Predicting the interactions of Weibo

YIXI ZHOU*
ShanghaiTech University
Shanghai, China
zhouyx2022@shanghaitech.edu.cn

QIANYI GONG
ShanghaiTech University
Shanghai, China
gongqy2022@shanghaitech.edu.cn

WENTAO YANG
ShanghaiTech University
Shanghai, China
yangwt2022@shanghaitech.edu.cn

***Abstract*—This research presents a comprehensive analysis and predictive modeling framework for social media engagement on Sina Weibo, China's largest microblogging platform. The study focuses on developing an accurate prediction model for post engagement metrics, including forwards, comments, and likes, within 24 hours of publication. We propose a systematic approach to feature engineering that encompasses three key dimensions: content characteristics, temporal patterns, and user behavior profiles.**

The research introduces several novel methodologies, particularly in user behavior analysis, where we developed a Zombie user identification system and comprehensive user profiling metrics. Our analysis demonstrates that user behavioral patterns are the most significant predictors of engagement success, followed by temporal and content features. The study evaluated multiple machine learning approaches, including Random Forest, XGBoost, Neural Networks with Random Forest emerging as the most effective model with a system named FRESH (Feature Random Forest Explained by SHAP Hyperboost) we create for choosing the best features, achieving a 31.51% accuracy rate on our weighted evaluation metric the same implication in Tianchi platform, ranking seventh among all participants in the engagement prediction challenge.

Through progressive model enhancement, we achieved a 57% improvement over the baseline model we set, with the most substantial gains coming from user profile integration. The findings highlight the critical role of user-centric features in engagement prediction and provide practical insights for content distribution optimization on social media platforms. This research contributes to both the theoretical understanding of social media engagement patterns and practical applications in content distribution systems.

***Index Terms*—Social Media Engagement, User Behavior Analysis, Machine Learning, Temporal Pattern Analysis, User Profiling, Feature Engineering**

I. Introduction

A. Background

In the contemporary digital landscape, social media platforms have become integral channels for information dissemination and social interaction. Sina Weibo, established in 2009, stands as China's predominant microblogging platform, serving over 582 million monthly active users as of 2023. Unlike traditional media channels, Weibo provides a dynamic environment where content value is directly reflected through user engagement metrics such as forwards, comments, and likes.

This is the final project for 24Fa CS182 Introduction to machine learning

B. Research Objectives

The primary goal of this research is to develop an accurate and efficient model for predicting post engagement levels on Sina Weibo within 24 hours of publication. Our objectives include:

1) Developing a comprehensive framework for analyzing and predicting user engagement patterns
2) Identifying key factors that influence content engagement success
3) Creating a practical prediction model that can be integrated into content distribution systems

C. Problem Statement

The challenge of engagement prediction in social media environments is multifaceted. First, engagement patterns are influenced by numerous factors, including content characteristics, temporal dynamics, and user behavior patterns. Second, the relationship between these factors and engagement outcomes is often non-linear and context-dependent.

Our dataset encompasses Weibo posts published between February and July 2015, including:

- Complete post content and metadata
- User interaction histories
- Temporal posting patterns
- Engagement metrics (forwards, comments, likes)
- Encrypted user and post identifiers for privacy protection

II. Feature Engineering and Analysis

We present our comprehensive approach to feature engineering and analysis for social media engagement prediction. Our framework encompasses four key dimensions: data preprocessing, content analysis, temporal patterns, and user behavior, each contributing unique and valuable insights to the prediction model.

A. Data Preprocessing Framework

The foundation of our analysis lies in systematic data preprocessing, which transforms raw social media data into structured, analyzable features. Our preprocessing pipeline handles both structured metadata and unstructured content, ensuring consistent data quality and format across all features.

For temporal information, we implemented a standardized processing system that decomposes timestamps into granular components (year, month, day, hour, minute, second). This

decomposition enables detailed temporal pattern analysis and facilitates the identification of time-based engagement trends.

Text content undergoes thorough preprocessing to remove noise while preserving meaningful information, which includes HTML tag removal, URL standardization, special character processing, handling forward and mention tags (like @) in Weibo, and text normalization.

The preprocessed data structure serves as the foundation for subsequent feature extraction and analysis stages, ensuring consistency and reliability in our modeling approach.

