Aryansh Tiwari

DATASET #19 : Creative Common Memes

February 9th, 2025

Cracking Reddit Meme Virality : Colour & Size Matters

Memes and their impact

Memes play a crucial role in shaping opinions, spreading ideas, and influencing culture. Predicting their virality helps brands boost engagement, politicians sway public opinion, and activists raise awareness. By leveraging data and AI, organizations can create highly shareable content, maximizing reach and impact in the digital age. Memes are powerful tools for communication, marketing, and social influence.

Introduction

What if the secret to meme virality isn't just the joke, but something more *subconscious*? Our analysis suggests that factors like **Hue, Saturation, and Thumbnail Size** have a significant impact—not necessarily because people *consciously* prefer them, but because they subtly influence engagement. A meme with **balanced colors and a prominent thumbnail** might be easier to notice, making it more likely to grab attention in a fast-scrolling feed. It's not that users actively seek out memes with the "perfect" hue or saturation, rather, their brains might be instinctively drawn to certain visual elements without them even realizing it.

While humor and relatability remain crucial, the way a meme is **visually presented** might play an unconscious role in determining whether it gets shared or ignored. Content creators looking to maximize engagement might not need to change the joke—but tweaking the **color balance, contrast, and thumbnail size** could make a surprising difference.

Dataset

Using the **PRAW** library for Reddit scraping, we collected **1,000 top posts** from multiple meme-related subreddits to analyze meme virality based on visual and textual attributes. The selected subreddits included **r/memes, r/meme, r/PoliticalHumour, r/ProgrammerHumour, r/RelationshipMemes, r/terriblefacebookmemes, r/dank_meme, r/dankmemes, r/Animememes, r/antimeme, r/HistoryMemes, r/wholesomememes, r/wholesomemes, and r/surrealmemes**. Although our initial extraction aimed for 13,000 posts, various issues—such as deleted posts, broken image links, and unavailable content—resulted in a **final dataset of 10,477 memes** (Figure-1).

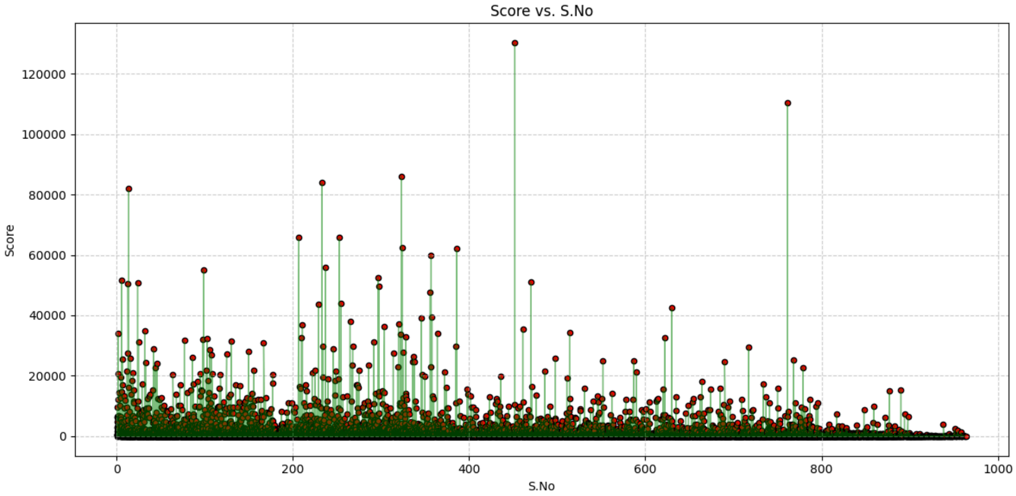

Figure - 1 : Post score plotted in an ordered manner (against serial number)

For each post, we gathered detailed metadata, including the **post title, URL, score, creation timestamp, number of comments, top comments, associated subreddit, and unique post ID**. Additionally, we extracted subreddit-level information such as **subreddit name, creation date, subscriber count, and description** to contextualize meme trends. To maintain consistency and focus on image-based content, we filtered out non-image posts, retaining only those in **JPG, JPEG, or PNG** formats. This refinement reduced the dataset to **9,399 valid meme posts**, ensuring a clean and structured dataset for further analysis.

