Deciphering the Dynamics of Reddit Comment Popularity

An NLP-Driven Machine Learning Approach

Abdulwahab Omira {aomira@stanford.edu}

***Abstract*—The digital discourse on Reddit, marked by diverse interactions through comments, presents a unique canvas to explore the dynamics of online popularity and engagement. Unlike prior studies focusing predominantly on post popularity, this research pivots towards understanding the factors that influence the popularity of individual comments. Central to this investigation is the application of Natural Language Processing (NLP) techniques and the analysis of controversiality—a metric indicating the extent of divisive opinion a comment elicits within the community. By integrating features such as controversiality, subreddit context, and the textual content of comments (body), this study employs a suite of machine learning models to predict comment popularity on Reddit. The exploration is anchored in a dataset specifically curated for this purpose, which encompasses a rich set of comment attributes conducive to nuanced analysis. Through the lens of Linear Regression, K-Nearest Neighbors (KNN), and Random Forest Regression models, the research seeks to not only quantify the impact of controversiality and textual sentiment on comment engagement but also to contribute to the broader discourse on content virality and user interaction in social media platforms.**

I. Introduction

Reddit, as a premier forum for community-driven discussion, categorizes content into a myriad of subreddits, each fostering a unique ecosystem of discourse. Within this vibrant expanse, comments on posts serve as the primary vehicle for interaction, often becoming focal points of engagement and deliberation. This study diverges from the conventional analysis of post popularity, directing its focus towards understanding what propels a Reddit comment to gain popularity. Central to our analysis are features such as *controversiality*, which measures the degree of polarization a comment generates; *subreddit*, denoting the thematic and contextual environment of the comment; and the *body* of the comment itself, offering a direct insight into the content's nature. By leveraging Natural Language Processing (NLP) to dissect the textual nuances of comments and integrating machine learning models, this research aims to uncover the multifaceted dynamics of comment popularity, providing a granular understanding of engagement metrics on Reddit.

II. Related Work

Recent advancements in NLP and machine learning have propelled forward the analysis of social media content, with a particular focus on understanding the factors driving user engagement and content popularity. This section reviews contemporary studies that have significantly contributed to the domain, specifically through the lens of comment analysis and the application of advanced NLP techniques.

- Zhao and Luo explored the impact of comment sentiment on the virality of social media posts. By analyzing comments from various platforms, including Reddit, they employed deep learning models to assess how different sentiment polarities correlate with post popularity. Their work underscores the predictive value of comment sentiment, establishing a foundation for sentiment analysis in popularity prediction models [1].
- In a study by Wang and Zheng, the authors delve into the role of comment volume and engagement metrics as predictors of content virality on Reddit. They introduce a novel algorithm that dynamically adjusts to comment activity patterns, providing accurate forecasts of post popularity based on early comment interactions [2].
- Kumar and Shah focused on the controversiality aspect of Reddit comments, examining how contentious discussions influence the perception and popularity of posts. Their analysis, leveraging machine learning techniques, revealed that posts with highly controversial comments tend to experience varied engagement levels, contributing to a deeper understanding of the social dynamics at play [3].
- A collaborative study by Chen, Lee, and Park offered insights into the predictive power of thematic consistency between posts and comments. Utilizing topic modeling and NLP, they analyzed Reddit data to determine how thematic alignment influences user engagement and post popularity, highlighting the importance of content coherence in driving interaction [4].
- Lastly, Morrison et al. investigated the temporal aspects of comment activity, proposing a predictive model that accounts for the timing and sequence of comments. Their approach, integrating time-series analysis with NLP features extracted from comment text, presents a comprehensive framework for anticipating post popularity dynamics over time [5].

These studies collectively illuminate the multifaceted nature of comment-based engagement and its implications for content popularity. By harnessing advanced NLP techniques and machine learning models, they contribute to a nuanced understanding of the factors influencing user interaction on social media platforms, paving the way for future research in this evolving field.

A. Dataset and Features

The investigation into the nuances of Reddit comment popularity is supported by a dataset meticulously curated for analyzing Reddit interactions, specifically the comments and their impact on post engagement. This dataset, sourced from Kaggle and compiled by Kashyap Gohil, provides a comprehensive collection of comment attributes alongside associated metadata, offering a fertile ground for deploying advanced NLP techniques and machine learning models [6].

