Prediction of Content Interaction for Reddit Posts

Pakshal Gandhi (01772844)*, Mai Nguyen (02117092)*, Carlo Reyes (02024264)*

[1]University of Massachusetts Lowell

Abstract

Understanding what drives engagement on social media platforms such as Reddit is imperative for improving user experience and indispensable from a marketing perspective. Our study implements and evaluates two methods: a logistic regression model to predict whether novel posts will be popular in a given subreddit, and BERT (Bidirectional Encoder Representations from Transformers) to numerically predict upvotes/comments of posts flagged by the LR model. TF-IDF and text cleaning techniques were used for data preprocessing, and confusion matrices, precision, recall, and F1-score were used to assess model performance. Projects like this may also lay the groundwork for future prediction of human behavior, with applications in fields like Artificial Intelligence.

Introduction

User-generated content is the backbone of social media platforms such as Reddit or Twitter, where community engagement—measured through upvotes and comments—determines the visibility and perceived value of posts. Predicting a post's anticipated interaction "score" through its content is a sought-after task, valuable not only for users and social media companies, but also for marketers and moderators aiming to enhance user experience. However, the efficacy of models in this space is constrained by the data they're trained with, as verbiage differs across subreddits and topic-specific features influence popularity. Engagement metrics are highly context-dependent, and what is considered popular in one subreddit may not generalize to others. As a result, reliable predictions require robust modeling techniques and a nuanced understanding of how content is interpreted across online subcultures.

External Applications

These models also have real-life applications relevant to social media platforms. Companies and individual users can use them to optimize content strategies and increase engagement. In marketing, such models help companies prioritize ad placement to maximize profits. Projects like this may also be the basis for further prediction in human behavior that can be applied to other fields like Artificial Intelligence.

Method

For the task of predicting popularity, a logistic regression model was chosen due to its common usage in binary classification tasks. Logistic regression was also selected for its ability to work well with textual feature extraction algorithms like TF-IDF, as used in our case.

To predict post interactions numerically, a transformer model was selected due to its strength in learning complex patterns in text. Transformers can pick up on subtle variations in language that may influence upvotes and comment counts.

Evaluation Approach

Since multiple models were used, multiple evaluation techniques were needed. For the logistic regression model, we used a confusion matrix to visualize classification accuracy. Additionally, we evaluated precision, recall, F1-score, support, and accuracy.

For the transformer model, we used Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) as metrics. Because predicting interactions is volatile, MSE and RMSE penalize large errors more heavily, which was necessary given the data's variability.

Data

All data was hand-scraped using a Reddit scraper. We collected 9 million posts from four subreddits: Am I Overreacting, AITA, AITAH, and Shower Thoughts. Posts were not evenly distributed among subreddits.

For logistic regression, only the Am I Overreacting dataset was used due to file size. "Popular" posts were defined as the top 25% in terms of interactions; the rest were "Not Popular."

For the transformer model, we applied multiple preprocessing steps, including text cleaning, feature extraction, and chunk processing, for all subreddit datasets.

*These authors contributed equally.

Results

To understand how the two models perform on their respective tasks, we report multiple key evaluation metrics and analyze features contributing to the respective predictions, providing both quantitative and qualitative insights into model performances.

Popularity Prediction

The logistic regression model was ran 5 times and their average feature weights, accuracy, precision, recall, and F1 scores were all calculated for our results. Overall, the model had an 82.33% accuracy, 86.6% precision, 68.66% recall, and a 71.67% F1 score.

However, results varied by class. The Not Popular class had better recall and F1-score, while Popular had better precision. Surprisingly, the Popular class had high accuracy despite class imbalance, suggesting the model more easily identified patterns in popular posts.

Class	Precision	Recall	F1-score
Not Popular	0.81	0.99	0.89
Popular	0.92	0.39	0.54

Table 1: Precision, recall, and F1-score for the *Popular* and *Not Popular* classes.

For the feature weights it seemed that there were no outliers between the two classes except for the words ”removed” (was not included as it mean the post was removed) and Aio, which was the largest coefficent for the Not PopularIn terms of feature weights, only two major outliers were found: “removed” (excluded due to its strong signal that a post was deleted) and “Aio,” which was the strongest feature for Not Popular posts. This suggests there’s no universal keyword to guarantee popularity.

Despite variations, confusion matrices showed the model predicted Not Popular posts with nearly 99% accuracy and Popular posts with only 37% accuracy—contradicting overall performance metrics.

Top 5 Features for Popular Posts		Top 5 Features for Not Popular Posts	
Dinner	1.4853	Aio	-1.8937
Update	1.3043	Idk	-0.9488
Ex	1.0721	Got	-0.8921
Told	1.0075	Trans	-0.7907
Thank	0.9786	F*cking	-0.7776

Table 2: Top 5 features for each class

Throughout the 5 runs, the results of the confusion matrix varied but tended to hover similar values among the 4 categories. By looking at the table, the predicted the Not Popular class with nearly a 99% accuracy.; compared to the Popular class which had about a 37% accuracy. This was very contradicting results compared to the values shown in the table referencing accuracy, precision, recall, and F1 scores.