To classify memes based on their virality, we applied a quantile-based labeling approach. For each subreddit, we selected the **top 5th quantile** of posts—those with the highest engagement metrics, such as post score and number of comments—and labeled them as **"Viral."** Similarly, the **bottom 5th quantile** of posts—those with the lowest engagement—were labeled as **"Non-Viral."** This filtering process ensured that we captured distinct differences in meme popularity while maintaining subreddit-specific trends. After applying this classification, our dataset was reduced to **1,180 data points** (Figure - 2)**,** providing a well-balanced and focused subset for analyzing the key factors influencing meme virality. This left us with 698 Viral and 482 Non-Viral posts (Figure - 3).

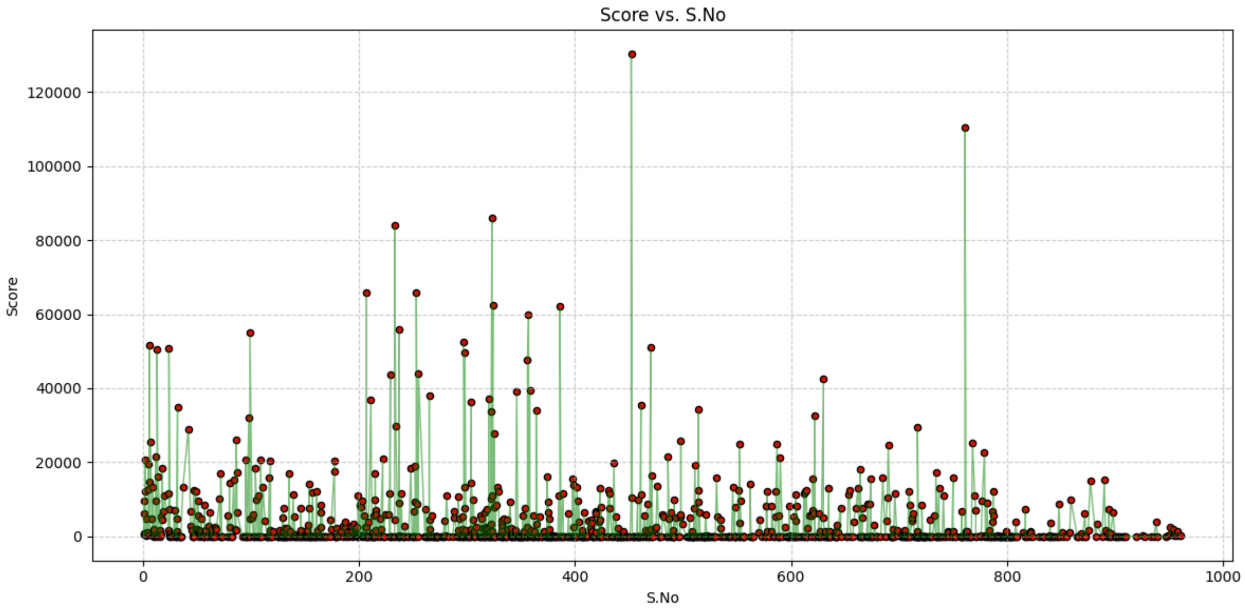

Figure -2 : Post score plotted in an ordered manner (Only Viral and Non-Viral Posts)

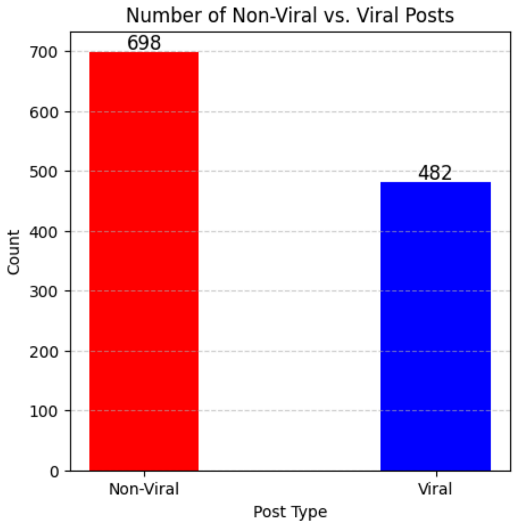

Figure -3 : Number of Viral vs Non-Viral posts

Methodology

To further analyze **meme virality**, we extracted various **image-related metadata** for each meme in our dataset. These features included **Thumbnail_Height, Thumbnail_Width, Top_Color, Top_Color_Name, Average_Hue, Average_Saturation, Average_Value, Average_Red, Average_Green, Average_Blue,** and **Hex_Color.** By quantifying the **visual properties** of each meme, we aimed to understand how **color composition** and **image structure** influence engagement levels. However, **18 images** encountered **format inconsistencies or extraction errors**, leaving us with a final dataset of **1,162 memes** for further analysis.