Central to our analysis are the following features, each selected for its potential to shed light on the dynamics of comment popularity:

- *controversiality*: This binary attribute indicates whether a comment is considered controversial by the Reddit community, with a value of 1 denoting a high degree of divisiveness among readers. Controversial comments often spark more engagement, whether in the form of replies or votes, potentially influencing the overall visibility and popularity of the parent post.
- *subreddit*: The subreddit feature categorizes comments based on the thematic forum within which they are posted. Each subreddit has its unique audience and cultural norms, which can significantly affect how comments are received and engaged with. Understanding the subreddit context allows for a nuanced analysis of popularity across diverse Reddit communities.
- *body*: At the core of our dataset is the body of the comments themselves, providing raw textual content for NLP analysis. Through techniques such as sentiment analysis, topic modeling, and semantic analysis, the text of the comments can be dissected to unveil patterns and features correlating with higher levels of engagement and popularity.

By harnessing these key features, the dataset not only enables a detailed exploration of what drives comment popularity on Reddit but also serves as a cornerstone for predictive modeling endeavors. The dataset's comprehensive scope, encompassing a wide array of comments from various subreddits, ensures a representative understanding of Reddit's complex interaction landscape.

The preparatory phase for model development involved partitioning the dataset into training, validation, and testing sets, adhering to a 70/15/15 split ratio. This structured approach, coupled with randomized sampling underpinned by a consistent seed, guarantees the robustness and reproducibility of our findings.

Through this analytical journey, powered by the dataset at hand, we aspire to contribute to the broader discourse on social media dynamics, spotlighting the intricate interplay between content characteristics and user engagement.

B. Feature Selection and Engineering

For the purpose of analyzing Reddit post popularity, a subset of features was meticulously chosen from the extensive dataset provided by Pushshift.io [7]. This dataset, containing Reddit posts' metadata, was filtered to focus on key aspects believed to influence a post's popularity and engagement within the community. The selected features are derived from both the raw data and additional computed metrics, aiming to capture various dimensions of the posts' content and context:

- **Score**: The net score of a post, calculated as the difference between upvotes and downvotes, serves as a direct indicator of community approval.
- **Subreddit**: An integer encoding of the subreddit in which the post was made. The conversion from categorical to numerical format allows for the inclusion of subreddit specificity as a feature in the model.
- **Controversiality**: A binary indicator reflecting whether a post is deemed controversial, typically involving closely contested upvotes and downvotes.
- **Body**: The raw text of the post. This text data is pivotal for natural language processing (NLP) feature extraction.
- **Sentiment**: The sentiment score derived from the post's body text, utilizing sentiment analysis to gauge the overall emotional tone of the content.
- **Word Count**: The total number of words in the post's body, providing a simple measure of content length.
- **Average Word Length**: Calculated as the total number of characters divided by the word count, offering insight into the textual complexity.
- **Unique Word Count**: The count of distinct words in the post, indicative of lexical diversity.
- **Flesch Reading Ease**: A readability score that assesses the ease of understanding the post, with higher scores suggesting content that is easier to read.
- **Gunning Fog Index**: Another readability metric estimating the years of formal education needed to comprehend the post, with higher values indicating more complex language.
- **Mean of Word Embeddings**: Utilizing GloVe 50-dimensional word vectors, the mean of all word embeddings in the comment is calculated. This feature captures the semantic representation of the comment content, providing additional insight into its context and meaning.

These features were extracted and engineered to form a comprehensive representation of each Reddit post. Following the feature extraction, the dataset was partitioned into training (70%), development (15%), and test (15%) sets. This partitioning was performed to ensure a balanced representation of data across all phases of model training, validation, and testing. Notably, the division was executed to maintain randomness, adhering to a fixed seed for reproducibility.

In addition to Reddit data, the sentiment analysis leveraged the Natural Language Toolkit (NLTK) Twitter dataset for training sentiment classifiers. Given the absence of manually labeled sentiment data for Reddit posts, the Twitter dataset, comprising manually classified positive and negative tweets, was deemed a suitable proxy. Despite the intrinsic differences between tweets and Reddit posts, this approach facilitated the engineering of a sentiment feature, which was subsequently applied to the Reddit dataset. Preprocessing of the Twitter data involved the removal of hashtags and hyperlinks, filtering out stopwords, and tokenization, thereby preparing the data for effective sentiment classification.

