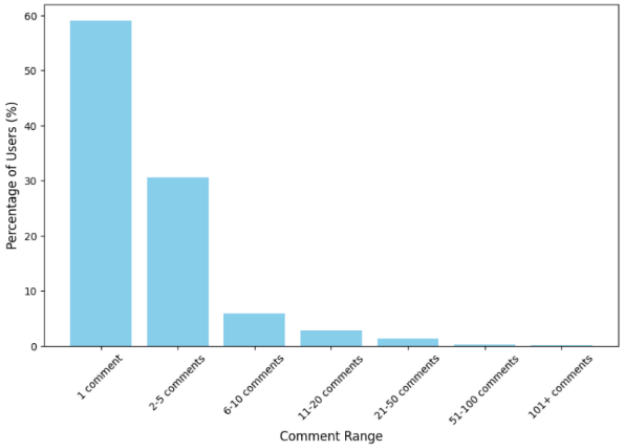


Figure 5: Distribution of User Comment Activity

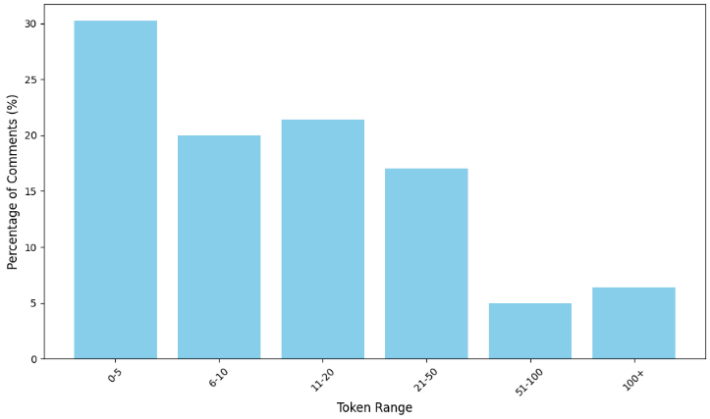


Figure 6: Distribution of Comment Token Lengths

4. Collaboration

4.1 GitHub

The GitHub repository for “reddit-memes” forms the central node around which all prescribed activities for the project take place. It allows easy collaboration and efficient version control among collaborators (https://github.com/martimesteves1/reddit-memes). In designing this project, GitHub was used to host this initiative because it could support a dynamic, iterative, and collaborative development framework yet still maintain high standards of organization, reproducibility, and accessibility (Perez-Riverol et al. 2016). The features of the platform established a dependable basis for managing the diverse demands of a research project characterized by significant scale and complexity, which required thorough data preprocessing, model implementation, and collaborative documentation.

The repository is systematically arranged into a flat layout directory structure that guarantees the distinct separation of essential functionalities while maintaining ease of navigation, therefore being well-suited for the collaborative practices adopted for this project. Inside the reddit memes folder, one can find modules dedicated to data acquisition tasks, like fetching

images or processing URLs and data preprocessing tasks with standardized cleaning and formatting methodologies — things that prove instrumental in constructing datasets for training and testing of the models.

The repository also implements dependency management using “uv”, making it very fast and simple to ensure consistent environments among contributors, reducing versioning conflicts.

Automated dependency checks and installation scripts are also included to make internal onboarding for new contributors easier. Furthermore, the repository incorporates a CI/CD pipeline through GitHub Actions to automate key processes such as running tests, syncing environments, and verifying the compatibility of feature branches. Any contribution made to the repository is thereby safeguarded against the possibility of integration errors while instituting code quality standards.

The repository uses a branching model aimed at allowing concurrent development. The main branch is the stable version of the codebase, and feature-specific branches like “download_images” or “mesteves-sandbox” give the possibility to a developer separate the work on common features from individual research. This practice of branching decreases the possibility of conflicts during the integration of contributions and ensures that the main branch always remains stable.

The scalability of the workflow on the platform itself finally proved to be quite an important asset in overcoming some of the project intricacies, therefore fostering proper collaboration and maintaining the quality of deliverables.

4.2. Programming Languages

This research project uses Python considering the flexible and rich libraries appropriate for data manipulation, distributed processing, and machine learning operations. The ability to enable modular programming and create reusable functions significantly facilitated the preparation of preprocessing scripts. The principles of object-oriented programming further contributed to the

development of modular and scalable components, ensuring the project's architecture remained robust and adaptable to large-scale data analysis. Additionally, Python's extensive library ecosystem proved instrumental in organizing the project workflow effectively (Shaik et al. 2021).

In this project, the centre of Python-based analytics stack is *Pandas*, *NumPy*, *Matplotlib*, *Stargazer*, *TensorFlow*, *scikit-learn*, and *Statsmodels*, where each contributes special capabilities that help to make data management, visualization, and modelling tasks easy. *Pandas* and *NumPy* are perfect for efficient data manipulation and numerical computing. *Matplotlib* makes the visualization of trends quick and intuitive. In addition, *TensorFlow* and *scikit-learn* are the tools used to build, train, and evaluate complex machine learning models, from neural networks to ensemble methods. *Statsmodels* and *Stargazer* complements these resources by supporting traditional statistical modelling, including the use of Ordinary Least Squares (OLS) to perform regression analysis. These libraries together ensure both scalability and efficiency, handling gracefully the difficulties of analysing large datasets from Reddit.

4.3. Code Rules and Quality

Ensuring high-quality code within the "reddit-memes" project is achieved through strict guidelines and automated processes. Pre-commit hooks automatically check for compliance with coding standards, type errors, dependency errors and failing tests before the code is committed. All this helps find and fix problems early within the development process and saves us from pushing errors to the main code repository.

Automated testing is part of the quality assurance process for the project. The repository uses Pytest, a flexible and powerful testing framework, to validate the functionality of individual elements in the codebase. Pytest excels at creating both simple unit tests and more involved functional tests, therefore proving to be an important tool in the verification of the reliability of the software components that are part of the project. Automated testing serves not only to

confirm the correctness of the code but also allows for the fast detection and fixing of errors during development (Hunt 2023, 175–186).

This repository uses MyPy, to enforce data validation and integrity with type annotations. Using MyPy ensures that data is correct according to predefined schemas and dramatically reduces errors caused by malformed or incorrect inputs, that could be more prevalent in commonly worked features.

This also becomes a relevant feature for a project like reddit-memes, since the data is coming from an ever-changing platform like Reddit. If this validation of data structures at runtime didn't happen, MyPy can help to make the data pipeline long-lasting and highly reliable (Hunt 2023, 38) This repository's architecture is based on a modular design philosophy, where scripts and functions are compartmentalized into distinct units performing specific tasks. This will make testing, debugging, and enhancements of the codebase easier to modulate. Further, extensive type annotations and exhaustive docstrings have been integrated throughout the code to enhance readability and provide explicit documentation for developers.

Finally, Ruff is used to lint all code, ensuring that all developed code follows proper code quality standards, making the final product efficient to read and understand between collaborators and future users.

4.4. Documentation

Comprehensive documentation is one of the main backbones of the reddit-memes project, as it makes everything transparent and accessible for all participants. Proper documentation encourages good communication, reduces waste, and helps to handle changes and issues, thus improving both the development process and overall user satisfaction. Modern research emphasizes that good documentation is at the core of successful software engineering practices since it ensures clear communication and simplifies the onboarding process for newcomers (Treude, Middleton, and Atapattu 2020). In addition, sufficient documentation should also be

maintained in agile contexts because this approach balances the need for flexibility with the need for transparency and consistency in project processes (Islam, Hasan, and Eisty 2023).

The README.md document contains detailed information on the project, its goals, results and other key characteristics. Meanwhile, the CONTRIBUTING.md contains detailed information about the workflow, including all the steps in setting up dependencies, initializing the development environment, and submitting any contribution. Besides, this doc emphasizes the importance of running the project in accordance with established coding standards and quality assurance controls.

Inline documentation, especially in the form of thorough docstrings for functions and classes, greatly enhances the accessibility of the codebase. The docstrings follow standard formats, such as NumPy or Google style, with unambiguous descriptions of the intent, parameters, and return values of each function or class. This level of documentation supports understanding and maintenance of the code by both current contributors and future developers. Although agile development generally values functional software over documentation, writing "just enough" documentation is acknowledged as the key to preserving both project efficiency and clarity.

By enforcing linter compliance with Ruff and pre-commit hooks, it is ensured that all code follows a baseline level of code quality, including the creation of proper documentation for the numerous modules, functions, classes and notebooks in the project.

With these methodologies, the highest levels of code quality and operational efficiency are achieved by the reddit-memes initiative. Synergy between extensive documentation, strong validation procedures, and smooth workflows guarantees that the project will retain accessibility, dependability, and scalability while pursuing its goal: creating a complete framework for analysing and predicting meme and comment engagement on Reddit.

4.5. Individual Methodology

The individual methodology was developed and implemented, targeting to meet specific project

goals and improve overall understanding.

In the GitHub repository, Jupyter Notebooks are used by each contributor to document their special research question and approach. A detailed record of preprocessing steps, several feature engineering techniques and multiple model implementations are provided by these notebooks. Not just the specific methodologies employed by the contributors are caught, but also many technical choices are drawing attention to that adjust with the overall objective.

This paper provides an in-depth explanation of the processes and choices made by each contributor above and beyond the code documentation, it helps readers understand their methodologies thoroughly. Many individual contributions become transparent and reproducible, and they easily integrate into the general analysis.

5. Beyond the Image: The Role of Text in Shaping Meme Engagement

5.1 Introduction

Building upon the collective exploration of factors influencing meme success on Reddit, this chapter delves into a new element - the text - exploring the role of text within a meme and its title on shaping engagement. By integrating Natural Language Processing (NLP) techniques, this analysis extracts, processes, and evaluates textual data. Specifically, it examines the text in memes by exploring its presence, length, and sentiment, while for titles, the focus is on length and sentiment due to their mandatory inclusion. Distinguishing between these two components is essential, as memes are primarily visual content with optional text overlaying the image, whereas titles are mandatory, purely textual, and prominently displayed at the top of Reddit posts. Analysing each separately ensures that their distinct roles in shaping engagement are accurately assessed. From this, the following research question was formulated:

RQ: How do memes' textual elements and titles influence meme engagement?

Previous studies have explored various aspects of textual elements in memes and their role in communication and virality. For instance, a study that extracted text using Tesseract Optical Character Recognition (OCR) and Google Cloud Vision API for refined analysis addressed thematic content and demographic sharing patterns. It found disparities in how different demographic groups express themselves through image-with-text memes compared to other forms of communication on Twitter. Some individuals preferred using memes with embedded text to convey their messages rather than traditional plain-text tweets, underscoring the importance of not neglecting such elements in analyses to fully understand how people communicate and engage online (Du, Masood, and Joseph 2020).

Another study focusing on virality prediction identified factors influencing memes' success using a Random Forest model. Among the factors analysed, such as composition, subjects, and audience, text was one of them and was extracted using OCR from the Google Vision API. It

was found that memes with fewer words are more likely to go viral, as excessive text can delay recognition and impair engagement. A threshold of 15 words was identified using the Kolmogorov-Smirnov test, above which the likelihood of virality decreases significantly (Ling et al. 2021). In contrast, a study examining the factors influencing meme virality on Facebook, using Negative Binomial Regression, measured text length by the number of lines and concluded that longer text positively contributes to a meme's likelihood of being shared (Nawawi et al. 2020).