B. Content Feature Analysis

Content analysis forms a critical component of our feature engineering framework, focusing on identifying and quantifying characteristics that influence engagement levels. Our content feature extraction process examines multiple dimensions of post content:

Text Structure Features:

- Word count
  This chart (Figure 1) illustrates the relationship between the proportion of posts with zero likes and the word count in Weibo. The blue bars represent the proportion of posts that received no likes, and the data indicates a decreasing trend in the proportion of zero likes as the word count increases. This phenomenon suggests that longer posts may be more engaging and capable of attracting user attention, resulting in higher interaction and recognition.

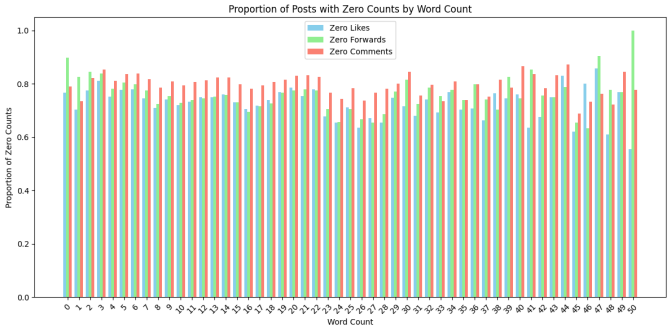

Fig. 1. Proportion of Posts with Zero Counts by Word Count. As the word count increases, the proportion of posts with zero likes decreases, indicating that longer posts tend to attract more engagement.

- Language type identification
  The analysis in (Figure 2) shows that posts with English content tend to have lower average counts for forwarding, commenting, and liking compared to those without. This suggests that the presence of English may positively influence engagement metrics.
- Title presence
  The presence of a compelling title is a crucial factor in driving user engagement on social media platforms like Weibo. Titles that resemble those found in news media often attract more attention and can lead to higher forwarding rates. These headlines are typically crafted to be concise, provocative, and informative, effectively capturing the essence of the content while sparking curiosity among readers.

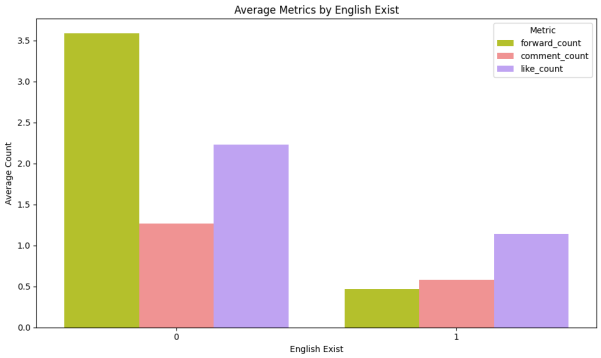

Fig. 2. Average forwarding, commenting, and liking counts based on the existence of English content.

Special Content Markers:

- Hashtag usage
  Hashtag usage is another pivotal element that can significantly enhance user engagement on social media platforms like Weibo. By categorizing content and making it easily searchable, hashtags help users discover relevant posts and topics of interest. Effective use of hashtags can increase the visibility of a post, attracting a broader audience and encouraging interactions. We draw the wordcrowd in Figure 3.

Fig. 3. Wordcrowd for the hashtag usage

- URL presence
  When users see a URL, it can pique their curiosity and motivate them to click through, leading to increased traffic for the linked content. Furthermore, URLs can enhance the perceived credibility of a post by directing users to authoritative sources or relevant information.
- User mentions
  User mentions play a significant role in fostering engagement on social media platforms like Weibo. When users mention others in their posts, it not only draws the attention of the mentioned individuals but also encourages them to interact with the content, potentially leading to higher rates of likes, shares, and comments.

- Media indicators (images, videos)
  Images can provide immediate context and make the content more relatable, while videos often convey information in an engaging and dynamic manner, encouraging viewers to share and comment.

Content Type Indicators:

- Promotional content
- Financial discussions
- Interactive elements
- Stock related
- Voting related

A set of Weibo posts consist of different areas and we separate them into different catogories according the keywords extracted from them in Figure 4.

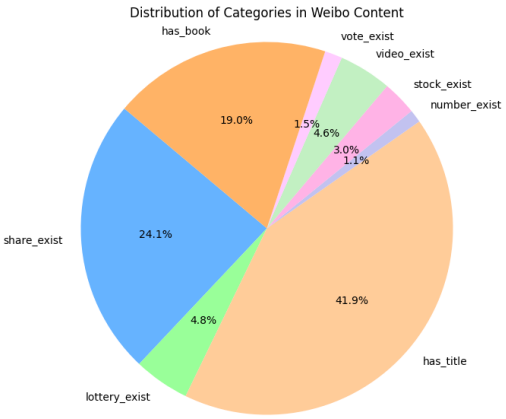

Fig. 4. Distribution of Categories in Weibo Content

This comprehensive content analysis enables our model to capture both explicit and implicit content characteristics that may influence engagement patterns.