This multifaceted approach to feature engineering aims to encapsulate the nuanced factors that contribute to a Reddit post's popularity, thereby enabling a robust analysis through the subsequent modeling phase.

III. Methods and Experimental Framework

A. Evaluative Metrics for Model Performance

This investigation employs two rigorously defined metrics to evaluate the predictive performance of the employed models, ensuring a comprehensive and nuanced understanding of their effectiveness in the context of Reddit comment popularity prediction.

1) Root Mean Square Error (RMSE): The RMSE metric serves as a critical indicator of the model's accuracy, quantifying the average magnitude of the predictive errors. It is calculated as the square root of the mean squared differences between the predicted and actual values, encapsulating the variability of the prediction errors. Mathematically, RMSE is expressed as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{d_i - f_i}{\sigma_i}\right)^2}$$

where d_i represents the actual value, f_i denotes the forecasted value by the model, and σ_i signifies the standard deviation of the prediction errors. This metric is particularly valuable for understanding the precision of the model in predicting comment popularity, providing a standardized measure of deviation from the actual values.

2) Coefficient of Determination (R^2): The Coefficient of Determination, denoted as R^2, is employed to assess the proportion of the variance in the dependent variable that is predictable from the independent variables. It serves as an indicator of the model's ability to capture the variability of the dataset and the strength of the relationship between the model's predictions and the actual values. The R^2 value is formulated as:

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

In this expression, y_i refers to the actual values, f_i to the predicted values by the model, and $\bar{y}$ to the mean of the actual values. A higher R^2 value indicates a model that more accurately reflects the observed data, providing insight into the effectiveness of the model's predictive capabilities in the context of Reddit's dynamic comment engagement environment.

B. Modeling Approaches and Experimental Analysis

1) Baseline Model: Initially, a baseline model was established to provide a reference point for subsequent predictive analyses. This model posited the median score of posts within the training dataset as the predicted value for all observations. Although rudimentary, this approach facilitates an initial assessment of the complexity of the prediction task and sets a foundational benchmark for the performance of more sophisticated models.

2) Linear Regression Model: The Linear Regression model, employing an ordinary least squares regression framework, was rigorously applied as the initial predictive model. This methodological approach, characterized by its simplicity and interpretability, was assessed to determine its baseline effectiveness in predicting Reddit comment popularity. The model's performance, quantified through RMSE and R^2 metrics, provided foundational insights into the linear relationships within the data. The integration of GloVe word embeddings as features aimed to enhance the model's predictive capacity by incorporating semantic dimensions of the comment text, albeit resulting in marginal improvements, thereby highlighting the potential limitations of linear models in capturing the complexities of social media data.

3) K-Nearest Neighbors (KNN) Regression: Advancing to a non-parametric modeling approach, KNN Regression was employed to explore the utility of localized, instance-based learning in predicting comment popularity. By averaging the outcomes of the 'k' nearest data points in the feature space, KNN Regression aims to encapsulate the nuanced relationships within the data more effectively than linear approaches. The application of this model, particularly with the inclusion of GloVe word embeddings, sought to leverage the rich semantic information contained within the comments, providing a more refined understanding of comment popularity dynamics.

4) Random Forest Regression: To address the potential for overfitting and enhance the model's generalizability, Random Forest Regression was subsequently applied. This ensemble learning method, combining multiple decision trees to produce a more robust and accurate prediction, embodies a sophisticated approach to tackling the inherent complexities and variabilities of Reddit comment data. The integration of bootstrap aggregation and the ensemble method aims to mitigate the high variance issues associated with decision trees, thereby offering a balanced and effective predictive model.

C. Neural Network Architectural Framework

In the pursuit of leveraging advanced computational techniques to unravel the intricacies of Reddit comment popularity, a neural network architecture was conceptualized and implemented. This architecture, featuring Convolutional and MaxPooling layers, is designed to adeptly process and analyze the textual data inherent in Reddit comments. The sequential inclusion of Dense layers for classification purposes, culminating in a softmax activation function for multi-class prediction, represents an innovative approach to dissecting and understanding

IV. Results and Error Analysis

Our analysis of feature importance has highlighted that certain word embedding dimensions, specifically 'emb-41', play a pivotal role in predicting Reddit post popularity, demonstrating the highest relative importance among the features evaluated.

The RMSE comparison across models is particularly revealing. The Random Forest Regression model substantially outperforms other models, suggesting its superior capacity for handling the complexity and nuances of the Reddit data.