Still focusing on structural elements of text but extending the analysis to sentiment, a study analysed textual, visual, and timing features to understand their role in driving meme virality on Reddit by applying logistic regression, where memes were classified as viral (top 5% by upvotes) or non-viral (bottom 5%). The text was extracted using OpenCV and EasyOCR, and sentiment analysis was conducted using the Hugging Face Transformers library with the DistilBERT model. The findings revealed that positive sentiment in meme text and titles significantly enhances virality, while negative sentiment, call-to-action phrases, and longer titles tend to reduce it (Sah and Jordan 2024). However, Tsugawa and Ohsaki (2017) concluded differently regarding text sentiment. Their analysis of text-only tweets on Twitter found that negative sentiment significantly increased the likelihood of content spreading widely, with negativity playing a stronger role in driving diffusion compared to positivity. However, positive sentiment still led to greater engagement than neutral sentiment, indicating that emotional tone, whether positive or negative, is more effective than neutrality in driving virality.

Although neither of these studies uses the exact same independent and dependent variables, their findings collectively set the foundation for developing this analysis. Reviewing prior studies allows this research to situate its results within a broader context, facilitating comparisons and identifying how its findings align with or diverge from existing knowledge.

5.2. Hypothesis

The analysis has been divided into a subset of hypotheses to ensure clarity and organisation, each targeting a different textual element.

Memes without text depend solely on their visual appeal, allowing them to be understood by audiences regardless of language or literacy barriers. It can also be argued that memes with text add context to the visuals, making them clearer and easier to understand. While these perspectives are not mutually exclusive, given that memes typically feature highly readable and concise text tailored to fit within limited space (Sherratt, Pimbblet, and Dethlefs 2023), it is expected that the inclusion of text will enhance engagement by effectively complementing the visual elements. For instance, the "Two Buttons" meme features a character sweating while choosing between two options (Appendix 6). It shows generic indecision without text, but captions like "Spend money on things I don't need" and "Save for the future" create a specific scenario where users can insert themselves, making the meme more relatable and engaging. Therefore, the following hypothesis was formed:

H1: Memes with text lead to a significantly larger (i) number of comments, (ii) score, and (iii) likelihood of receiving awards of awards compared to memes without text.

In this part of the analysis, the presence of text in memes was examined without considering its length. Since the dimension of text length is thought to potentially impact engagement, the next hypothesis explores how text length, measured by word count, influences engagement metrics. Indeed, as previously mentioned, memes typically contain short text, which is expected to complement the visual elements effectively, thereby enhancing engagement. However, when the text is long, it likely means that most of the meme is dominated by text, overshadowing the image and overloaded with information. Humans process visual information faster than text; therefore, longer text in the meme and its title demand a greater attention span. Since memes are designed to be consumed quickly, excessive text interrupts their fast-paced and easily

digestible format. This can overwhelm audiences, leading them to skip over lengthy memes. Following this rationale, engagement is expected to increase with text length up to a certain point, after which it begins to decline, leading to the next hypothesis:

H2: Higher word count in (a) memes and (b) titles leads to an increase in (i) the number of comments, (ii) score, and (iii) the likelihood of receiving awards up to a certain point, after which these metrics begin to decline.

Moving on to the last hypothesis, this aims to analyse the content of the text itself, rather than its structure, through sentiment analysis. Sentiment is anticipated to influence engagement metrics in distinct ways. For scores, memes with positive sentiment are expected to receive the highest scores, as they generate likes, followed by neutral sentiment, which tends not to provoke strong reactions, and negative sentiment, which gets the lowest scores due to dislikes. The same pattern is expected for the likelihood of receiving awards, as positive sentiment fosters appreciation, neutral sentiment has a moderate effect, and negative sentiment reduces the probability of recognition. Regarding the number of comments, memes with negative sentiments are anticipated to generate the most discussion. Negative sentiment often provokes strong reactions, debates, or disagreements, encouraging users to engage in conversation. Memes with positive sentiment are expected to elicit fewer comments, as they inspire agreement or appreciation rather than lengthy discussions. Neutral sentiment, lacking emotional intensity, will likely spark the fewest comments since it does not evoke strong reactions or drive users to interact extensively. From this reasoning, the following hypothesis is proposed:

*H3: Sentiment in (a) meme text and (b) titles influences engagement metrics, with (i) the number of comments increasing with negative sentiment and decreasing with neutral sentiment, while (ii) scores and (iii) the likelihood of receiving awards increasing with positive sentiment and decreasing with negative sentiment***.**

5.3. Methodology

5.3.1. Data Refinement

The dataset used to evaluate the hypotheses comprises 76213 memes, representing a subset of the 2023 r/memes collection. Additional processing is required to prepare it for analysis.

Since this research question focuses on the textual elements of memes, it was essential to determine whether memes contained text before proceeding with further analysis. This step was crucial not only because analysing textual elements requires the presence of text but also because identifying text-containing memes in advance improves efficiency by eliminating unnecessary text extraction attempts on images already flagged as not containing text.

To address this, a dataset of 374 images was manually labelled by visually inspecting each image and categorising it as either containing text or not so that these labels could serve as the ground truth for evaluating the performance of three open-source text detection models: EAST, Tesseract, and PaddleOCR. Each model was applied to the labelled dataset, and their predictions were compared against the ground truth to assess their performance using metrics such as Precision, Recall, and F-measure. PaddleOCR demonstrated the best overall performance, achieving the highest F-measure (0.991087).

Next, using PaddleOCR as the detector, Tesseract and PaddleOCR were employed to extract text from the images. Unlike the previous step, the 374 images were not labelled with the extracted text itself, as this would have been too time-consuming. Instead, the extracted text was inspected qualitatively to evaluate the models' performance, revealing a recurring issue: the models often failed to correctly interpret spaces, frequently combining adjacent words into a single, continuous string. This limitation could significantly underestimate the word count in subsequent text length analysis, leading to distorted results.

The search for models was extended, and Google Vision was selected for its high accuracy, which is well-ranked among other OCR tools (Dilmegani 2024), as well as for its suitability for

this specific task. It includes a specialized text detection feature, designed specifically to extract text from images, distinguishing it from conventional document text extraction methods that primarily handle structured and dense text formats. This feature is particularly useful in meme analysis, as meme text is often embedded in noisy backgrounds. Google Vision was used without incurring costs by using the initial free credit provided by the platform. Still, it is important to note that this is a paid model, which becomes particularly relevant if more images are analysed in the future.

Besides extracting text, Google Vision can identify language as a multilingual model, which is relevant to this analysis. For instance, differences in language structure, such as verb conjugations or word order, can lead to variations in sentence length and word count when expressing the same idea (Dudău and Sava 2021). To ensure consistency and simplify the analysis, memes with text not in English, which accounted for 7% of the dataset (Appendix 7), were eliminated. Another justification for this decision is the existence of memes where part of the text is in English, and other parts are in a different language, such as Chinese, often used intentionally not to be understood but instead function more like an image, contributing to the overall aesthetic or humour. This aligns with the expression "It's all Greek to me," commonly used to describe something incomprehensible. Similarly, in these memes, the non-English text serves as a visual design element intended to symbolise something incomprehensible regardless of its actual linguistic meaning. Including such memes in the analysis would distort results, particularly for sentiment analysis.

After completing the text extraction process, the results were merged with relevant columns from the original dataset. These included the title, engagement metrics, and control variables such as the quarter, whether the post was removed or locked, and whether the author had premium status. This ensured that only data pertinent to the research question was retained. Feature engineering was then performed to create variables required for the analysis, including

a binary indicator for text presence, length features, and sentiment analysis for text in memes and titles.

Text length was measured as word count, with spaces used as the criterion to separate words. A logarithmic transformation was applied to the text length variables, consistent with the approach used for certain dependent variables. This transformation was necessary to address the high skewness observed during exploratory data analysis (EDA), further discussed in the EDA section. For sentiment analysis, a pre-trained multilingual BERT model was used, the same one employed in other research questions for consistency. While it is common to classify text as positive, neutral, or negative or assign a score in a range such as -1 to 1, this model outputs sentiment as star ratings, ranging from 1 star (most negative) to 5 stars (most positive).

5.3.2. Selected Variables

Having established the hypotheses to be tested and identified the necessary variables, Table 3 summarises their roles in the analysis, outlining whether each was used as an independent or dependent variable and indicating the associated hypothesis.

Table 3: Description of independent and dependent variables used in each hypothesis

Variable	Description	Type	Independent	Dependent
has_text	Indicates whether the meme contains text (1 = Yes, 0 = No)	Numerical - Binary	H1	
log_text_count	Logarithm of the word count in the text extracted from the meme	Numerical - Continuous	H2	
log_title_count	Logarithm of the word count in the text extracted from the meme	Numerical - Continuous	H2	
text_sentiment	Sentiment score of the meme text, rated from 1 star (negative) to 5 stars (positive)	Categorical - Ordinal	H3	
title_sentimemt	Sentiment score of the title text, rated from 1 star (negative) to 5 stars (positive)	Categorical - Ordinal	H3	
log_score	Logarithm of the engagement score (e.g., upvotes) for the meme	Numerical - Continuous		All
log_comments	Logarithm of the number of comments on the meme post	Numerical - Continuous		All
has_awards	Indicates whether the meme received awards (1 = Yes, 0 = No)	Numerical - Binary		All

5.3.3. Exploratory Data Analysis

Although an exploratory data analysis was previously conducted, performing a focused EDA on the columns most relevant to the specific research question being addressed in this chapter is important.

After performing preliminary checks, it was verified that only the column indicating the text in the meme, as well as the column capturing its sentiment, contained null values, as expected for image-only memes, and it was confirmed that all column data types were correct for analysis.

Moving on to the study of the distribution of the variables, the presence of text is highly imbalanced (Appendix 8). 93.6% of memes contain text, aligning with the notion that meme creators find value in incorporating text into their memes to enhance communication.

Both text word count (Figure 7) and title word counts (Figure 8) display heavily positively skewed distributions for continuous variables. Indeed, most memes contain a few words and have short titles, as expected, with medians of 14 and 4 words and means of approximately 20 and 5 words, respectively. However, the range is substantial, with the most extended text containing 2478 words and the longest title reaching 150 words. This suggests the presence of extreme outliers in both columns, as confirmed by the boxplots (Figures 9 and 10). A logarithmic transformation was applied to handle the high skewness of these variables, similar to the approach used for the dependent variables. This transformation reduced the skewness, resulting in more symmetric distributions that are more appropriate for analysis (Figures 11 and 12).