Beyond visual attributes, we conducted **sentiment analysis** on the **top comments and post titles** using the **TextBlob library.** The top **20 comments** for each post were categorized based on their **sentiment polarity**, and we recorded the count of **positive, negative, and neutral comments.** Additionally, each **post title** received a **sentiment score**, capturing its **emotional tone.** These sentiment-related variables provided key insights into **audience reactions** and the role of emotional cues in **meme virality.**

To predict whether a meme would be **Viral** or **Non-Viral**, we trained a **LightGBM (LGBM) classifier** using a feature set that combined **visual and sentiment-based attributes.** The input features included **Positive_Count, Negative_Count, Neutral_Count, Overall_Sentiment_Label_Encoded, Thumbnail_Height, Thumbnail_Width, Average_Hue, Average_Saturation, Average_Value, Average_Red, Average_Green, Average_Blue,** and **Post_Title_Sentiment_Value.** This diverse feature set allowed the model to analyze both **visual composition and audience sentiment,** improving its ability to predict meme popularity.

The **LGBM model** demonstrated **strong predictive performance**, achieving an **accuracy of 80.69%** and an **ROC AUC Score of 0.85.** The **ROC Curve** (Figure - 4) illustrates the model's ability to separate viral and non-viral memes, demonstrating a **high true positive rate with minimal false positives.**

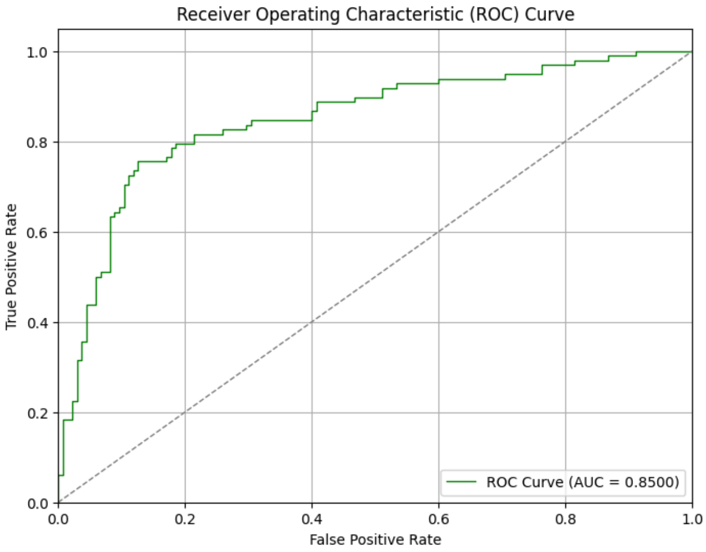

Figure - 4 : Receiver Operating Characteristic Curve for LGBM Predictions

Additionally, the **Confusion Matrix** (Figure - 5) provides a breakdown of **correctly and incorrectly classified memes,** further confirming the model's effectiveness. These results highlight the **significant role of visual aesthetics and sentiment-driven engagement** in meme virality, offering valuable insights for **content creators, marketers, and researchers** seeking to optimize meme-based communication strategies.

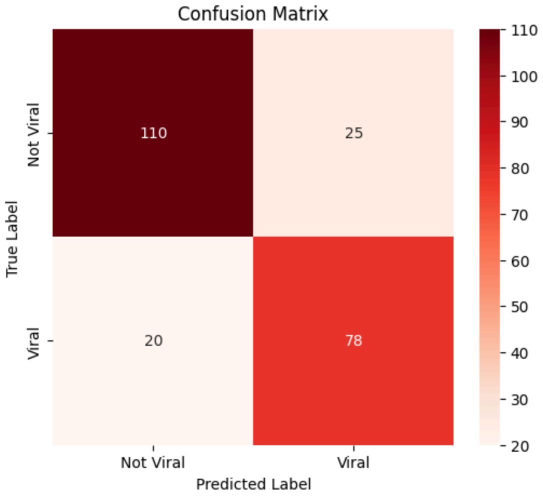

Figure - 5 : Confusion Matrix for LGBM Predictions

Interesting Findings

The **feature importance plot from the LightGBM model** (Figure - 6) illustrates the significance of various features in predicting **meme virality**. The x-axis represents the **feature importance score**, while the y-axis lists the features ranked by their contribution to the model's predictions. The color bar on the right provides an additional visual cue for **importance levels**, with higher values indicating greater influence on the model's decision-making.