C. Temporal Pattern Analysis

Temporal analysis in our framework focuses on identifying patterns and relationships between posting time and engagement levels. Our approach examines temporal patterns at multiple granularities:

Daily Patterns: The analysis focuses on user activity and content posting on Weibo throughout the day, as illustrated in the provided chart in Figure 5 and Figure 6.
The total number of users exhibits a clear diurnal pattern.

- Peak Usage: A noticeable peak occurs in the early hours of the day, particularly around 8 AM, indicating high user activity as individuals start their day.
- Low Activity Period: User engagement drops significantly during late night to early morning hours (1 AM to 6 AM), reflecting lower overall engagement.
- Moderate Activity: Throughout the day, user counts fluctuate but remain relatively stable post-morning peak.

The total posts demonstrate a more consistent pattern compared to the total users.

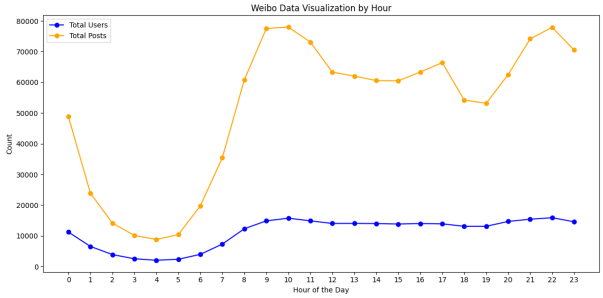

Fig. 5. Weibo Data Visualization by Hour

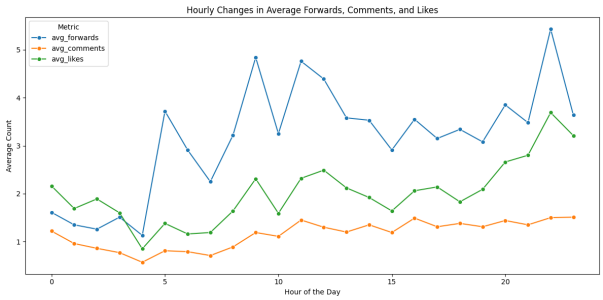

Fig. 6. The chart indicates that average forwards and comments exhibit significant fluctuations throughout the day, peaking in the early morning and evening, while average likes show a steady increase, reflecting varying levels of user engagement.

- Lower Activity: The number of posts is low during night and early morning, consistent with user activity levels.
- Increase in Posts: A rise in posts aligns with user peaks around 8 AM, with a secondary peak around 12 PM, suggesting that increased user presence correlates with higher posting activity, which is aligned with a report from Weibo data analysis [5].
- Evening Engagement: There is another peak around 7 PM, indicating active user engagement post-work or school hours.

Monthly Patterns:

- Weekly trends The average number of posts during weekdays is generally higher than on weekends. Weekdays show averages ranging from approximately 6,700 to 7,800 posts, while weekend averages are significantly lower. This finding implies that content strategies should focus on weekdays to maximize engagement in Figure 7.

- Monthly variations
  We draw the average data from to July overtime in Figure 8. The plot displays three lines representing the average forwards, comments, and likes over time on a social media platform. Each line has a different color: blue for forwards, green for comments, and orange for likes. If there are noticeable spikes in any of the metrics, it could correlate with specific events or campaigns that were run on those dates. Analyzing the content shared on those dates could provide insights into what drives

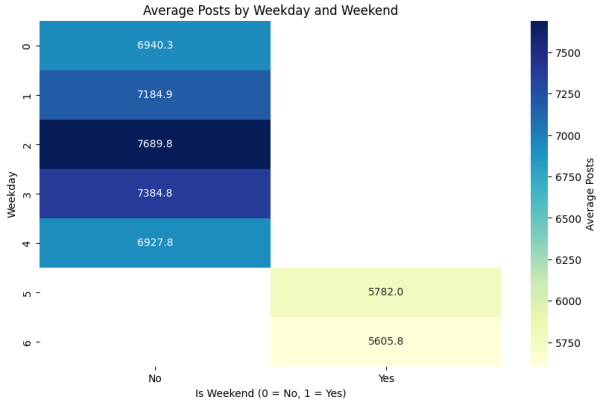

Fig. 7. Total Posts by Weekday and Weekend

higher engagement.