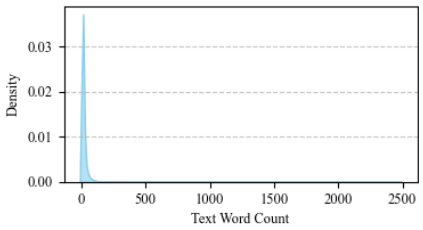


Figure 7: Density plot of text word count

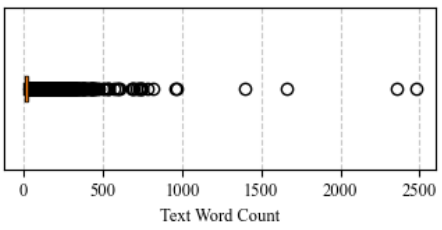


Figure 9: Boxplot of text word count distribution

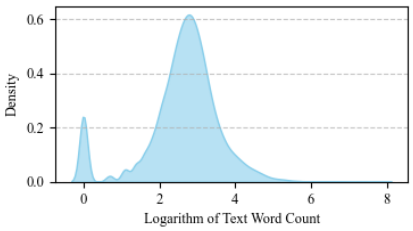


Figure 11: Density plot of text word count

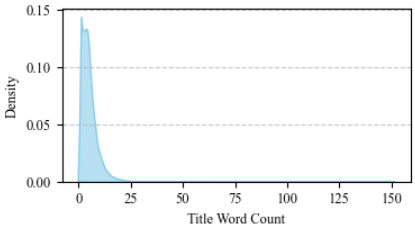


Figure 8: Density plot of text word count

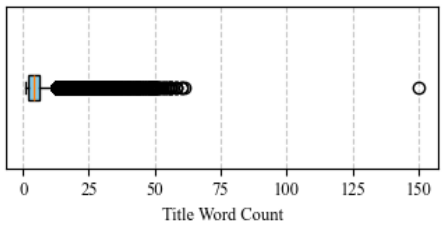


Figure 10: Boxplot of title word count distribution

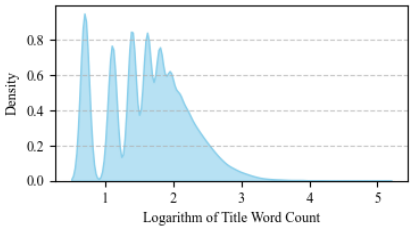


Figure 12: Density plot of logarithm of title word count

Examining the categorical columns, the sentiment distributions of both text and title reveal a clear polarisation, with most meme ratings concentrated at the extremes, aligning with the notion that memes are crafted to evoke strong emotional responses. However, the text and title distributions lean in different directions. This polarisation skews negatively for text sentiment (Figure 13), with 51.6% of memes rated 1 star, reflecting highly negative sentiment, compared to 26.1% rated five stars. In contrast, title sentiment (Figure 14) skews more positively, with 42.3% of titles rated five stars, while 27.3% fall into the 1-star category.

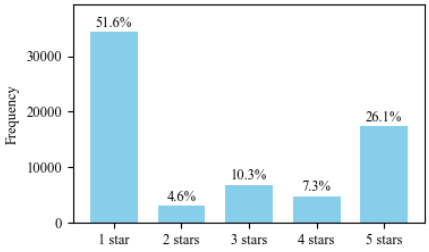


Figure 13: Distribution of text sentiment

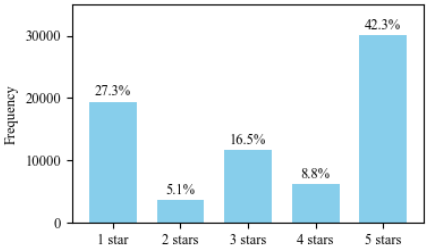


Figure 14: Distribution of title sentiment

To better understand the relationship between the text and the title, the percentage of words in the title also present in the text was computed. The resulting distribution reveals a heavily skewed pattern, with most memes exhibiting little to no overlap (Appendix 9). This suggests that, despite Reddit requiring users to write a title, creators generally avoid recycling the text from the meme itself, resulting in them having complementary rather than redundant roles.

The correlation analysis reveals another interesting insight: a negative correlation between text count and text sentiment, as well as title count and title sentiment (Appendix 10). While the correlations are not strong, they suggest an inverse relationship where longer texts and titles are associated with more negative emotions. One possible explanation is that longer texts might

reflect rants and detailed criticisms, often linked to negative sentiment.

5.3.4. Hypothesis Assessment Framework

Before delving into regression analysis, a preliminary study assessed relationships between variables. Scatter plots were used to examine trends and potential non-linear patterns for continuous independent variables (logarithm of text and title word count) and continuous dependent variables (logarithm of number of comments and score). However, the visualisations did not reveal any clear or strong relationships, suggesting that any potential associations may be weak (Appendix 11).

For categorical independent variables (presence of text and text and title sentiment), the Shapiro-Wilk test for normality and Levene's test for homogeneity of variances revealed deviations from normality and unequal variances (Appendix 12), justifying the use of non-parametric methods. The Kruskal-Wallis test was employed for continuous dependent variables and consistently revealed significant differences in distributions across categories (Appendix 13). For the categorical dependent variable (presence of awards), the Chi-Squared test of independence also identified significant associations between the categorical variables (Appendix 14). Despite the Kruskal-Wallis test not providing information about the magnitude of the differences and the Chi-Squared test not indicating the strength of associations, their significant results suggest the presence of notable differences and relationships worth further investigation.

Hypothesis testing was conducted using regression models, which quantify how changes in independent variables affect the average value of the dependent variable (Gordon 2015). This aligns with the goal of measuring the impact of text features on meme engagement. Moreover, regression allows for control variables and can handle non-linear relationships, making it ideal for testing hypotheses like H2 and H3, which assume non-linear effects.

More specifically, linear, quadratic, and logistic regression models were employed, depending

on the nature of the hypotheses (Table 4). Linear regression was applied for hypotheses expecting a straightforward, proportional relationship between variables, such as the impact of having text on engagement. Quadratic regression was used for hypotheses H2 and H3, which assume non-linear effects that vary depending on the variable. For instance, word count is hypothesised to follow an inverted U-shaped relationship with all engagement metrics, while sentiment is expected to show a U-shaped relationship with comments. Given that the presence of awards is a binary variable, logistic regression was employed to examine the likelihood of receiving awards while accounting for the dichotomous nature of the data (Berger 2017).

Table 4: Regression models used for hypothesis testing

			Dependent Variables		
		Hypothesis	*log_comments*	*log_score*	*has_awards*
Independent Variables	*has_text*	H1	Linear regression	Linear regression	Logistic regression
	log_text_count	H2	Quadratic regression	Quadratic regression	Quadratic logistic regression
	log_title_count	H2	Quadratic regression	Quadratic regression	Quadratic logistic regression
	text_sentiment	H3	Quadratic regression	Linear regression	Logistic regression
	title_sentiment	H3	Quadratic regression	Linear regression	Logistic regression

5.4. Results

The first hypothesis proposed that memes with text would lead to significantly greater engagement across three metrics: number of comments, score, and likelihood of receiving awards, compared to memes without text. The results validate this hypothesis for two of the three metrics (Figures 15, 16 and 17) at the considered 5% significance level: the number of comments and score. For these metrics, the coefficients are positive, indicating that the presence of text in memes is associated with increased engagement. Specifically, memes with text are associated with a 10.85% increase in the number of comments and a 30.47% increase in score compared to textless memes.

has_text	0.103***
	(0.073 , 0.132)
author_premium	0.188***
	(0.154 , 0.222)
locked	0.316***
	(0.289 , 0.344)
removed	-0.730***
	(-0.745 , -0.715)
quarter_1	0.296***
	(0.276 , 0.315)
quarter_2	0.159***
	(0.138 , 0.181)
quarter_3	-0.026**
	(-0.046 , -0.006)
Observations	71081
R^2	0.141
Adjusted R^2	0.141
Residual Std. Error	0.968 (df=71073)
F Statistic	1668.957*** (df=7; 71073)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 15: OLS for the presence of text on the logarithm of number of comments

has_text	0.266***
	(0.220 , 0.312)
author_premium	0.165***
	(0.112 , 0.218)
removed	-1.634***
	(-1.657 , -1.611)
quarter_1	1.196***
	(1.165 , 1.227)
quarter_2	1.045***
	(1.012 , 1.078)
quarter_3	-0.547***
	(-0.578 , -0.516)
Observations	71081
R^2	0.338
Adjusted R^2	0.338
Residual Std. Error	1.511 (df=71074)
F Statistic	6057.766*** (df=6; 71074)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 16: OLS for the presence of text on the logarithm of score

has_text	0.740*
	(-0.013 , 1.493)
author_premium	0.894***
	(0.616 , 1.172)
removed	-1.729***
	(-1.988 , -1.470)
Observations	71081
Pseudo R^2	0.061
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 17: OLS for the presence of text on the presence of awards

The second hypothesis suggested a non-linear relationship between word count and engagement metrics, proposing that increasing word count in (a) memes and (b) titles would initially enhance engagement. However, beyond a certain threshold, the effect was expected to diminish and eventually reverse, leading to a decline in the engagement metrics. The coefficients in the models provide evidence supporting the hypothesised U-shaped relationship (Figures 18-23): the linear term is positive, indicating an initial increase in engagement with rising word count, while the squared term is negative, reflecting a turning point where additional words reduce engagement It can also be observed that the linear term is higher for titles than for text within memes, indicating that titles have a stronger immediate impact on engagement metrics as word count increases compared to meme text. This is likely because titles are prominently displayed and are often the first element users notice. If a title fails to engage, users are less likely to view or interact with the meme itself, making the title's impact on engagement particularly significant. However, the squared coefficient's absolute value is larger for titles than for meme text in the models for comments and score, suggesting that while titles initially boost engagement more strongly, their effect diminishes more steeply as word count grows.

log_text_count	0.074***
	(0.053 , 0.094)
log_text_count_sq	-0.007***
	(-0.011 , -0.003)
author_premium	0.185***
	(0.151 , 0.220)
locked	0.310***
	(0.283 , 0.338)
removed	-0.724***
	(-0.739 , -0.709)
quarter_1	0.294***
	(0.274 , 0.314)
quarter_2	0.159***
	(0.138 , 0.181)
quarter_3	-0.026**
	(-0.046 , -0.006)
Observations	71081
R^2	0.142
Adjusted R^2	0.142
Residual Std. Error	0.967 (df=71072)
F Statistic	1474.348*** (df=8; 71072)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 18: OLS for text word count on the logarithm of number of comments

log_text_count	0.253***
	(0.221 , 0.285)
log_text_count_sq	-0.039***
	(-0.046 , -0.033)
author_premium	0.160***
	(0.107 , 0.213)
removed	-1.618***
	(-1.641 , -1.595)
quarter_1	1.193***
	(1.162 , 1.224)
quarter_2	1.042***
	(1.009 , 1.076)
quarter_3	-0.546***
	(-0.577 , -0.515)
Observations	71081
R^2	0.340
Adjusted R^2	0.340
Residual Std. Error	1.509 (df=71073)
F Statistic	5231.776*** (df=7; 71073)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 19: OLS for the logarithm of text word count the logarithm of score

log_text_count	0.889***
	(0.348 , 1.430)
log_text_count_sq	-0.145***
	(-0.242 , -0.047)
author_premium	0.886***
	(0.607 , 1.164)
removed	-1.692***
	(-1.951 , -1.433)
Observations	71081
Pseudo R^2	0.063
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 20: OLS for text word count on the presence of awards

log_title_count	0.095***
	(0.041 , 0.148)
log_title_count_sq	-0.031***
	(-0.046 , -0.015)
author_premium	0.189***
	(0.155 , 0.223)
locked	0.319***
	(0.291 , 0.347)
removed	-0.734***
	(-0.750 , -0.719)
quarter_1	0.296***
	(0.276 , 0.316)
quarter_2	0.159***
	(0.137 , 0.181)
quarter_3	-0.028***
	(-0.048 , -0.007)
Observations	71081
R^2	0.141
Adjusted R^2	0.141
Residual Std. Error	0.968 (df=71072)
F Statistic	1456.165*** (df=8; 71072)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 21: OLS for title word count on the logarithm of number of comments

log_title_count	0.609***
	(0.526 , 0.692)
log_title_count_sq	-0.169***
	(-0.192 , -0.145)
author_premium	0.164***
	(0.111 , 0.217)
removed	-1.610***
	(-1.634 , -1.586)
quarter_1	1.189***
	(1.158 , 1.220)
quarter_2	1.033***
	(1.000 , 1.067)
quarter_3	-0.555***
	(-0.586 , -0.524)
Observations	71081
R^2	0.339
Adjusted R^2	0.339
Residual Std. Error	1.510 (df=71073)
F Statistic	5208.751*** (df=7; 71073)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 22: OLS for title word count on the logarithm of score

log_title_count	1.244**
	(0.235 , 2.253)
log_title_count_sq	-0.314**
	(-0.589 , -0.039)
author_premium	0.892***
	(0.614 , 1.170)
removed	-1.673***
	(-1.939 , -1.407)
Observations	71081
Pseudo R^2	0.061
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 23: OLS for title word count on the presence of awards

The sensitivity of engagement to title length is further emphasised by the turning points, which occur at much lower word counts for titles than for text within memes in both the number of comments and score (Table 5).

