Project Report: Text Mining and Predictive Modeling on TikTok Creator Bios

Jiaxi Xian, Liuqiang Huang, Chunning Zhang,
Zichen He, Ziming Mo, Luxin Zhang

1. Introduction

In this project, we analyze a dataset of TikTok creator profiles to examine whether profile-level textual and behavioral signals encode predictive information about creator visibility. Specifically, we study how biography text and aggregated engagement metrics relate to a creator's follower status. Because video-level content is not observed in the dataset, this project does not attempt to model content quality or short-term content success. Instead, it focuses on whether observable profile representations contain sufficient signal to distinguish creators with high follower counts.

The objectives of the project are twofold. First, we apply text mining techniques to extract structure from short, informal creator biographies. Second, we build predictive models to evaluate which textual and numerical features are most informative for classifying high-follower creators. The dataset includes roughly one thousand TikTok accounts with variables such as follower count, engagement metrics, and free-text bios. Although bios are brief, they function as compact self-presentation devices, making them a useful setting for studying text preprocessing, vectorization, and model interpretability.

The modeling framework incorporates concepts covered in class, including tokenization, bag-of-words, TF-IDF, logistic regression, and random forests, and extends them through clustering and model comparison. By comparing linear and non-linear classifiers, the project assesses both interpretability and predictive robustness in a high-dimensional, text-rich setting.

2. Data Preparation and Text Processing

Our analysis began with the Kaggle TikTok User Profiles Dataset, which served as the foundation for the project. Several cleaning steps were required to ensure reproducibility and compatibility with the modeling pipeline. Numerical fields with impossible values, such as negative follower counts, were removed. Because follower counts were highly skewed, a log transformation was applied to improve interpretability.The biography field underwent standard preprocessing, including lowercasing, removal of URLs, punctuation, and non-informative characters, and consolidation of whitespace. Emojis were preserved because they carry communicative meaning and frequently appear in social media bios. Tokenization and vocabulary construction were completed using TensorFlow's TextVectorization layer with a 2,000-token limit. This produced a sparse bag-of-words matrix with dimensions of 995×2000.

TF-IDF representations were then generated to highlight creator-specific language patterns for downstream modeling.

3. Exploratory Text Analysis

A first pass through the tokenized corpus involved computing the global frequency of each token across all bios. This identifies the types of words creators rely on most (see Appendix Figure 1).

The most frequent words relate to cross-platform promotion (e.g., “instagram,” “youtube”), creator identity (“artist,” “music”), and calls to action such as “follow.” This reflects the functional use of TikTok bios as spaces for signaling niche and directing audiences to other platforms. We also reviewed engagement metrics like engagement rate; although not visualized, these variables are used later as inputs for predictive modeling.

In addition to textual patterns, examining the distribution of engagement-related variables shows there is a strong, approximately linear relationship between the followers and likes (see Appendix Figure 2). Verified accounts tend to cluster in the upper-right region of the plot, suggesting that verification status is associated with higher overall engagement and visibility.

4. Predictive Modeling Approach

4.1 Logistic Regression

The first predictive task classified creators into two groups based on follower count, where high-follower status (top quartile) is used as a proxy for long-term creator visibility rather than content-level success. Textual features derived from TF-IDF representations of creator biographies were combined with log-transformed numeric engagement metrics and derived interaction variables to form a unified feature matrix. All numeric features were standardized prior to model estimation. Model performance was evaluated using accuracy, precision, recall, F1-score, and the area under the ROC curve (see Appendix Figure 3). The ROC curve provides a threshold-independent summary of the model’s discriminative ability.

The logistic regression model demonstrated reliable predictive performance, with the ROC curve indicating meaningful separation between high- and low-follower creators. To interpret the drivers of this classification, coefficient estimates were examined (see Appendix Figure 4). Coefficient signs and magnitudes provide a transparent interpretation: positive coefficients increase the predicted probability of being classified as a high-follower creator, while negative coefficients reduce it.