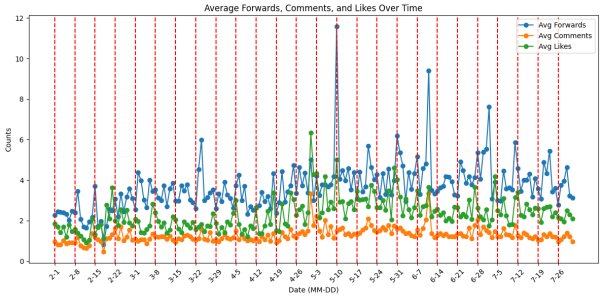

Fig. 8. Average Forwards, Comments, and Likes Over Time

- Holiday effects
  We also notify if holidays show higher averages in forwards, comments, and likes, this could indicate that users are more active during these times, perhaps due to leisure time or special promotions.

This temporal analysis provides crucial insights into optimal posting times and helps identify periods of maximum engagement potential.

D. User Behavior and Profile Analysis

User behavior emerges as a crucial predictor of engagement success in our analysis. Our user profiling framework encompasses:

Historical Performance Metrics:

- Average engagement rates
  The K-Means clustering analysis [4] provides a meaningful segmentation of less active users based on their engagement metrics. By examining the clusters and their centroids, insights can be drawn about user behavior, which can inform content strategies and user engagement initiatives. We use the K-Means algorithm resulting in 10 distinct clusters of users based on their average comments, forwards, and likes in Figure 9. For instance, some clusters may represent users who primarily comment but do not forward or like content, while others may show a

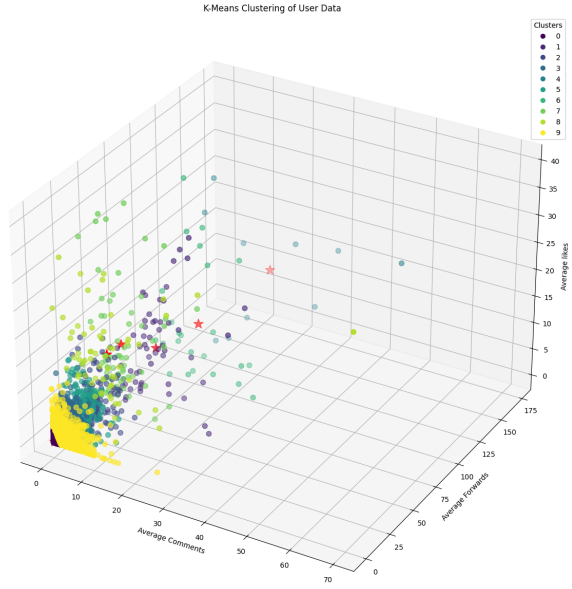

Fig. 9. K-Means Clustering of User Data

mix of behaviors.

PCA is employed to reduce the dimensionality of the original dataset while retaining as much variance as possible. This allows for easier visualization and interpretation of complex data. The plot in Figure 10 shows distinct groupings of points, each corresponding to different clusters. The colors represent the clusters identified in the previous K-Means analysis.

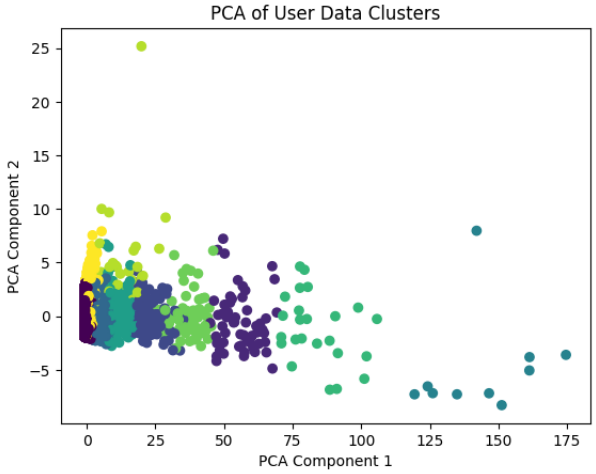

Fig. 10. PCA of User Data Clusters, Some clusters are well-separated, indicating that they represent distinct user segments based on their engagement metrics. Others may overlap, suggesting similarities in user behavior across those clusters.

- Engagement variance and Peak performance indicators
  The purpose of this radar chart is in Figure 11 to provide a comprehensive overview of user engagement metrics in a single visual representation. By plotting these

metrics, researchers can easily compare different users' engagement levels and identify patterns or anomalies.