Table 5: Turning points of the independent variables across the dependent variables

Variable	Number of Comments	Score	Likelihood of awards receipt
Text Word Count	198	26	21
Title Word Count	5	6	7

These findings suggest that users tolerate higher word counts in meme text than titles, likely because the text often explains or clarifies the meme's meaning. By providing essential context or enhancing the humour, longer text within memes can add value to the overall content, making users more willing to engage with it despite its length. This highlights the importance of achieving a critical balance in title length, as excessively long titles can negatively affect engagement, a sensitivity that is less prominent for meme text.

The third hypothesis proposed that sentiment in (a) meme text and (b) titles would influence engagement metrics, with (i) the number of comments increasing with negative sentiment and decreasing with neutral sentiment, while (ii) scores and (iii) the likelihood of receiving awards would increase with positive sentiment and decrease with negative sentiment.

The results of the analysis revealed that the hypothesised influence of sentiment on engagement metrics was largely insignificant, with only the effect of text sentiment on the likelihood of receiving awards showing statistical significance (Figure 24). The observed negative coefficient for text sentiment on the likelihood of receiving awards does not validate the initial hypothesis.

text_sentiment	-0.091***
	(-0.150 , -0.031)
author_premium	0.885***
	(0.605 , 1.166)
removed	-1.730***
	(-1.992 , -1.467)
Observations	66506
Pseudo R^2	0.058
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 24: OLS for text sentiment on the presence of awards

Unlike upvotes and downvotes, which represent a form of engagement accessible to all users, awarding memes is a less frequent action undertaken by a specific subset of users. The negative coefficient suggests that this group may be particularly drawn to memes with darker or more critical tones, finding them thought-provoking or uniquely humorous. As a result, awards may serve as a way for them to express deeper appreciation for the impact or resonance of such memes, even when the sentiment is negative.

Despite these results, it would be premature to conclude that sentiment does not influence other engagement metrics. Several limitations in the data and methodology may have contributed to the lack of significant findings, as will be further discussed in the limitations section. For instance, the distribution of sentiment in the dataset is not balanced, with certain sentiment categories potentially underrepresented, making it difficult to detect clear relationships. Additionally, extracting sentiment from meme text poses unique challenges, as much of the text is often descriptive rather than explicitly emotional. Meme text frequently serves to clarify the image or set up a joke rather than express sentiment in the traditional sense, which can result in sentiment analysis tools misclassifying or failing to capture the subtleties of the content. These factors likely obscure the true relationship between sentiment and engagement, suggesting that further research is needed to draw firmer conclusions.

5.5. Challenges

Text extraction relied on Google Vision, chosen for its high accuracy and suitability. However, extracting text from memes presents unique challenges due to the nature of this medium. Memes are often designed with creative distortions, unconventional fonts, and overlapping text and visual elements. While these features are intentionally crafted to enhance the humour and aesthetic value of memes, they complicate the ability of OCR systems to detect and accurately extract text. Unlike structured sources such as documents or plain images with clear typography, meme text often lacks a standardised format, further increasing the complexity of the task.

This complexity can result in some words not being extracted, potentially omitting key parts of the meme's text. Conversely, an equally challenging issue arises when OCR systems extract irrelevant text such as watermarks, numerical markings on a speedometer in the background, or decorative elements like a mouse pad with a world map. Although these are indeed text, they are better interpreted as visual elements rather than linguistic content. Since this research

focuses on the role of textual elements in engagement, such extraneous text introduces irrelevant data into the results, distorting the intended analysis.

Another limitation stems from the exclusion of non-English text, a decision aimed at ensuring analytical consistency and addressing the complexities of handling multilingual data. Without tailored processing, incorporating multiple languages could undermine the reliability of the analysis due to linguistic variations and structural differences. While this approach enhances the accuracy of findings within the context of English-language memes, it narrows the study's scope, omitting potential insights into meme engagement across diverse linguistic and cultural contexts.

Finally, sentiment analysis introduces its own set of challenges. Meme text is often exceptionally short, typically consisting of just a few words or phrases. In this dataset, the median text length is only 14 words, which underscores the brevity of the content. This poses a limitation for many sentiment analysis models, which often require a minimum amount of linguistic input to generate reliable classifications. Additionally, the meaning of meme text is heavily intertwined with accompanying visual elements, making it highly context-dependent. Analysing the text in isolation strips away crucial cues provided by the imagery, hindering the ability of sentiment models to interpret emotional tone accurately (Behera and Ekbal, 2020). The informal nature of meme language further complicates sentiment analysis as meme text frequently incorporates cultural references and sarcasm, which can mislead sentiment models (Hazman, McKeever, and Griffith 2023). Sarcasm is particularly problematic, as it often conveys a sentiment opposite to the literal meaning of the words, increasing the likelihood of misclassification.

6. Discussion

6.1. Human Faces and Meme Formats: Complementary Roles

The visual presence of human faces and the structural characteristics of meme formats emerged as interconnected factors influencing engagement. While faces alone had mixed effects—modestly increasing scores (+2.15%, coefficient: 0.0215) but slightly reducing comments (-2.03%, coefficient: -0.0203)—their impact was strongly mediated by the meme's format and textual context. For instance, **'Emotional Reaction'** memes, where the main focus is the presence of a face expressing an emotion in reaction to a textual element, have a stronger engagement in comments than meme formats that rarely or never have faces, like **'Drawing'**, **'Photo'** and **'Screenshot'**, but did not have statistically significant effects in the score engagement. **'Situational'** and **'Template'** commonly have prominent facial and textual elements in their construction, and both had stronger performance in score and comment engagement than other categories.

The size and diversity of faces further influenced these factors. Larger faces reduced engagement (coefficient: -0.3838 for score; -0.1951 for comments), likely overshadowing textual and visual harmony. In contrast, emotional diversity in faces, captured by the "mixed emotions" variable, positively influenced both scores and comments, reinforcing the value of visual complexity in driving engagement. **'Situational'** memes that often have mixed emotions have consistent results, as they drive more score and comment engagement than meme formats like **'Emotional Reaction'** and **'Drawing'** that don't often mix emotions.

6.2. Text as a Mediator Between Visual and Structural Elements

Text played a pivotal role in amplifying or modulating the impact of visual and structural features. Memes with text significantly outperformed those without, boosting scores (+30.47%) and comments (+10.85%). Text length exhibited a non-linear relationship, with moderate lengths maximising engagement before diminishing returns set in. The previous pattern aligns

with the role of text in situational memes, where short captions contextualise the visuals without overwhelming them.

Formats that inherently integrate text, such as the **'Template'** and **'Situational** formats, also had significantly more engagement in score and comments than other formats. These rely heavily on textual-enhanced humour, which interacts with visual elements to generate engagement. Conversely, screenshots and photos, which often lack text, had the weakest engagement metrics, underscoring the importance of textual enhancement in visual content.

Effective engagement arises not from any single feature but from the harmonious integration of visual appeal, textual context, and structural clarity.

6.3. Emotional Dynamics and Visual-Textual Interactions

Sentiment analysis showed how emotional and sentimental content in comments interacts with the visual and textual components of memes. Posts that received a higher proportion of highly negative comments, driven by users' strong reactions, were the posts that attracted more engagement in the form of discussions. Positive sentiment, on the other hand, was associated with higher scores and awards, in line with the supportive tone of many situational memes.

These findings illustrate feedback loops: memes that can successfully pair relatable images with succinct text tend to create emotional responses, which then encourage user engagement through commentary, ratings, and recognitions. High-arousal emotions like joy consistently increased all engagement metrics when compared to low-arousal emotions. Anger, as a high-arousal emotion, also attracted engagement in the form of discussions by significantly increasing the number of comments. However, this engagement was not necessarily done with a positive intent. As a result, users, despite being engaged, were less motivated to upvote or award the post. In contrast, sadness as a low-arousal emotion reflected its limited capacity to stimulate meaningful interaction or recognition. Memes that featured emotionally diverse

content, such as mixed emotion faces or polarizing comment sentiment, achieved wider engagement, indicating that diversity in emotional expressions promotes relatability and attracts discussion.

6.4. Exploring Linear and Non-Linear Trends

The machine learning analysis really helped to shift the focus from descriptive results to predictive insights to find linear and non-linear patterns in engagement metrics. Unlike previous analyses, which looked only at direct relationships, this approach used models like Random Forest to explain complex interactions and dependencies between different features. The Random Forest Regressor performed well in predicting log-transformed scores (R-squared: 0.85) and comment counts (R-squared: 0.63), capturing complex relationships. More specifically, the models showed that the interaction between text length, subreddit size, and emotional diversity is non-linear, where certain thresholds increase engagement, while others yield diminishing returns.

Significantly, these models measured the underlying impact of visual characteristics in conjunction with textual and platform-related variables. For example, although an increase in text length initially increases engagement, overextension in length decreases its effect—an effect that is moderated by meme structure and emotional variation. Similarly, the infrequent appearance of awards showed a strong correlation with the diversity of sentiment (entropy), proving the ability of machine learning in identifying subtle factors driving interactions. Such cases were, however, much more challenging for classification tasks. While having high accuracy of 0.99, the models performed poorly with F1-scores of 0.08.

In a nutshell, these machine learning frameworks bring forth the complex, non-linear relationships between visual-textual interactions and emotional dynamics. Taken together with the descriptive analyses, these results show the importance of accounting for the direct and

dynamic effects of multimodal features on engagement.

6.5. Multimodal Integration

The results indicate that meme engagement is driven by the dynamic interplay of visual, textual, and emotional elements. Visual features, such as human faces, increase relatability but are highly dependent on contextual integration with textual and structural elements. Text acts as a mediator, contextualizing visuals and eliciting emotions, whose effectiveness is modulated by format and length. Emotional dynamics are amplified in these interactions, while diversity and polarisation bring about broader engagement and deeper user involvement. Adding another layer of understanding, machine learning exposes non-linear patterns and shows how these multimodal features interact in very diverse contexts.

In conclusion, meme engagement is a complex phenomenon where successful meme engagement calls for the integrated function of visual, textual, and emotive elements. These results put into focus the importance of understanding multimodal interactions in creating content for digital platforms—a most valuable message for content creators and researchers.

7. Limitations and Challenges

Despite the increasing complexity of multimodal analytical methodologies, investigations of meme interactions on Reddit still face numerous constraints and challenges. First, memes are inherently dynamic and context-dependent. As cultural markers that frequently rely on humour, societal critique, and relevant references, their impact can shift rapidly with evolving events, linguistic trends, and changes in user demographics (Shifman 2014). Items considered interesting and shareable one day may lose their significance just weeks later, making the task of building reliable predictive models highly challenging. Additionally, advancements in technology and data collection methods continuously expand the range of available engagement metrics. This progress can render older studies less comprehensive or less relevant over time,

as they may not account for newly available insights or evolving user behaviours. Such variability necessitates constant updates to methods, feature sets, and analysis frameworks to ensure that the insights drawn remain accurate and relevant (Adami and Kress 2014).