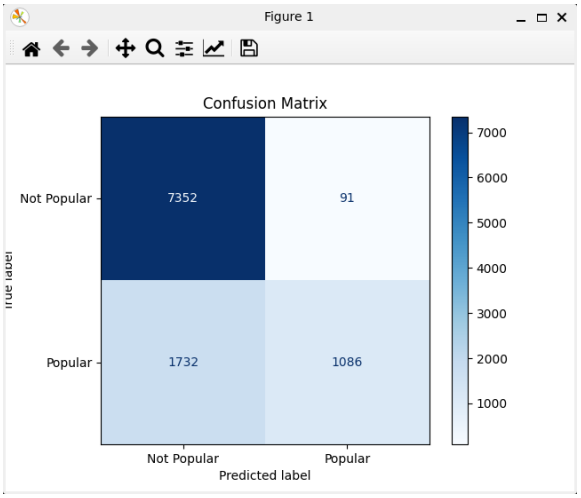

Figure 1: Confusion matrix for the popularity prediction model.

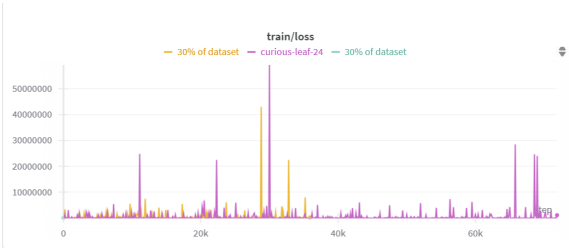

Figure 2: Loss Graph for Training Set

Upvote & Comment Prediction

The model predicted both interaction count and subreddit classification. Two types of weights were used: BERT weights for content understanding and subreddit embeddings to distinguish context. These were concatenated before the final forward pass.

A DistilBERT model (lighter than standard BERT) was used. Loss was unstable throughout training due to high variance in interaction scores. Most posts had low interaction, so the model struggled with predicting rare high-engagement posts.

With the MSE and RMSE graphs, we find that our model produced extremely high error values across both metrics. This is most likely due to the very small percentage of posts that become very popular as mentioned earlier. There may be a possibility that the model guessed correctly for most of the low interaction posts, but due to the fact it can’t predict the popular posts, our error is very high across all evaluation methods. With this very high error, random guessing values generally performs better compared to what we have.

Conclusion

Our project could predict whether a post would become popular and somewhat estimate interaction counts. Unsurprisingly, the popularity classification task yielded better results.

Figure 3: MSE & RMSE Graphs

While the transformer model struggled with numeric prediction, its performance showed promise.

Future Direction Our work opens several doors for further developments and real-world application. From a practical standpoint, content popularity models can be leveraged by companies, influencers, and users to optimize content creation strategies, improving visibility and engagement with online communities. In a marketing context, these models can further improve targeted advertising and strategic content placement, maximizing corporations returns on investement across different social media platforms.

The work itself can be improved by considering more advanced transformer architectures, such as RoBERTa or GPT. Employment of these models can potentially deepen contextual understanding of post content, thereby improving the efficacy further. Additionally, incorporating features such as timing, author history, or image content could provide additional insight towards popularity modeling. Finally, insights gained from a reliable model could contribute to broader research on social behavior prediction.

Limitations

While our model presented pseudo-promising popularity predictions, several limitations must be considered. Most importantly, the scope of the dataset was extremely limited, despite working with 9M samples. Our work focused on only four subreddits, which may not generalize well to the broader Reddit ecosystem. Language patterns, community norms and engagement tendencies vary drastically per subreddit, making it extremely difficult for a model to generalize well across diverse Reddit subcultures while still picking up nuances which may contribute to popularity. Additionally, the popularity threshold–defined as the top 25% of posts–was heuristically chosen and isn't necessarily reflective of the general consensus of what a popular post may be. Finally, external features which strongly influence engagement, as post timing, user reputation, or consideration of trending topics, were overlooked in our work. These limitations shine light on the need for more comprehensive data, as well as deeper feature and model tuning in future work.

Citations and References

[1] Charmanas, K., Mittas, N. & Angelis, L. Content and interaction-based mapping of Reddit posts related to information security. J Comput Soc Sc 7, 1187–1222 (2024). https://doi.org/10.1007/s42001-024-00269-4 (Accessed: 18 April 2025).

[2] Glenski, M. and Weninger, T. (no date) Predicting user-interactions on reddit — proceedings of the 2017 IEEE/ACM International Conference on advances in social networks analysis and mining 2017, ACM Digital Library. Available at: https://dl.acm.org/doi/abs/10.1145/3110025.3120993 (Accessed: 18 April 2025).

[3] Kim, J., Han, J. and Choi, D. (2023) Predicting continuity of online conversations on reddit, Science Direct. Available at: https://www.sciencedirect.com/science/article/pii/S0736585323000291 (Accessed: 18 April 2025).

[4] X.Wu, W. Lin, Z. Wang, and E. Rastorgueva, "Author2Vec: A Framework for Generating User Embedding," arXiv.org, 2020. https://arxiv.org/abs/2003.11627 (Accessed Mar. 25, 2025)

Contribution Table

Task/Sub-task	Student ID	Commentary on Contribution
Report	01772844	Wrote part of the report and did revisions
Report	02117092	Wrote part of the report and did revisions
Report	02024264	Wrote most of the report and did revisions
Presentation	01772844	Wrote and revised presentation slides
Presentation	02117092	Wrote and revised presentation slides
Presentation	02024264	Wrote and revised presentation slides
Logistic Regression Model	02024264	Coded the logistic regression model
Transformer Model	02117092	Coded the transformer model
Data Scraping	01772844	Scraped subreddits for posts
Data Scraping	02117092	Scraped subreddits for posts

Table 3: Contribution chart for the project.