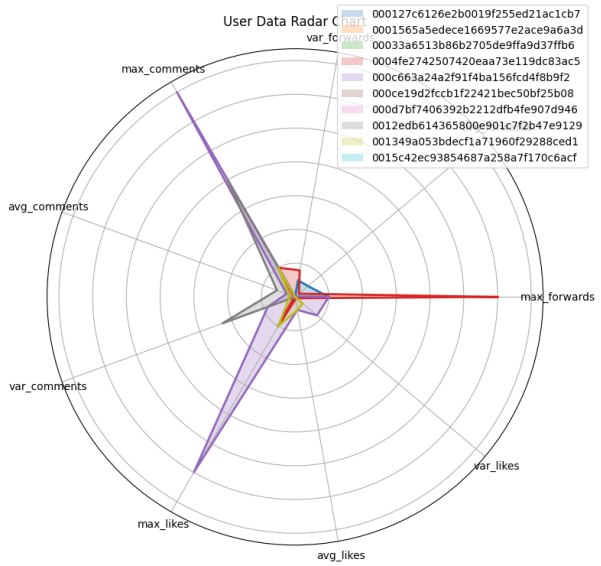

Fig. 11. User Data Radar Chart

In the analysis of Weibo data, four categories of users are defined based on their engagement metrics in Table I: the average number of comments, forwards, and likes. The zombie user identification system represents a particular innovation in our approach, automatically identifying and handling consistently low-engagement accounts to improve overall prediction accuracy.

TABLE I
USER ENGAGEMENT CATEGORIES ON WEIBO

User Category	Description
Zombie Users	These users are defined as individuals who have zero activity across all three metrics, meaning their average comments, average forwards, and average likes are all equal to zero.
Most Zombie Users	This category includes users who exhibit inactivity in two of the three metrics, indicating that two of their average metrics are equal to zero while one is greater than zero.
Likely Zombie Users	Users in this category have one of the three engagement metrics equal to zero, suggesting minimal interaction but some activity in the other two metrics.
Active Users	These individuals have non-zero values in all three engagement categories, demonstrating consistent interaction with the platform.

The Venn diagram in Figure 12 illustrates the relationship between three variables: average comments, average forwards, and average likes. Each variable can either be zero or non-zero, and the diagram visually represents the different combinations of these states.

This comprehensive feature engineering framework provides the foundation for our subsequent modeling efforts, enabling more accurate and reliable engagement predictions. The combination of content, temporal, and user behavior features creates a rich feature space that captures the complex dynamics of social media engagement.

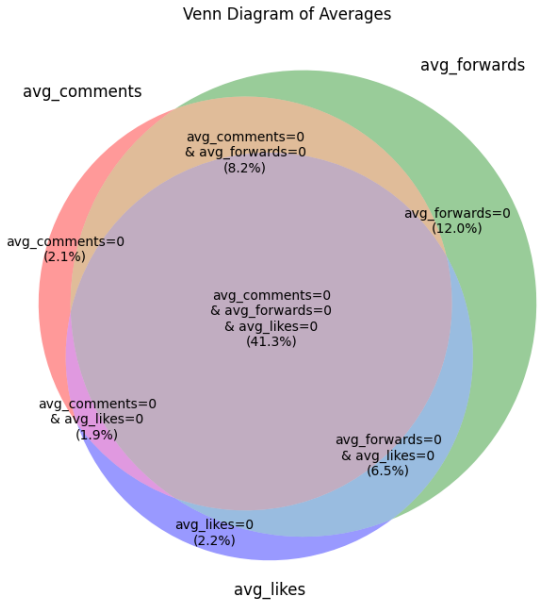

Fig. 12. Venn Diagram of Averages: The categories represent different combinations of user engagement metrics, indicating how many instances have zero average comments, forwards, or likes, highlighting various patterns of interaction with the content.

III. BASELINE MODEL IMPLEMENTATION

In establishing the foundational framework for predicting user engagement on Weibo posts, we implemented a baseline model that integrates temporal and content-based features. This initial approach serves as a benchmark for evaluating more sophisticated modeling techniques and provides insights into the fundamental patterns of user engagement behavior.

A. Feature Engineering

The baseline model incorporates two primary categories of features:

1) Temporal Features: We extracted chronological components from post timestamps, including year, month, day, hour, minute, and second. These temporal features are crucial for capturing potential patterns in posting times that might influence user engagement levels. The granularity of temporal features allows the model to identify both broad seasonal trends and specific time-of-day effects on user interaction patterns.