Similarly, this study was constrained by the engagement metrics available in the dataset, which dictated how engagement could be measured. Ideally, more granular metrics such as separate upvote and downvote counts, time spent viewing a meme, or sharing behaviour could have been included to provide a more comprehensive understanding of user interactions. One of the primary metrics used was score, which, while useful, has its limitations. Score, defined as the difference between upvotes and downvotes, provides an indication of a meme's success but does not fully encapsulate engagement. This is because downvotes, although a form of user interaction, reduce the overall score, potentially underrepresenting the true level of activity. For instance, a meme with high upvotes and equally high downvotes may appear to have low engagement when it has elicited significant user reactions. However, this limitation may be somewhat mitigated by the type of memes typically posted on r/memes. Given the subreddit focus on humour and light-hearted content, it is reasonable to assume that posts are less likely to provoke strong negative reactions that would lead to significant downvoting. This assumption is further supported by the fact that no posts in the dataset have negative scores, suggesting that downvoting is relatively uncommon on this subreddit. As a result, while score remains an imperfect measure of engagement, this focus on generally agreeable content may reduce the distortion caused by downvotes.

A notable characteristic of the dataset is the presence of duplicate memes—identical content posted multiple times but showing varying engagement metrics. This reflects the natural reposting and sharing behaviour common on platforms like Reddit, where memes are frequently recycled for different audiences or timeframes. However, it introduces challenges for analysis

by potentially over-representing certain meme formats or themes, which could distort conclusions about engagement trends. The variations in engagement metrics for identical memes highlight the influence of external factors, such as the timing of the post or the level of competition with other content at the moment of posting, complicating efforts to draw clear conclusions about the intrinsic qualities of memes that drive engagement.

Another factor that may influence engagement metrics is the potential activity of bots on Reddit. Bots can artificially manipulate metrics such as upvotes, downvotes, and awards, distorting the true representation of user engagement. Bots can be sophisticated, mimicking human behaviour in ways that make them difficult to distinguish from genuine users, which makes their identification challenging and not entirely accurate. These automated accounts could either inflate or deflate engagement metrics, for example, by artificially upvoting or downvoting memes or indiscriminately awarding posts. Although some bot accounts were identified in comments during the study, it is difficult to determine the extent of their presence on the dataset, introducing an additional layer of complexity to the interpretation of engagement metrics.

Still on the data perspective, one of the most persistent challenges is the significant skewness in engagement metrics, where a small fraction of memes garners a disproportionately large number of upvotes, comments, and shares, while the majority receives minimal attention. This can lead to two distinct challenges depending on how the dataset is structured (Haibo He and Garcia 2009). In a dataset where high-engagement memes are rare and most memes receive low engagement, machine learning models may lean toward predicting low engagement for most cases to optimise accuracy. On the other hand, if the dataset is artificially balanced—by oversampling high-engagement memes or under sampling low-engagement ones—the model risks overemphasising the trends associated with popular memes. In this scenario, the model may disproportionately focus on the distinct patterns of high-engagement memes, reducing its

ability to generalise to the broader meme ecosystem, which primarily consists of low-engagement memes. This challenge was particularly evident for the presence of text, as the number of memes receiving awards was significantly smaller compared to other engagement metrics. Indeed, in many models, the results for this variable were not statistically significant, likely due to the rarity of memes with awards.

Moreover, the multimodal nature of memes—combining images, overlaid text, and sometimes other media—adds to their analytical complexity. Each modality provides distinct cues, yet the interplay between them often conveys the true essence of a meme message. For instance, an image that appears innocuous on its own may become humorous or politically charged only when paired with accompanying text. Capturing this interdependence requires models capable of effectively integrating heterogeneous data while accounting for the interactions between visual, linguistic, and cultural markers (Zhong et.al. 2024). Without a nuanced understanding of these relationships, predictions of meme engagement risk being oversimplified, potentially overlooking the multidimensional humour and culturally grounded references that drive user interactions.

Finally, another significant limitation of this study lies in the computational challenges associated with analysing meme datasets. Memes are inherently multimodal, combining images, text, and sometimes additional media, which results in large and complex datasets. Processing and analysing such data require substantial computational resources, particularly for tasks like image analysis, text extraction, and sentiment evaluation. Expanding the scope to include multiple subreddits or additional years would dramatically increase the volume of data, further straining computational capacity. These challenges are particularly pronounced when working with image-heavy datasets, as images require more storage, processing power, and advanced feature engineering compared to purely textual data. Consequently, while focusing

the main analysis on r/memes and a single year ensures a manageable dataset and consistent scope, it also limits the study's ability to generalize findings or explore broader trends across platforms and timeframes.

8. Applications and Directions for Future Research

The various applications and avenues for future research on Reddit memes are all predicated on an understanding of the cross-modal interactions between visual, textual, and emotional elements. This study has brought to light the critical nature involved in combining these different facets in meme interactions prediction and evaluation. Of course, there remains further improvement and expansion potential using more sophisticated methodologies that could also test a wider range of contextual factors and engagement metrics.

Future research should centre on the dynamics of content amplification and the effects of meme formats and emotional engagement on interaction with users. Although this paper has emphasized direct analysis of the engagement metrics—scores, comments, and awards—there is still some potential to examine and predict further engagement indicators. More detailed and comprehensive metrics, such as the duration of viewing, patterns of sharing behaviour, or distinct tallies of upvotes and downvotes, might come in handy. Integration of these metrics would help researchers better grasp the complex nature of engagement dynamics and deliver actionable insights for content creators and platform developers.

Although the way that these algorithms work to create or diminish posts based on emotional arousal and content categories was not analysed in this study, investigation of these systems may yield very substantial insights into the incentive structures that guide content creation and dissemination. The existence of algorithmic biases, toward highly emotionally charged or divisive content, certainly warrants more research into their effects on user interactions and platform dynamics in general. For example, evidence has been found that ranking algorithms

based on engagement can amplify emotionally charged, out-group hostile content that may affect user perceptions and interactions (Milli et al. 2023).

One important area for further research is the ethical implications of content that evokes powerful emotions and negativity. This study highlighted the ability of memes to provoke high emotional reactions, both in terms of anger and joy, as well as their subsequent effects on engagement metrics. However, more research is required to understand the long-term effects on individual well-being, societal discourse, and platform engagement. With the rise in emotionally engaging memes, there is a danger that these will further polarize or amplify harmful narratives, which raises critical questions about content platform responsibilities in moderating such phenomena. Research has shown that the diffusion of emotionally charged messages in social media may engender changes in societal values and even have an impact on political attitudes, raising ethical concerns about what content is spread (Steinert 2021).

Building further upon the machine learning methods used in this study, future research should explore more advanced techniques to better capture the non-linear relationships found. For instance, more complex models, such as gradient boosting methods or neural networks, or ensemble models, could potentially offer a deeper understanding of the relationship between sentiment and engagement in different contexts. These models might be particularly well-suited to predicting a broader range of engagement metrics and overcoming some of the limitations of existing models, such as the identified inference disparities and the challenges of handling complicated relationships among multimodal features. Recent multimodal sentiment analysis studies have shown it to be capable of integrating visual and linguistic features to display complex emotions anda contexts contained within memes, evidencing the importance of advanced analysis methodologies (Hazman, McKeever, and Griffith 2023).

The expansion of the analysis concerning various types of memes offers a significant avenue

for exploration. Existing research indicates complex interactions among meme formats, textual components, and emotional responses. Subsequent investigations might utilize multi-agent inference methodologies to enhance findings further.

For example, sequential inference methods that classify memes as being ultimate or base according to attributes, or model comparisons with better contextual understanding, might bring forth underlying relationships that exist between meme formats and user engagement. More importantly, such approaches would offer the ability to specify exactly how different types of memes produce disparities in forms of engagement, therefore fostering a more lucid view of their overall impact.

Research on multimodal recontextualization of political memes has shown that presentation modes strongly impact users' perception and forwarding intentions, implying the role of format in meme engagement in a strong way (Bülow and Johann 2023).

Research should also consider the temporal and cross-platform dynamics: as cultural artefacts, memes are likely to evolve with societal trends, whose applicability often depends on timing and context. Longitudinal studies, which utilize years of data or datasets from the multiple social media platforms and other subreddits, could make these dynamics clearer. These papers might outline the changes in meme engagement over time and across users of different demographics. Certainly, bringing advanced computational resources to deal with large-scale complexity in multimodal datasets will add further depth and real-world relevance to such analysis. Subsequent research could look at the influence of cultural, linguistic, and regional differences on meme engagement to find community-specific patterns. Looking at how emotional responses and preferred meme formats differ across different demographic groups could also provide useful information for marketers and platform developers. Finally, understanding the interaction between meme engagement and broader societal events—for

example, political and historical events or cultural trends—could help situate the patterns found and explain how memes function as indicators of changes in society. The interaction among user feedback mechanisms, such as comment sentiment and subsequent engagement indicators, needs much more investigation. For instance, knowing whether divisive comments inspire re-engagement or discourage further participation could help platforms develop strategies for productive discussion moderation. Equally, understanding whether an increase in engagement levels with specific meme formats leads to the increased adoption of similar formats may provide insight into changing trends within meme culture.

In summary, future research efforts should not only strive to overcome the limitations acknowledged in this work but also further expand its scope by applying novel methodologies and interdisciplinary approaches. By allowing for a more fine-grained examination and prediction of an expanded range of engagement indicators and exploring details in multimodal interactions, researchers can achieve a better understanding of the digital culture surrounding memes and their impact on users' behaviour, platforms' ecosystems, and social discourse.

9. References

Adami, Elisabetta, and Gunther Kress. 2014. ‘Introduction: Multimodality, Meaning Making, and the Issue of “Text”’. *Text & Talk* 34 (3). https://doi.org/10.1515/text-2014-0007.

Ahmed, Md. Rekib, Neeraj Bhadani, and Ishita Chakraborty. 2021. ‘Hateful Meme Prediction Model Using Multimodal Deep Learning’. In *2021 International Conference on Computing, Communication and Green Engineering (CCGE)*, 1–5. Pune, India: IEEE. https://doi.org/10.1109/CCGE50943.2021.9776440.

Ananth, Mahesh. 2001. ‘Book Review: Explaining Culture: A Naturalistic Approach’. *Philosophy of the Social Sciences* 31 (4): 563–71. https://doi.org/10.1177/004839310103100407.

Arthur Heitmann. 2023. "Arctic Shift Photon Reddit”. https://arctic-shift.photon-reddit.com/download-tool.

Barnes, Kate, Tiernon Riesenmy, Minh Duc Trinh, Eli Lleshi, Nóra Balogh, and Roland Molontay. 2021. ‘Dank or Not? Analyzing and Predicting the Popularity of Memes on Reddit’. *Applied Network Science* 6 (1): 21. https://doi.org/10.1007/s41109-021-00358-7.

Basch, Corey H., Zoe Meleo-Erwin, Joseph Fera, Christie Jaime, and Charles E. Basch. 2021. ‘A Global Pandemic in the Time of Viral Memes: COVID-19 Vaccine Misinformation and Disinformation on TikTok’. *Human Vaccines & Immunotherapeutics* 17 (8): 2373–77. https://doi.org/10.1080/21645515.2021.1894896.