The results indicate that both textual content and engagement-related variables contribute significantly to creator popularity. In particular, coefficients associated with cross-platform promotional terminology (e.g., references to Instagram or external contact information) tend to be positive, suggesting that creators who explicitly promote off-platform presence are more

likely to belong to the high-follower group. These findings highlight the role of deliberate self-presentation and branding strategies alongside measurable engagement metrics.

4.2 Random Forest Classifier

To incorporate non-linear interactions ignored by logistic regression, a random forest classifier was fitted using the same feature set. Because TF-IDF generates high-dimensional representations, only the top tokens by variance were included to reduce sparsity. Random forest feature importance highlights variables with substantial predictive value, offering a complementary interpretation to coefficient-based analysis.

The random forest model confirmed several patterns found by logistic regression, while also elevating some text-based features that interacted with structured variables. This indicates that creator bios contain meaningful information but require flexible models to capture subtle signaling effects (see Appendix Figure 5).

Importantly, the random forest model confirms the relevance of creator bios while demonstrating that their predictive value is context-dependent. Textual signals appear more influential when combined with behavioral indicators of engagement and consistency, suggesting that bios function as subtle signals of professionalism, branding, or audience targeting rather than as standalone predictors. Together, these results indicate that flexible, non-linear models are better suited for uncovering the nuanced ways in which self-presentation and engagement jointly shape creator popularity.

4.3 Boosting

Building on the random forest results, gradient boosting and XGBoost models were estimated to further assess model robustness and predictive stability. Both models produced results that were highly consistent with the random forest, with feature importance rankings dominated by engagement-related variables such as total likes, follower ratios, and likes per video.
In the gradient boosting model, feature importance was heavily concentrated on a single dominant predictor—the total number of likes. While this concentration contributed to strong predictive performance, it offered limited additional explanatory insight relative to the random forest, as fewer secondary variables meaningfully influenced the classification outcome.
XGBoost slightly improved predictive performance while maintaining consistent interpretation with the random forest.
Overall, the consistency of results across random forest, gradient boosting, and XGBoost indicates that the identified drivers of creator popularity are robust and not sensitive to the specific choice of tree-based modeling approach. This convergence strengthens confidence in the conclusion that engagement behavior is the dominant determinant of high-follower status, while textual self-presentation plays a secondary but complementary role.

5. Discussion

Across both models, biography text plays an important role in distinguishing creator types. Although TikTok bios are short, they still convey information about branding, identity, and platform strategy. Cross-platform terms consistently appear as influential features, supporting the idea that creators who integrate TikTok into a broader content ecosystem tend to gain more visibility. The comparison of logistic regression and random forest shows how linear and non-linear approaches capture different aspects of the data. Logistic regression provides clear interpretability, while the random forest detects interactions without manual specification. A key limitation is that bios are often fewer than 15–20 words and may include emojis or stylized text, which reduces the effectiveness of traditional NLP methods. Future work could explore character-level models, embeddings, or clustering to capture richer semantic patterns.

6. Conclusion

This project shows that creator biographies, although short, contain meaningful structure and predictive value at the profile level. Using text preprocessing, tokenization, vectorization, and supervised learning, we identified linguistic patterns and engagement-related features that help distinguish high-follower creators from others. When combined with aggregated behavioral metrics, bio text contributes additional predictive signal, particularly in non-linear models.

The proposed modeling framework is reproducible and modular, allowing new features or alternative text representations, such as embeddings, to be incorporated without altering the overall pipeline. A key limitation of this analysis is that video-level content is unobserved; therefore, the results should be interpreted as identifying profile-level correlates of creator visibility rather than causal drivers of content success. Despite this limitation, the findings have practical relevance for creator analytics and platform-level profiling, especially in large-scale settings where content inspection may be costly or unavailable.