2) Content-Based Features: For processing textual content, we implemented Term Frequency-Inverse Document Frequency (TF-IDF) vectorization using a parallel processing approach. The parallel implementation was necessary to efficiently handle the large volume of text data while maintaining computational feasibility. This approach transforms the unstructured text content into a structured numerical representation that captures the relative importance of terms within each post.

B. Model Architecture

The prediction framework utilizes a Random Forest Regressor as the core modeling component.

The model was trained on an 80-20 split of the dataset, with 80% used for training and 20% reserved for testing and validation.

C. Performance Evaluation

We developed a custom evaluation metric that accounts for the relative importance of different engagement types. The scoring function S for a post p is defined as:

$$S_p = 1 - (0.5 \cdot \delta_f + 0.25 \cdot \delta_c + 0.25 \cdot \delta_l)$$

where:

- $\delta_f = \frac{|f_p - f_a|}{f_a + 5}$ represents the normalized forward count error
- $\delta_c = \frac{|c_p - c_a|}{c_a + 3}$ represents the normalized comment count error
- $\delta_l = \frac{|l_p - l_a|}{l_a + 3}$ represents the normalized like count error

The final model performance is calculated as:

$$Performance = \frac{\sum_{i=1}^{n}(min(e_i + 1, 101) \cdot I(S_i > 0.8))}{\sum_{i=1}^{n}(min(e_i + 1, 101))}$$

where e_i is the total engagement count for post i, and I is an indicator function that equals 1 when the prediction accuracy exceeds 80%.

And the final score of the baseline is about 20.08%

IV. Enhanced Model with Zombie User Detection

Building upon our baseline implementation, we introduced a significant enhancement by incorporating zombie user detection, which led to a substantial improvement in model performance, increasing the prediction accuracy from 20.08% to 25.81%.

A. Zombie User Identification

In the context of social media engagement, we define zombie users as accounts that consistently show zero engagement across all interaction metrics (forwards, comments, and likes). This pattern suggests either inactive accounts or automated accounts that do not generate meaningful social interactions. We implemented the following identification process:

1) We identified users whose posts consistently received zero engagement across all three metrics (forwards, comments, and likes) in the training dataset
2) We maintained a distinct set of user identifiers (UIDs) for accounts that exhibited this zero-engagement pattern
3) These identified zombie users were segregated from the main training process to prevent them from introducing noise into the model's learning patterns

B. Modified Training Strategy

The enhanced model employs a two-pronged approach:

- For regular users, we maintain the original Random Forest Regressor with temporal and content-based features
- For identified zombie users, we automatically predict zero engagement across all metrics, bypassing the main prediction model

This stratification of users allows for more focused training on accounts that demonstrate actual engagement patterns, while efficiently handling predictions for consistently non-engaging accounts.

C. Performance Improvement

The introduction of zombie user detection led to a significant improvement in model performance:

- Baseline Model Performance: 20.08%
- Enhanced Model Performance: 25.81%
- Relative Improvement: 28.5%

This substantial improvement suggests that explicitly accounting for user engagement patterns through zombie user detection is an effective strategy for refining engagement predictions. The enhancement particularly benefits the model's ability to accurately predict cases of zero engagement, which form a significant portion of the dataset.

This improvement validates our hypothesis that distinguishing between engaging and non-engaging users is crucial for accurate engagement prediction in social media platforms. The results suggest that future improvements might be achieved through more sophisticated user categorization schemes.

V. Content-Based Feature Enhancement

Following the implementation of zombie user detection, we further refined our model by incorporating sophisticated content preprocessing and feature extraction techniques. This iteration focused on extracting structured information from the unstructured text content of Weibo posts.

A. Content-Based Model

Our initial approach focused on analyzing post content through Natural Language Processing (NLP) techniques. We implemented several content-specific features:

- Basic content indicators (topics, user mentions, URLs)
- TF-IDF vectorization of cleaned content (100 features)
- Temporal features (timestamp components)

This baseline content-focused model achieved a performance of 25.75%, highlighting the limitations of relying solely on content analysis.

B. User Behavior Integration

Following the content-based approach, we incorporated user behavioral patterns:

- Historical engagement metrics (average forwards, comments, likes)
- Engagement variability (maximum values, variance)
- Temporal activity patterns

This enhancement improved model performance to 30.61%, demonstrating the importance of user behavioral features.