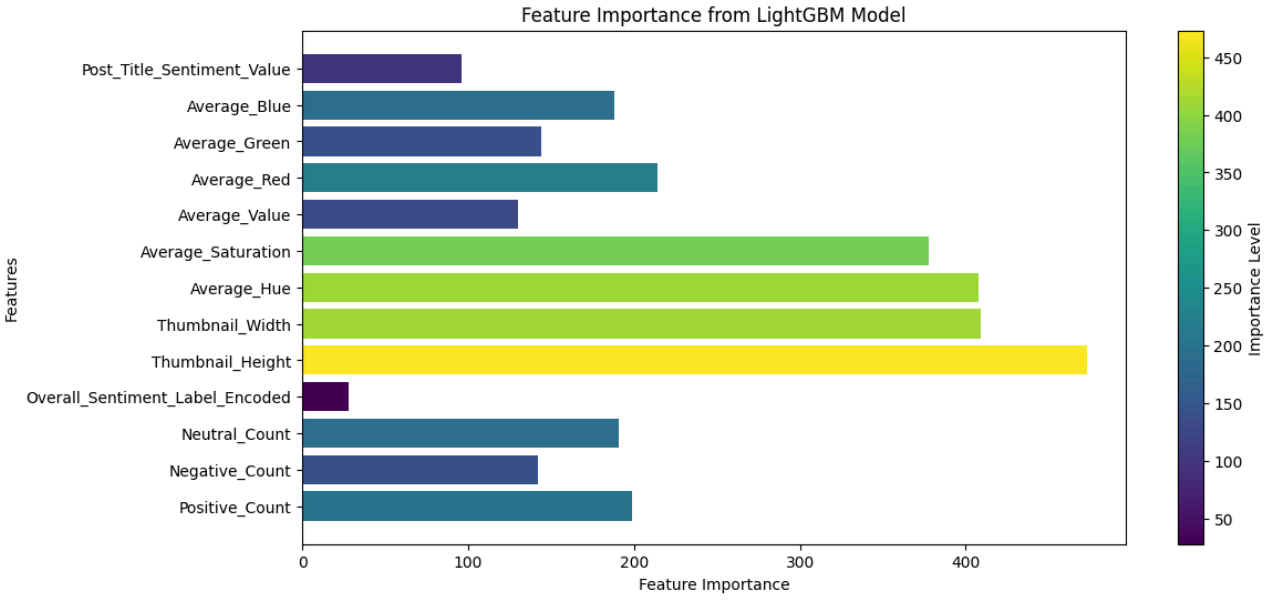

Figure - 6 : Feature Importance Plot for LGBM

As shown in Figure - 6, **Thumbnail_Height** is the most influential feature, suggesting that larger meme thumbnails contribute significantly to virality. Other critical features include **Thumbnail_Width, Average_Hue, and Average_Saturation**, reinforcing the importance of **image properties** in meme success. Additionally, **Average_Red, Average_Blue, and Average_Green** have notable impacts, highlighting the role of **color composition** in viral memes.

In terms of **text-based sentiment**, **Positive_Count and Neutral_Count** show strong influence, indicating that the number of positive and neutral audience reactions is crucial in determining meme virality. Meanwhile, **Post_Title_Sentiment_Value** and **Overall_Sentiment_Label_Encoded** have relatively lower importance, suggesting that while sentiment matters, **visual and engagement-based factors** play a more dominant role.

Overall, Figure - 6 emphasizes that a combination of **image attributes, sentiment scores, and audience engagement metrics** determines meme virality. This insight can be leveraged to optimize content creation strategies for higher engagement.

Ultimately, meme virality isn't just about humor—it's a blend of **content and subconscious appeal**. While people may not actively think about a meme's **hue, saturation, or thumbnail size**, these elements can subtly **influence visibility and engagement**. This suggests that the art of meme-making isn't just about crafting the funniest caption, but also about understanding how visual cues shape online interactions. As digital spaces become more crowded, creators who optimize both **the message and the medium** might just have the edge in making their memes truly unforgettable.