7. Appendix: Tables and Figures

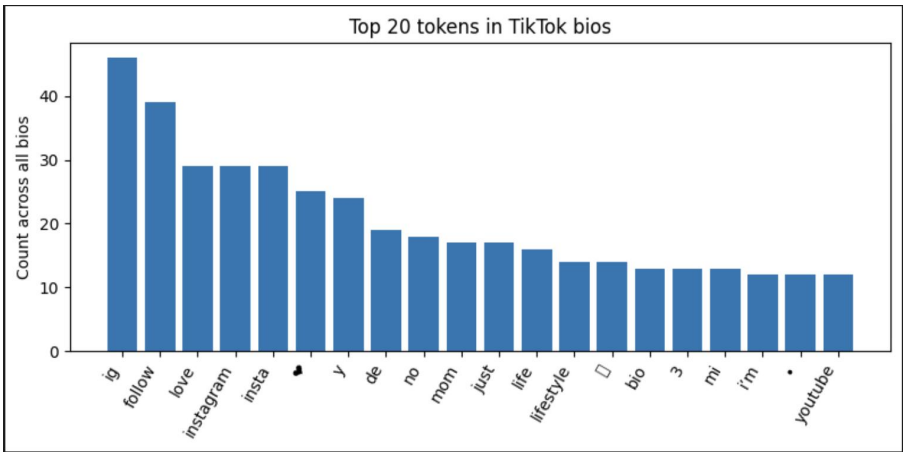

Figure 1: Top 20 tokens in TikTok bios

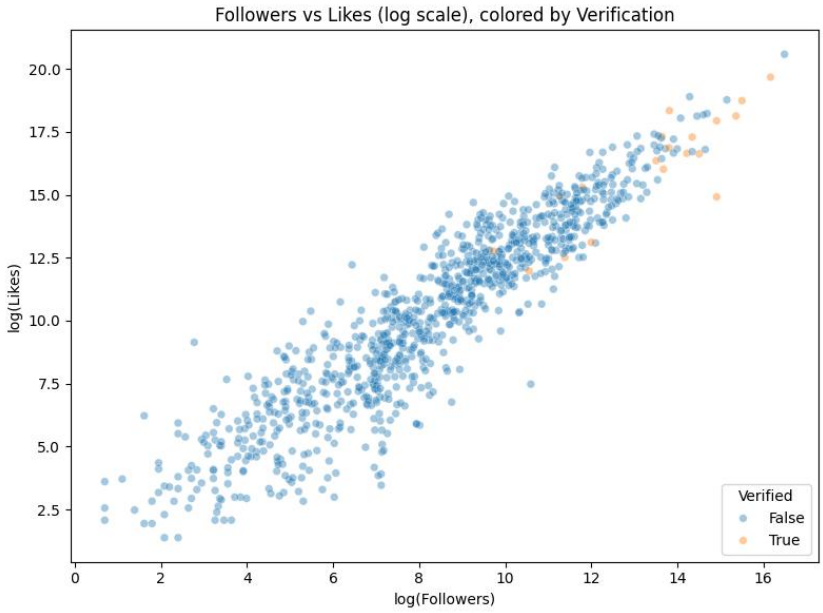

Figure 2: Relationship between followers and likes

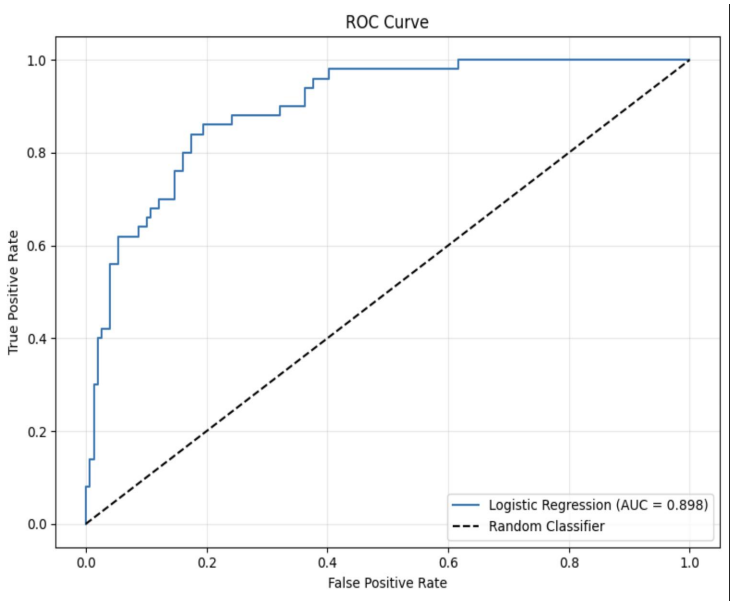