C. Feature Engineering Optimization

We conducted systematic feature engineering in three key areas:

1) Enhanced Content Features: This analysis focuses on various aspects of language and user engagement. It includes linguistic features such as word count, language detection, numerical patterns, and punctuation use. Structural indicators assess content types, interaction patterns, and topic density. Temporal pattern analysis examines hour-of-day effects, seasonal trends, and time-based engagement. Lastly, user engagement metrics cover historical statistics, interaction patterns, and regularities in user activity.

In the following part, we will discuss how we choose the best feature using a system called **FRESH**.

D. Linear Regression Results

We also experimented with linear regression as a baseline approach, which achieved an accuracy of 22.71%. This relatively poor performance was expected given the inherently non-linear relationships in user engagement data. However although the model is simple enough, with more careful feature extraction it performs better than the baseline with just few unhandled inputs.

E. Model Optimization for Randomforest

The number of estimators, which refers to the number of trees in the forest, is a critical hyperparameter that influences the model's accuracy and performance. This analysis examines how different values for the number of estimators impact the model's score, providing insights into optimal parameter selection.

A series of experiments were conducted with the number of estimators ranging from 30 to 60. The performance of the model was evaluated using a scoring metric, which quantifies its predictive accuracy. The following scores were recorded in the Figure 13. The results indicate a peak performance at

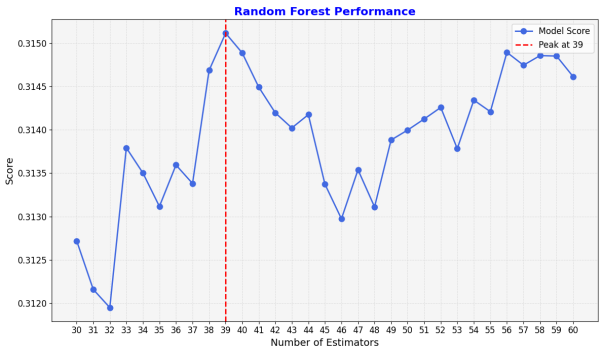

Fig. 13. Random Forest Performance

nmuber of estimators = 39, where the score reached 0.3151. This is followed closely by scores for 38 and 40 estimators, suggesting that increasing the number of estimators beyond this point may not yield significant improvements in accuracy.

TABLE II
MODEL PERFORMANCE EVOLUTION

Estimators	Accuracy	Features Added
25	31.15%	Basic feature set
30	31.19%	Enhanced temporal features
35	31.21%	Additional content indicators
40	31.49%	Complete feature integration
39	**31.51**%	Complete feature integration

F. Final Results

Our final optimized model achieved 31.51% accuracy, representing a significant improvement over the baseline:

- Initial content-only model: 25.75%
- User behavior integration: 30.61%
- Final optimized model: 31.51%
- Total improvement: +5.74 percentage points

This progression demonstrates that while content analysis provides valuable insights, the combination of user behavior patterns, temporal features, and optimized model parameters yields the most effective prediction system. The final model's performance suggests that engagement prediction in social media requires a holistic approach that considers multiple aspects of user behavior and content characteristics.

VI. XGBOOST IMPLEMENTATION

A. Model Selection and GPU Acceleration

Following our Random Forest implementation, we explored XGBoost (eXtreme Gradient Boosting) [3] as an alternative approach.

B. Performance Analysis

The XGBoost implementation achieved a performance score of 30.47%, which, while competitive, did not surpass our best Random Forest model (31.49%). However, the significant reduction in training time presents an important practical advantage:

TABLE III
MODEL COMPARISON SUMMARY

Model	Accuracy	Key Advantage
Random Forest	31.49%	Higher accuracy
XGBoost (GPU)	30.47%	Faster training

This exploration demonstrates that while our Random Forest implementation remains the most accurate model, the XGBoost implementation offers a compelling alternative when considering the balance between prediction accuracy and computational efficiency.