Beer, Jeff. 2019. ‘Inside the Secretly Effective--and Underrated--Way Netflix Keeps Its Shows and Movies at the Forefront of Pop Culture’. Fast Company. 28 February 2019. https://www.fastcompany.com/90309308/by-any-memes-necessary-inside-netflixs-winning-social-media-strategy/.

Behera, Pranati, Mamta, and Asif Ekbal. 2020. "Only Text? Only Image? Or Both? Predicting Sentiment of Internet Memes." In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, 444–452. Indian Institute of Technology Patna, Patna, India: NLP Association of India (NLPAI). https://aclanthology.org/2020.icon-main.60.

Berger, Dale. 2017. ‘Categorical Data Analysis’.

Blackmore, Susan. 2001. ‘EVOLUTION AND MEMES: THE HUMAN BRAIN AS A SELECTIVE IMITATION DEVICE’. *Cybernetics and Systems* 32 (1–2): 225–55. https://doi.org/10.1080/019697201300001867.

Block, Ned. 1993. Review of *Review of Consciousness Explained.*, by Daniel C. Dennett. *The Journal of Philosophy* 90 (4): 181–93. https://doi.org/10.2307/2940970.

Bülow, Lars, and Michael Johann. 2023. ‘Effects and Perception of Multimodal Recontextualization in Political Internet Memes. Evidence from Two Online Experiments in Austria’. *Frontiers in Communication* 7 (January):1027014. https://doi.org/10.3389/fcomm.2022.1027014.

Carreño, Ander, Iñaki Inza, and Jose A. Lozano. 2020. ‘Analyzing Rare Event, Anomaly, Novelty and Outlier Detection Terms under the Supervised Classification Framework’. *Artificial Intelligence Review* 53 (5): 3575–94. https://doi.org/10.1007/s10462-019-09771-y.

Caton, Hiram. 1997. “Review of Explaining Culture: A Naturalistic Approach, by Dan Sperber.” Politics and the Life Sciences 16 (2): 319–323. https://www.jstor.org/stable/4236374.

Cauteruccio, Francesco, Enrico Corradini, Giorgio Terracina, Domenico Ursino, and Luca Virgili. 2022. ‘Investigating Reddit to Detect Subreddit and Author Stereotypes and to Evaluate Author Assortativity’. *Journal of Information Science* 48 (6): 783–810. https://doi.org/10.1177/0165551520979869.

Chen, Wuxing, Kaixiang Yang, Zhiwen Yu, Yifan Shi, and C. L. Philip Chen. 2024. ‘A Survey on Imbalanced Learning: Latest Research, Applications and Future Directions’. *Artificial Intelligence Review* 57 (6): 137. https://doi.org/10.1007/s10462-024-10759-6.

Colacrai, Ernesto, Federico Cinus, Gianmarco De Francisci Morales, and Michele Starnini. 2024. ‘Navigating Multidimensional Ideologies with Reddit’s Political Compass: Economic Conflict and Social Affinity’. arXiv. https://doi.org/10.48550/arXiv.2401.13656.

Colbaugh, Richard, and Kristin Glass. 2011. ‘Detecting Emerging Topics and Trends via Predictive Analysis of ‘Meme’ Dynamics’. In *Proceedings of 2011 IEEE*

International Conference on Intelligence and Security Informatics, 192–94. Beijing, China: IEEE. https://doi.org/10.1109/ISI.2011.5983999.

Dawkins, Richard. 2016. *The Selfish Gene: 40th Anniversary Edition.* Oxford: Oxford University Press.

Dilmegani, Cem. 2024. "OCR Benchmarking: Text Extraction/Capture Accuracy." AI Multiple Research. Accessed November 1, 2024. https://www.aimultiple.com.

Du, Yuhao, Muhammad Aamir Masood, and Kenneth Joseph. 2020. 'Understanding Visual Memes: An Empirical Analysis of Text Superimposed on Memes Shared on Twitter'. *Proceedings of the International AAAI Conference on Web and Social Media* 14 (May):153–64. https://doi.org/10.1609/icwsm.v14i1.7287.

Dudău, Diana Paula, and Florin Alin Sava. 2021. 'Performing Multilingual Analysis With Linguistic Inquiry and Word Count 2015 (LIWC2015). An Equivalence Study of Four Languages'. *Frontiers in Psychology* 12 (July):570568. https://doi.org/10.3389/fpsyg.2021.570568.

Elkin-Koren, Niva. 2010. "User-Generated Platforms." In *Working Within the Boundaries of Intellectual Property*, edited by Rochelle Dreyfuss, Diane L. Zimmerman, and Harry First. Oxford: Oxford University Press. Forthcoming. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1648465.

Ferrara, Emilio, and Zeyao Yang. 2015. 'Quantifying the Effect of Sentiment on Information Diffusion in Social Media'. *PeerJ Computer Science* 1 (September):e26. https://doi.org/10.7717/peerj-cs.26.

Fillmore, Herr Andrew. 2015. 'The Effect of Daily Internet Usage on a Short Attention Span and Academic Performance'. urn:nbn:de:bsz:mit1-opus4-73076.

Ford, Trenton W., Rachel Krohn, and Tim Weninger. 2023. 'Competition Dynamics in the Meme Ecosystem'. *ACM Transactions on Social Computing* 6 (3–4): 1–19. https://doi.org/10.1145/3596213.

Glăveanu, Vlad Petre, and Constance De Saint Laurent. 2021. 'Social Media Responses to the

Pandemic: What Makes a Coronavirus Meme Creative'. *Frontiers in Psychology* 12 (March):569987. https://doi.org/10.3389/fpsyg.2021.569987.

Gordon, Rachel A. 2015. 'REGRESSION ANALYSIS FOR THE SOCIAL SCIENCES'.

Greene, Viveca S. 2019. '"Deplorable" Satire: Alt-Right Memes, White Genocide Tweets, and Redpilling Normies'. *Studies in American Humor* 5 (1): 31–69. https://doi.org/10.5325/studamerhumor.5.1.0031.

Hamilton, Phillip. 2023. "The Barbenheimer Phenomenon: How Memes and Viral Marketing Impacted the Success of the 'Barbie' and 'Oppenheimer' Movies." *Know Your Meme*. Accessed 10 November 2024. https://knowyourmeme.com/editorials/insights/the-barbenheimer-phenomenon-how-memes-and-viral-marketing-impacted-the-success-of-the-barbie-and-oppenheimer-movies.

Hazman, Muzhaffar, Susan McKeever, and Josephine Griffith. 2023. 'Meme Sentiment Analysis Enhanced with Multimodal Spatial Encoding and Facial Embedding'. In , 1662:318–31. https://doi.org/10.1007/978-3-031-26438-2_25.

Holland, Charlotte R. 2020. 'Just a Joke? The Social Impact of Internet Memes'. https://doi.org/10.13140/RG.2.2.17476.04485.

Hossain, Eftekhar, Omar Sharif, and Mohammed Moshiul Hoque. 2022. 'MemoSen: A Multimodal Dataset for Sentiment Analysis of Memes'.

Hunt, John. 2023. *Advanced Guide to Python 3 Programming*. Undergraduate Topics in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-40336-1.

Illowsky, Barbara, and Susan Dean. 2023. 'Introductory Statistics'.

Islam, Md Athikul, Rizbanul Hasan, and Nasir U. Eisty. 2023. 'Documentation Practices in Agile Software Development: A Systematic Literature Review'. arXiv. https://doi.org/10.48550/arXiv.2304.07482.

Kulkarni, Anushka. 2017. 'Internet Meme and Political Discourse: *A Study on the Impact of*

Internet Meme as a Tool in Communicating Political Satire'. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3501366.

Ling, Chen, Ihab AbuHilal, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. 2021. 'Dissecting the Meme Magic: Understanding Indicators of Virality in Image Memes'. arXiv. https://doi.org/10.48550/arXiv.2101.06535.

Literat, Ioana, and Sarah Van Den Berg. 2019. 'Buy Memes Low, Sell Memes High: Vernacular Criticism and Collective Negotiations of Value on Reddit's MemeEconomy'. *Information, Communication & Society* 22 (2): 232–49. https://doi.org/10.1080/1369118X.2017.1366540.

Liu, Henry X., and Shuo Feng. 2024. 'Curse of Rarity for Autonomous Vehicles'. *Nature Communications* 15 (1): 4808. https://doi.org/10.1038/s41467-024-49194-0.

Lynch, Michael P. 2022. 'Memes, Misinformation, and Political Meaning'. *The Southern Journal of Philosophy* 60 (1): 38–56. https://doi.org/10.1111/sjp.12456.

Maalouf, Maher, and Theodore B. Trafalis. 2011. 'Rare Events and Imbalanced Datasets: An Overview'. *International Journal of Data Mining, Modelling and Management* 3 (4): 375. https://doi.org/10.1504/IJDMMM.2011.042935.

McGrady, Marisa, and Kathryn Elizabeth Hamm. 2019. 'A Meme Is Worth a Thousand Words: Universal Communication Through Memes'. https://spiral.lynn.edu/facpubs/751/.

Md. Rekib, Neeraj Bhadani, and Ishita Chakraborty. 2021. 'Hateful Meme Prediction Model Using Multimodal Deep Learning.' In 2021 *International Conference on Computing, Communication and Green Engineering (CCGE),* 1–5. https://doi.org/10.1109/CCGE50943.2021.9776440.

Medvedev, Alexey N., Renaud Lambiotte, and Jean-Charles Delvenne. 2019. 'The Anatomy of Reddit: An Overview of Academic Research'. In *Dynamics On and Of Complex Networks III*, edited by Fakhteh Ghanbarnejad, Rishiraj Saha Roy, Fariba Karimi, Jean-Charles Delvenne, and Bivas Mitra, 183–204. Springer Proceedings in Complexity. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-14683-2_9.

Memon, Mumtaz Ali. 2024. ‘Control Variables: A Review and Proposed Guidelines’. *Journal of Applied Structural Equation Modeling* 8 (2): 1–14. https://doi.org/10.47263/JASEM.8(2)01.

Milli, Smitha, Micah Carroll, Yike Wang, Sashrika Pandey, Sebastian Zhao, and Anca D. Dragan. 2023. ‘Engagement, User Satisfaction, and the Amplification of Divisive Content on Social Media’. arXiv. https://doi.org/10.48550/arXiv.2305.16941.

Milosavljević, Ilija. 2020. ‘THE PHENOMENON OF THE INTERNET MEMES AS A MANIFESTATION OF COMMUNICATION OF VISUAL SOCIETY - RESEARCH OF THE MOST POPULAR AND THE MOST COMMON TYPES’. *MEDIA STUDIES AND APPLIED ETHICS* 1 (1): 9–27. https://doi.org/10.46630/msae.1.2020.01.

Moody-Ramirez, Mia, and Andrew B Church. 2019. ‘Analysis of Facebook Meme Groups Used During the 2016 US Presidential Election’. *Social Media + Society* 5 (1): 2056305118808799. https://doi.org/10.1177/2056305118808799.

Morabito, Greg. 2019. ‘Depression Shouldn’t Be a #Brand Engagement Strategy’. Eater. 4 February 2019. https://www.eater.com/2019/2/4/18211094/sunny-d-tweet.

Nawawi, Elly Atiqah, Muhammad Faiz Izzudin, Nur Atiqah Mohd Zulkafli, and Siti Sarah Januri. 2020. ‘Identifying the Factors Affecting Internet Memes to Become Viral on Social Media’.