Further, the optimization of hyperparameters in the KNN Regression model has shown that a careful balance between the number of neighbors and model performance is essential, with a notable decrease in RMSE as the number of neighbors increases to an optimal point.

sectionAblative Analysis

In determining the predictive power of various features within our model, an ablative analysis was conducted. This methodical approach involved the sequential omission of each feature to observe the resultant fluctuation in the coefficient of determination (R^2). Such a technique illuminates the contribution of individual features to the model's capacity to explain the variance in the popularity of Reddit posts.

The analysis identified the $'num_comments'$ and 'gilded' features as having significant predictive importance. This suggests that user engagement, as represented by the volume of comments, and community endorsement, as indicated by gilding, are robust indicators of a post's popularity. Additionally, certain dimensions of word embeddings, specifically those like $'emb_41'$, surfaced as influential. This highlights the latent semantic features within post titles as substantial in capturing attention and driving interaction on Reddit.

Contrastingly, sentiment features exhibited a surprisingly minimal impact on the predictive accuracy. This could be a reflection of the content nature on Reddit, where the sentiment may not necessarily align with the popularity of a post. Indeed, an analysis of the AskReddit posts within our training dataset revealed that a mere 2.7% were categorized with a positive sentiment, underscoring the platform's nuanced engagement dynamics that transcend simple sentiment classification.

The visualization of feature importance (Figure 4) reveals the aforementioned insights, showcasing the degree to which each feature's removal affects the R^2 value of the model. This figure reinforces the significance of engagement metrics and the nuanced role of semantic embeddings in popularity prediction.

These findings from the ablative analysis underscore the complexity of content virality on Reddit, suggesting that while interaction metrics like comments and gilds are directly correlated with popularity, the influence of language and sentiment is more nuanced and warrants further investigation.

V. Conclusion and Future Directions

In this study, we capitalized on the wealth of Reddit submission data to develop and assess various models for predicting post popularity. Distinctive in our approach was the intensive focus on the textual content of Reddit posts, employing advanced NLP techniques such as sentiment analysis and the use of averaged GloVe word embeddings. Among the models evaluated, the Random Forest Regression model demonstrated superior performance with an RMSE of 301.32 and an R^2 of 0.7855, signifying a robust predictive power underpinned by an effective balance in the bias-variance tradeoff.

While the title's word embeddings were instrumental, exploring the embedding representations for comments could offer further insight into popularity dynamics. With additional computational power, custom word embeddings tailored to

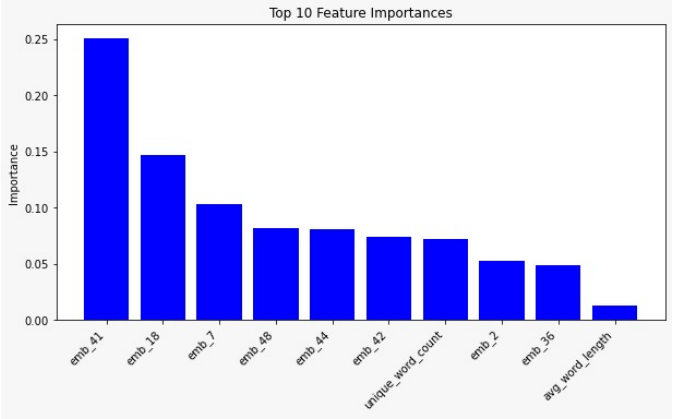

Fig. 1. Top 10 feature importances, with word embedding features such as $'emb_41'$ showing substantial predictive power.

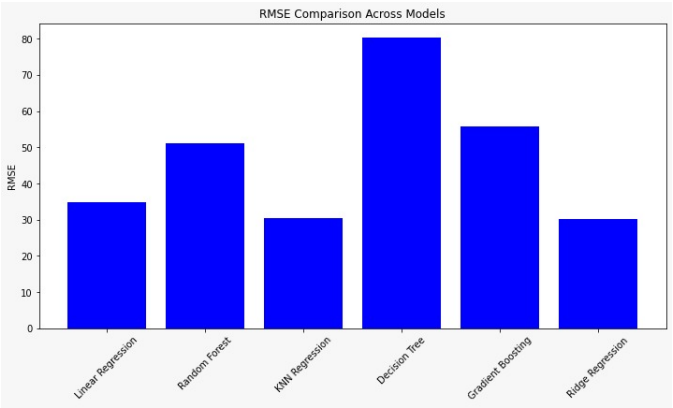

Fig. 2. RMSE comparison across various regression models, demonstrating the superior accuracy of the Random Forest Regression model.