Figure 3: ROC-AUC curve for logistic regresion

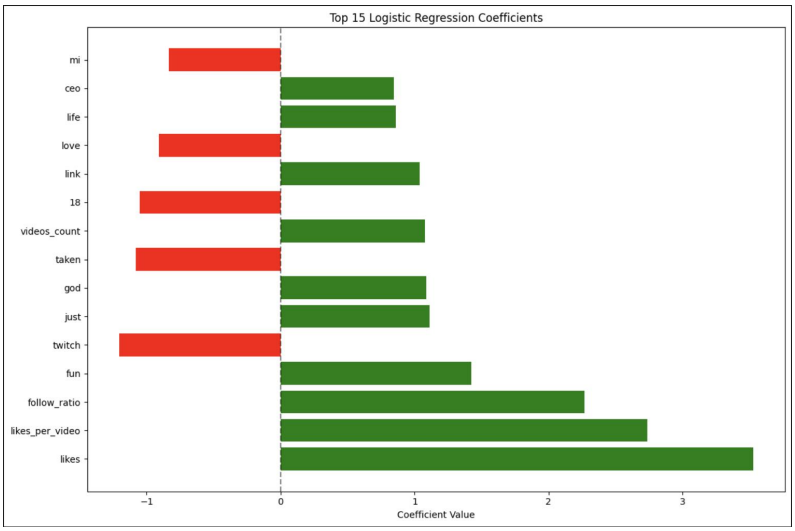

Figure 4: Top 25 Logistic Regression coefficients

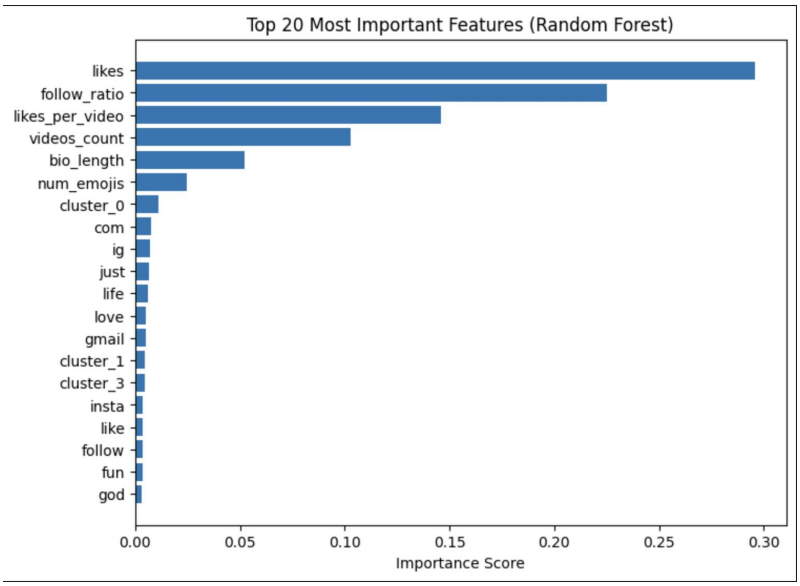

Figure 5: Top 20 most important features using Random Forest