Nissenbaum, Asaf, and Limor Shifman. 2018. ‘Meme Templates as Expressive Repertoires in a Globalizing World: A Cross-Linguistic Study’. *Journal of Computer-Mediated Communication* 23 (5): 294–310. https://doi.org/10.1093/jcmc/zmy016.

Newton, Giselle, Michele Zappavigna, Kerryn Drysdale, and Christy E. Newman. 2022. ‘More than Humor: Memes as Bonding Icons for Belonging in Donor-Conceived People’. *Social Media + Society* 8 (1): 205630512110690. https://doi.org/10.1177/20563051211069055.

Ngo, Triet Minh. 2021. ‘Meme Marketing: How Viral Marketing Adapts to the Internet Culture’, 76. https://scholarworks.uni.edu/hpt/484/.

OpenAI. 2021. "DALL·E: Creating Images from Text.". https://openai.com/index/dall-e/?utm_source=chatgpt.com.

Perez-Riverol, Yasset, Laurent Gatto, Rui Wang, Timo Sachsenberg, Julian Uszkoreit, Felipe Da Veiga Leprevost, Christian Fufezan, et al. 2016. 'Ten Simple Rules for Taking Advantage of Git and GitHub'. Edited by Scott Markel. *PLOS Computational Biology* 12 (7): e1004947. https://doi.org/10.1371/journal.pcbi.1004947.

Pranesh, Raj Ratn, and Ambesh Shekhar. 2020. 'MemeSem:A Multi-Modal Framework for Sentimental Analysis of Meme via Transfer Learning'.

Proferes, Nicholas, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. 'Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics'. *Social Media + Society* 7 (2): 20563051211019004. https://doi.org/10.1177/20563051211019004.

Rathi, Navrang, and Pooja Jain. 2023. 'The Power and Pitfalls of Internet Memes in Promoting Brand Awareness and Engagement on Social Media'.

Ross, Andrew S., and Damian J. Rivers. 2017. 'Digital Cultures of Political Participation: Internet Memes and the Discursive Delegitimization of the 2016 U.S Presidential Candidates'. *Discourse, Context & Media* 16 (April):1–11. https://doi.org/10.1016/j.dcm.2017.01.001.

Sah, Tanmay, and Kayden Jordan. 2024. 'Decoding Reddit Memes Virality'. https://doi.org/10.21203/rs.3.rs-4351378/v1.

Shaik, Afroz, Rahul Arulkumaran, Ravi Kiran Pagidi, and Dr S P Singh. 2021. 'Utilizing Python and PySpark for Automating Data Workflows in Big Data Environments' 5 (4).

Sherratt, Victoria, Kevin Pimbblet, and Nina Dethlefs. 2023. 'Multi-Channel Convolutional Neural Network for Precise Meme Classification'. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, 190–98. Thessaloniki Greece: ACM. https://doi.org/10.1145/3591106.3592275.

Shifman, Limor. 2014. *Memes in Digital Culture*. MIT Press Essential Knowledge.

Cambridge, Massachusetts: The MIT Press.

Shyalika, Chathurangi, Ruwan Wickramarachchi, and Amit P. Sheth. 2024. 'A Comprehensive Survey on Rare Event Prediction'. *ACM Computing Surveys*, October, 3699955. https://doi.org/10.1145/3699955.

Srinidhi, Ketavarapu, Ala Harika, Srujan Sai Voodarla, and Jarugula Abhiram. 2024. 'Sentiment Analysis of Social Media Comments Using Natural Language Procesing'. In *2024 International Conference on Inventive Computation Technologies (ICICT)*, 762–68. https://doi.org/10.1109/ICICT60155.2024.10544849.

Statista. 2024. "Reddit - Statistics & Facts". https://www.statista.com/topics/5672/reddit/#topicOverview.

Steinert, Steffen. 2021. 'Corona and Value Change. The Role of Social Media and Emotional Contagion'. *Ethics and Information Technology* 23 (S1): 59–68. https://doi.org/10.1007/s10676-020-09545-z.

Tobias Schmidt, Felix Schiller, Maximilian Götz, and Christian Wolff, "A Corpus of Memes from Reddit: Acquisition, Preparation and First Case Studies," in *INFORMATIK 2023. Designing Futures: Zukünfte gestalten* (Gesellschaft für Informatik e.V., 2023), 795–804, https://doi.org/10.18420/inf2023_89.

Treude, Christoph, Justin Middleton, and Thushari Atapattu. 2020. 'Beyond Accuracy: Assessing Software Documentation Quality'. arXiv. https://doi.org/10.48550/arXiv.2007.10744.

Tsugawa, Sho, and Hiroyuki Ohsaki. 2017. 'On the Relation between Message Sentiment and Its Virality on Social Media'. *Social Network Analysis and Mining* 7 (1): 19. https://doi.org/10.1007/s13278-017-0439-0.

Vassio, Luca, Michele Garetto, Carla Chiasserini, and Emilio Leonardi. 2021. 'Temporal Dyamics of Posts and User Engagement of Influencers on Facebook and Instagram'. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 129–33. Virtual Event Netherlands: ACM.

https://doi.org/10.1145/3487351.3488340.

Veselovsky, Veniamin, and Ashton Anderson. 2023. ‘Reddit in the Time of COVID’. arXiv. https://doi.org/10.48550/arXiv.2304.10777.

Wagener, Albin. 2021. ‘The Postdigital Emergence of Memes and GIFs: Meaning, Discourse, and Hypernarrative Creativity’. *Postdigital Science and Education* 3 (3): 831–50. https://doi.org/10.1007/s42438-020-00160-1.

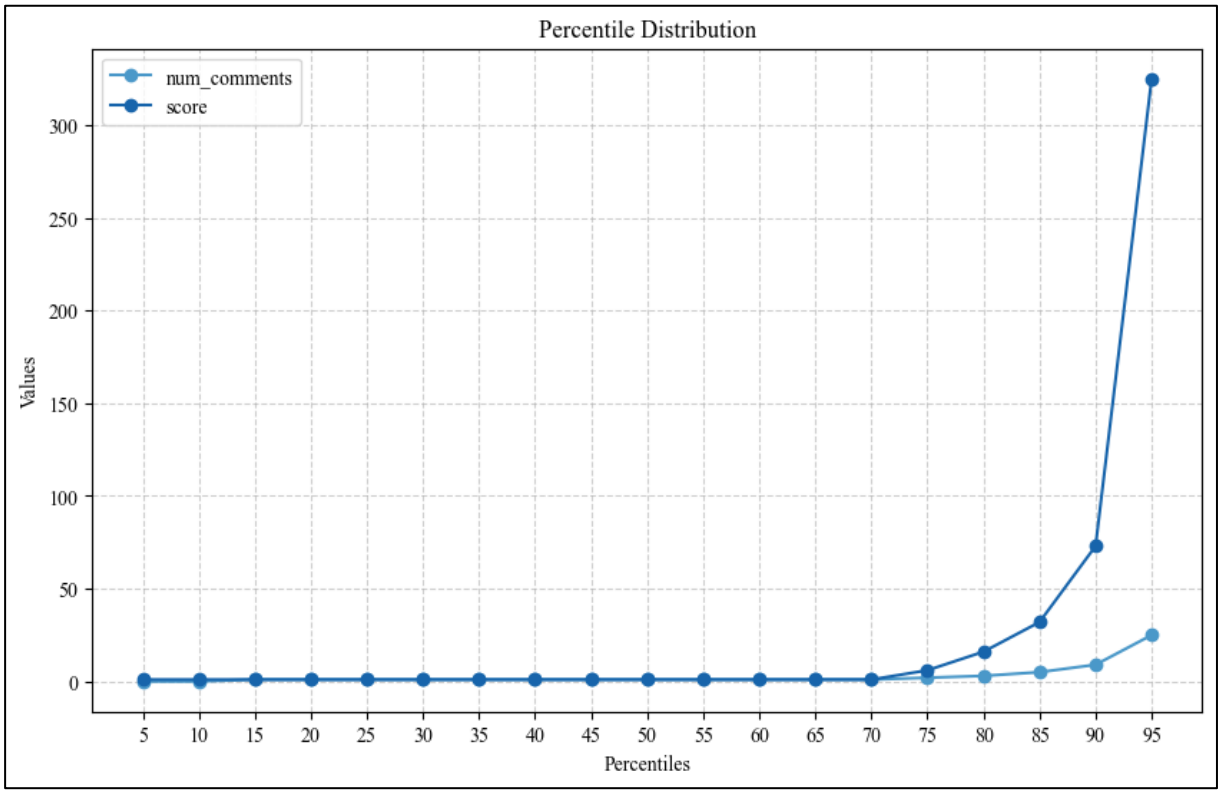
Walther, Joseph B., and Jeong-woo Jang. 2012. ‘Communication Processes in Participatory Websites’. *Journal of Computer-Mediated Communication* 18 (1): 2–15. https://doi.org/10.1111/j.1083-6101.2012.01592.x.

Weng, Lilian, Filippo Menczer, and Yong-Yeol Ahn. 2013. ‘Virality Prediction and Community Structure in Social Networks’. *Scientific Reports* 3 (1): 2522. https://doi.org/10.1038/srep02522.

Weng, Lilian, Filippo Menczer, and Yong-Yeol Ahn. 2014. ‘Predicting Successful Memes Using Network and Community Structure’. *Proceedings of the International AAAI Conference on Web and Social Media* 8 (1): 535–44. https://doi.org/10.1609/icwsm.v8i1.14530.

Zhong, Yang, and Bhiman Kumar Baghel. 2024. ‘Multimodal Understanding of Memes with Fair Explanations’. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2007–17. Seattle, WA, USA: IEEE. https://doi.org/10.1109/CVPRW63382.2024.00206.