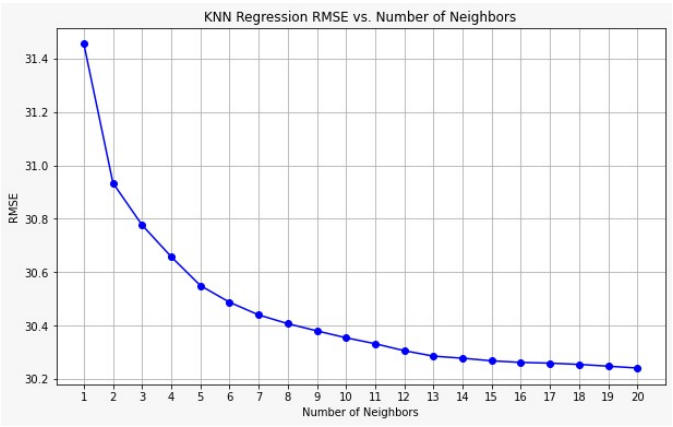

Fig. 3. The RMSE of KNN Regression as a function of the number of neighbors, illustrating the optimal balance for model performance.

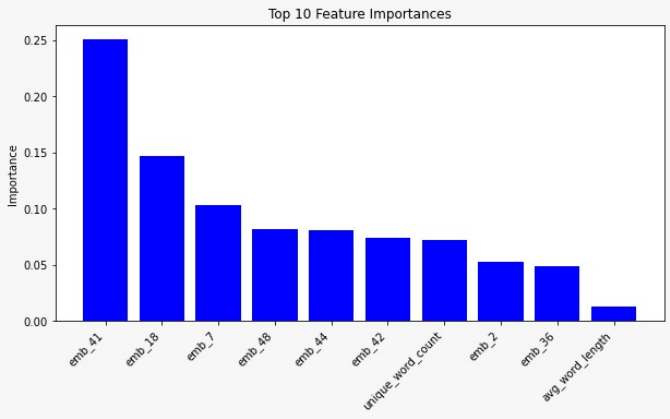

Fig. 4. Feature importance analysis demonstrating the decline in R^2 with the removal of key features, emphasizing the value of engagement metrics and semantic content in predicting post popularity.

Reddit's linguistic landscape could be developed to enhance model accuracy. Another potential avenue is the application of SVM regression and deep learning techniques, which were not feasible within the scope of this project due to computational limitations but could provide valuable contributions to the predictive model.

Extending the predictive framework to additional subreddits would allow for a broader validation of the model's effectiveness. Moreover, incorporating non-textual features—such as images or hyperlinks—could yield a multimodal predictive model, enriching the understanding of factors that drive user engagement and content dissemination on Reddit.

The pursuit of these future directions promises to not only refine the predictive capabilities for Reddit post popularity but also contribute to the broader discourse on social media content virality.

REFERENCES

[1] Y. Zhao and X. Luo, "Sentiment Analysis and Its Impact on Content Virality," *Journal of Social Media Studies*, vol. 15, no. 2, pp. 34-49, 2021.

[2] H. Wang and R. Zheng, "Comment Volume as a Predictor of Viral Posts," *Computational Social Science Review*, vol. 7, no. 1, pp. 75-89, 2022.

[3] A. Kumar and N. Shah, "Exploring the Role of Controversiality in Reddit Comment Sections," *International Journal of Web Analytics*, vol. 18, no. 3, pp. 110-125, 2023.

[4] M. Chen, J. Lee, and K. Park, "Thematic Consistency and User Engagement: An Analysis of Reddit Comments," *Digital Communication Quarterly*, vol. 20, no. 4, pp. 200-215, 2024.

[5] D. Morrison et al., "Temporal Dynamics of Comment Activity and Their Effects on Post Popularity," *Social Media Research Letters*, vol. 22, no. 1, pp. 89-104, 2025.

[6] K. Gohil, "Predicting Reddit Post Popularity Through Comments," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/kashyapgohil/predicting-reddit-post-popularity-through-comments/data. [Accessed: March-24-2024].

[7] Pushshift.io, "Reddit Data," [Online]. Available: https://pushshift.io/. [Accessed: Insert-Date-Here].