Reddit Comment Engagement Analysis

DATA 410 Project Proposal

Group Members: Aleric Govender, Jordi Capdevila Maso, Aarav Gosalia

Date: Feb 11th, 2025

1. Introduction

Reddit is one of the largest social media platforms where users engage in discussions through comments. Understanding what drives engagement in Reddit comments can provide insights into online discourse and community interaction. In this project, we analyze comment scores and replies across multiple subreddits, identifying key linguistic and metadata-based predictors. Using web scraping and regression analysis, we aim to uncover factors influencing comment popularity, such as sentiment, length, and timing.

2. Dataset Description

Our dataset is curated through Reddit posts and comments, focusing on 10 diverse subreddits, including r/AskReddit, r/datascience, r/technology, r/movies, r/gaming, r/books, r/health, r/Showerthoughts, r/UnpopularOpinion, and r/FloridaMan. This mix ensures a variety of informative, casual, and controversial discussions, capturing a range of engagement patterns.

Currently, we have web scraped data from r/AskReddit, r/datascience, and r/technology. The remaining subreddits will be included in the final dataset as additional data is collected.

The dataset consists of 17 features, categorized as:

- Continuous Variables: Sentiment score, comment score, text length, word count, comment age, parent score, user karma, account age.
- Categorical Variables: Subreddit name, comment hour, comment day.
- Binary Variables: Contains emoji, contains question, contains profanity, is early comment (within one hour of parent).

Comment engagement is assessed through comment score (upvotes minus downvotes) and number of replies, which serve as the dependent variables in our regression analysis. The dataset currently consists of 19,067 observations spanning 17 attributes, providing a comprehensive foundation for our study.

3. Data Collection Process

The data is scraped in real-time using the PRAW API, ensuring up-to-date information. The collection process involves extracting top posts from each subreddit, retrieving associated comments, applying text processing techniques, and storing the structured data in CSV format.

For sentiment analysis, we use VADER (Valence Aware Dictionary and Sentiment Reasoner), a lexicon-based sentiment analysis tool optimized for social media text, which assigns a sentiment score to each comment based on the presence of positive, negative, and neutral words. Additionally, SpaCy is used for natural language processing (NLP) tasks, such as detecting whether a comment contains a question by analyzing sentence structure. To maintain data quality, we will perform basic cleaning to remove deleted or bot-generated comments before analysis.

4. Scientific Questions

Our project seeks to answer the following key questions:

1. Does sentiment (positive/negative) influence upvotes and replies?
2. How do metadata features like comment timing and user karma affect engagement?

5. Challenges & Limitations

- **API Rate Limits:** Reddit imposes restrictions on the number of requests per minute, limiting data collection speed.
- **Data Bias**: Engagement may vary significantly across subreddits, leading to subreddit-specific trends.
- **Noise in Text Data:** Comments may include sarcasm, slang, and emojis, making sentiment analysis challenging.
- **Outliers & Manipulation:** Viral comments or artificially boosted comments (e.g., bots) may distort engagement patterns.

6. Preliminary Exploratory Data Analysis (EDA)

Summary Statistics :

Statistic	Comment Score	Sentiment Score	Text Length	Word Count	User Karma	Account Age (days)
Min	-136	-0.9981	1	1	-100	0
1st Quartile	1	-0.3294	47	8	3,740	743.2
Median	3	0.0000	98	18	20,156	2,142
Mean	44.74	0.02321	177.1	31.23	71,216	2,381.2
3rd Quartile	11	0.4019	206	37	71,440	4,036
Max	22,077	0.9995	5,181	898	6,272,141	7,041
Missing (NAs)	-	-	-	-	73	73

The Comment Score, our main response variable, represents upvotes minus downvotes and shows a highly skewed distribution. The median score is 3, but the mean is 44.74, indicating a few viral comments significantly raise the average. Scores range from -136 (heavily downvoted) to 22,077 (highly upvoted), with most comments receiving modest engagement (1st quartile = 1, 3rd quartile = 11).

To enhance the robustness of our analysis, further data collection and a more even distribution across subreddits are necessary. Mitigating this skewness through log transformation or outlier handling will improve the reliability of our model in capturing the key drivers of Reddit comment engagement.