10. Appendices



Appendix 1: Percentile Distribution of Posts for Score and the Number of Comments Original 2023 Dataset

	score								num_comments								awards							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
quarter																								
Q1	20862.000	328.495	2391.008	0.000	1.000	4.000	41.000	84714.000	20862.000	18.062	135.103	0.000	1.000	1.000	4.000	8383.000	20862.000	0.008	0.090	0.000	0.000	0.000	0.000	1.000
Q2	16187.000	369.313	2861.488	0.000	1.000	1.000	25.000	88465.000	16187.000	12.133	82.368	0.000	1.000	1.000	2.000	2774.000	16187.000	0.005	0.073	0.000	0.000	0.000	0.000	1.000
Q3	20567.000	34.255	894.764	1.000	1.000	1.000	1.000	59791.000	20567.000	8.792	67.025	0.000	1.000	1.000	2.000	3517.000	20567.000	0.008	0.090	0.000	0.000	0.000	0.000	1.000
Q4	18597.000	138.473	1424.295	1.000	1.000	1.000	1.000	50697.000	18597.000	8.485	65.074	0.000	1.000	1.000	2.000	2355.000	18597.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Appendix 2: Descriptive Statistics for Quarter - Score, Number of Comments and Awards 2023 Dataset

	score								num_comments								awards							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
author_premium																								
False	73493.000	190.837	1844.759	0.000	1.000	1.000	6.000	88465.000	73493.000	11.091	84.884	0.000	1.000	1.000	2.000	6724.000	73493.000	0.005	0.071	0.000	0.000	0.000	0.000	1.000
True	2720.000	766.789	4556.197	0.000	1.000	1.000	40.250	77359.000	2720.000	35.552	220.192	0.000	1.000	2.000	6.000	8383.000	2720.000	0.022	0.146	0.000	0.000	0.000	0.000	1.000

Appendix 3: Descriptive Statistics for Author Premium - Score, Number of Comments and Awards 2023 Dataset

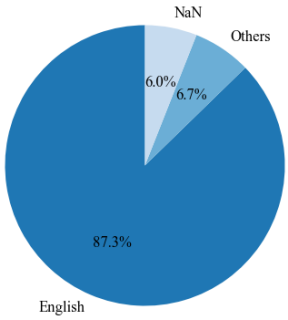
	score								num_comments								awards							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
removed_post																								
0	32178.000	403.686	2859.022	0.000	1.000	7.000	46.000	88465.000	32178.000	19.942	116.504	0.000	1.000	3.000	6.000	6724.000	32178.000	0.011	0.104	0.000	0.000	0.000	0.000	1.000
1	44035.000	70.876	980.605	0.000	1.000	1.000	1.000	77359.000	44035.000	6.135	71.107	0.000	1.000	1.000	1.000	8383.000	44035.000	0.002	0.041	0.000	0.000	0.000	0.000	1.000

Appendix 4: Descriptive Statistics for Removed Post - Score, Number of Comments and Awards 2023 Dataset

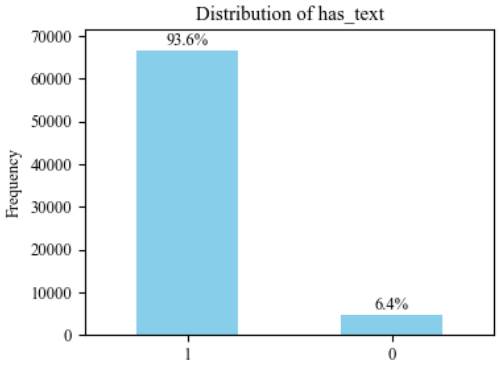
	score								num_comments								awards							
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max
locked																								
False	70452.000	211.352	2018.055	0.000	1.000	1.000	6.000	88465.000	70452.000	11.679	92.947	0.000	1.000	1.000	3.000	8383.000	70452.000	0.006	0.075	0.000	0.000	0.000	0.000	1.000
True	5761.000	211.881	1886.626	0.000	1.000	2.000	9.000	84714.000	5761.000	15.459	97.020	0.000	1.000	1.000	3.000	2568.000	5761.000	0.005	0.073	0.000	0.000	0.000	0.000	1.000

Appendix 5: Descriptive Statistics for Locked - Score, Number of Comments and Awards 2023 Dataset

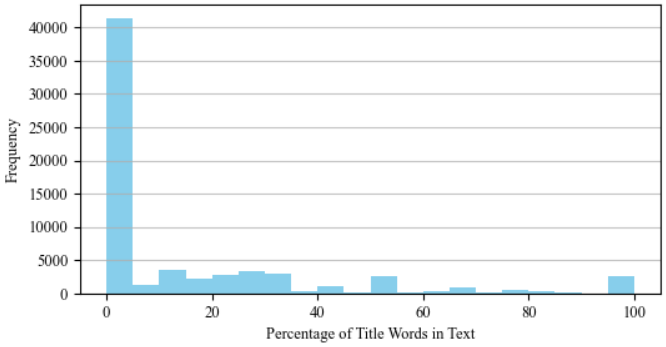
Appendix 6: "Two Buttons" meme



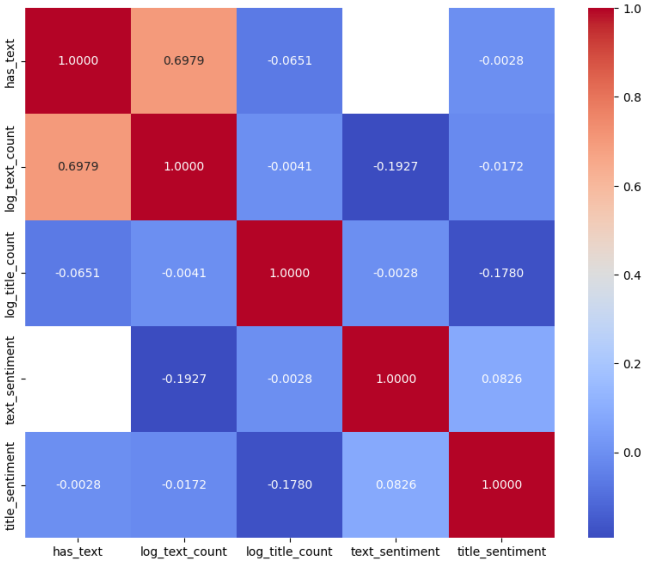
Appendix 7: Language Distribution: no text, English and Other



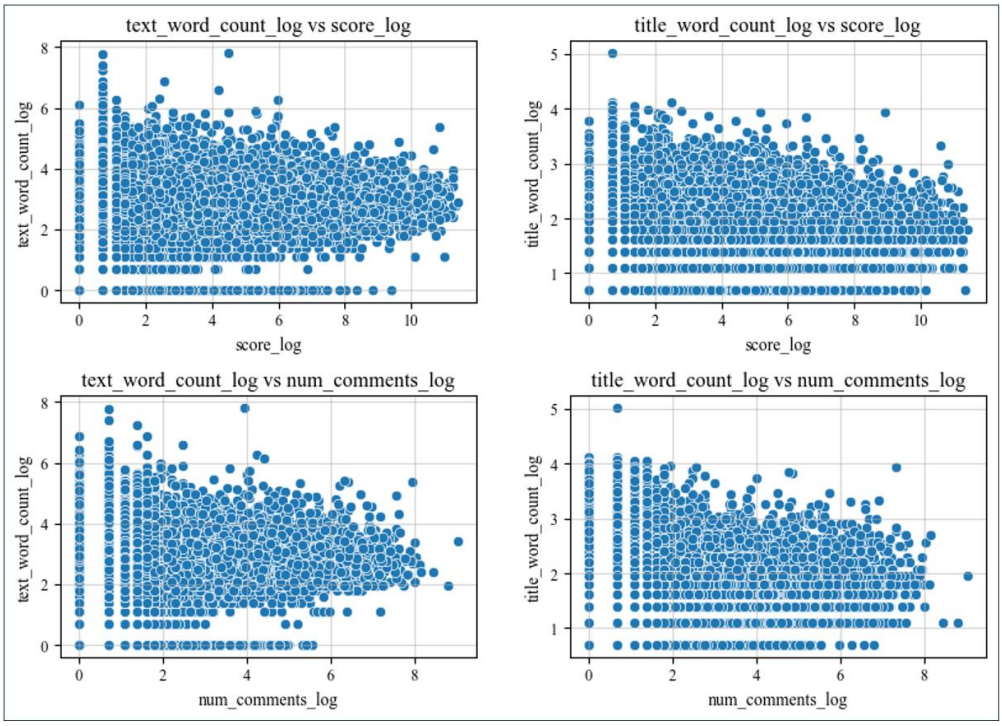
Appendix 8: Distribution for the presence of text



Appendix 9: Distribution for the percentage of title word in text in meme



Appendix 10: Correlation matrix for the independent variables



Appendix 11: Scatter Plots of Log-Transformed Word Counts vs Engagement Metrics

For the presence of text on the number of comments:

```
Shapiro-Wilk Normality Test Results:
Category: 1, W-statistic: 0.7112, p-value: 0.0000
Category: 0, W-statistic: 0.5401, p-value: 0.0000

Levene's Test for Homogeneity of Variance:
W-statistic: 697.4109, p-value: 0.0000
```

For the presence of text on score:

```
Shapiro-Wilk Normality Test Results:
Category: 1, W-statistic: 0.6449, p-value: 0.0000
Category: 0, W-statistic: 0.3143, p-value: 0.0000

Levene's Test for Homogeneity of Variance:
W-statistic: 908.5528, p-value: 0.0000
```

For the sentiment of text on the number of comments:

```
Shapiro-Wilk Normality Test Results:
Category: 4.0, W-statistic: 0.7137, p-value: 0.0000
Category: 3.0, W-statistic: 0.7106, p-value: 0.0000
Category: 5.0, W-statistic: 0.7120, p-value: 0.0000
Category: 1.0, W-statistic: 0.7078, p-value: 0.0000
Category: 2.0, W-statistic: 0.7425, p-value: 0.0000

Levene's Test for Homogeneity of Variance:
W-statistic: 6.6579, p-value: 0.0000
```

For the sentiment of text on score:

```
Shapiro-Wilk Normality Test Results:
Category: 4.0, W-statistic: 0.6522, p-value: 0.0000
Category: 3.0, W-statistic: 0.6493, p-value: 0.0000
Category: 5.0, W-statistic: 0.6526, p-value: 0.0000
Category: 1.0, W-statistic: 0.6358, p-value: 0.0000
Category: 2.0, W-statistic: 0.6779, p-value: 0.0000

Levene's Test for Homogeneity of Variance:
W-statistic: 4.3075, p-value: 0.0017
```

For the sentiment of title on the number of comments:

```
Shapiro-Wilk Normality Test Results:
Category: 4.0, W-statistic: 0.7132, p-value: 0.0000
Category: 1.0, W-statistic: 0.7170, p-value: 0.0000
Category: 5.0, W-statistic: 0.6796, p-value: 0.0000
Category: 3.0, W-statistic: 0.6994, p-value: 0.0000
Category: 2.0, W-statistic: 0.7722, p-value: 0.0000

Levene's Test for Homogeneity of Variance:
W-statistic: 95.9574, p-value: 0.0000
```

For the sentiment of title on score:

```
Shapiro-Wilk Normality Test Results:
Category: 4.0, W-statistic: 0.6447, p-value: 0.0000
Category: 1.0, W-statistic: 0.6541, p-value: 0.0000
Category: 5.0, W-statistic: 0.5880, p-value: 0.0000
Category: 3.0, W-statistic: 0.6470, p-value: 0.0000
Category: 2.0, W-statistic: 0.7195, p-value: 0.0000

Levene's Test for Homogeneity of Variance:
W-statistic: 82.8957, p-value: 0.0000
```

Appendix 12: Shapiro-Wilk and Levene's Test Results

For the presence of text on the number of comments:

```
Kruskal-Wallis Test Result:
H-statistic: 560.8818, p-value: 0.0000
```

For the presence of text on score:

```
Kruskal-Wallis Test Result:
H-statistic: 1001.1895, p-value: 0.0000
```

For the sentiment of text on the number of comments:

```
Kruskal-Wallis Test Result:
H-statistic: 30.0932, p-value: 0.0000
```

For the sentiment of text on score:

```
Kruskal-Wallis Test Result:
H-statistic: 31.4180, p-value: 0.0000
```

For the sentiment of title on the number of comments:

```
Kruskal-Wallis Test Result:
H-statistic: 157.6746, p-value: 0.0000
```

For the sentiment of title on score:

```
Kruskal-Wallis Test Result:
H-statistic: 358.8846, p-value: 0.0000
```

Appendix 13: Kruskal-Wallis Test Results

For the presence of text on presence of awards:

```
Chi-Square Test Result:
Chi2 statistic: 14.6621, p-value: 0.0001
```

For the sentiment of text in meme on presence of awards:

```
Chi-Square Test Result:
Chi2 statistic: 9.6968, p-value: 0.0459
```

For the sentiment of title on presence of awards:

```
Chi-Square Test Result:
Chi2 statistic: 13.2500, p-value: 0.0101
```

Appendix 14: Chi-Square Test Results