C. Analysis of Feature Impact with SHAP

1) Introduction to SHAP Methodology: The SHAP (SHapley Additive exPlanations) method [2] is a powerful tool for interpreting machine learning models, providing insights into how different features contribute to predictions made by complex algorithms like XGBoost. Based on cooperative game theory, SHAP values assign each feature a value reflecting its

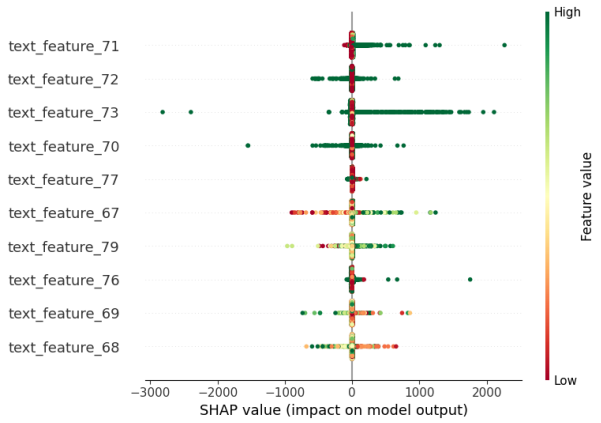

Fig. 14. Distribution of SHAP values for specific features.

contribution to the prediction for an individual instance. Figure 14 illustrates the distribution of SHAP values for specific features, highlighting their impact on the model's output. The x-axis represents SHAP values, indicating whether each feature contributes positively or negatively to the predicted outcome. The y-axis lists features , with each point showing how a particular instance's feature value influences the model's prediction. The color gradient (from blue to red) reflects feature values, where blue indicates low values and red indicates high values.

2) Explanation of the Figures and Results: In this study, we utilized the XGBoost model to predict the number of forwards, comments, and likes on Weibo posts. To interpret the model's predictions, we employed SHAP values, quantifying the contribution of each feature to the predicted outcomes.

Figure 15 reveals significant differences in contributions among various features. For example:

- In predicting forwards, features such as average forward count exhibit higher SHAP values, indicating their substantial impact.
- For comment prediction, features like all punctuation existing feature show notable influence.
- In predicting likes, average forward count again stands out, highlighting its significance across all metrics.

3) Insights and Benefits for Feature Selection: The analysis of SHAP values provides several insights and benefits for feature selection in subsequent modeling efforts, particularly for both XGBoost and Random Forest models.

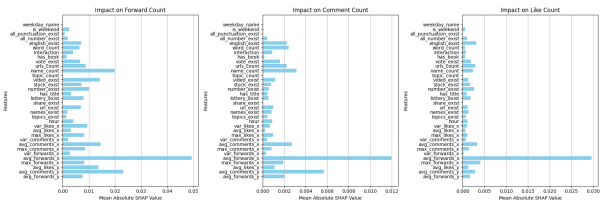

Fig. 15. Impact of various features on model predictions.

VII. FRESH: A Novel Framework for Feature Selection and Model Interpretation Using XGBoost, Random Forest, and SHAP

After testing the Random Forest and XGBoost, we attemp to find a solution for both of their drawbacks. Random Forest is widely recognized for its robustness and superior performance on structured datasets, yet it often suffers from longer training times and limited interpretability regarding feature contributions. Conversely, XGBoost offers faster training and improved accuracy, while SHAP provides a powerful method for interpreting model outputs. The FRESH (Feature Random Forest Explained by SHAP Hyperboost) framework aims to address these challenges by integrating these methodologies to streamline feature selection and enhance interpretability.

A. Framework Overview

The FRESH framework consists of the following key components:

1) Feature Analysis with XGBoost: The process begins by training an XGBoost model on the dataset, leveraging its speed and accuracy to generate predictions.

2) SHAP Value Analysis: Following the training, SHAP values are computed to quantify the contribution of each feature to the model's predictions. This step facilitates a comprehensive understanding of feature importance and interactions.

3) Feature Selection for Random Forest: Based on SHAP analysis, the most significant features are identified and selected for further modeling. This targeted approach reduces complexity and enhances the efficiency of the subsequent model.

4) Random Forest Model Training: Finally, a Random Forest model is trained using the selected features, capitalizing on its robustness and ability to manage non-linear relationships in the data.

B. Expected Outcomes

The FRESH framework is anticipated to yield several key benefits:

- **Enhanced Interpretability**: The integration of SHAP values allows for intuitive visualizations of feature contributions, aiding stakeholders in understanding model decisions.
- **Improved Performance**: By utilizing only the most relevant features, the Random Forest model can achieve higher accuracy and reduced training times.
- **Synergistic Strengths**: FRESH effectively combines the advantages of XGBoost for feature analysis and SHAP for interpretability, resulting in a powerful and efficient modeling approach.

C. Improvement in Model Performance

After implementing the FRESH framework for feature selection and model interpretation, we observed a notable improvement in the model's performance. Specifically, the carefully selected features led to an increase in the final score

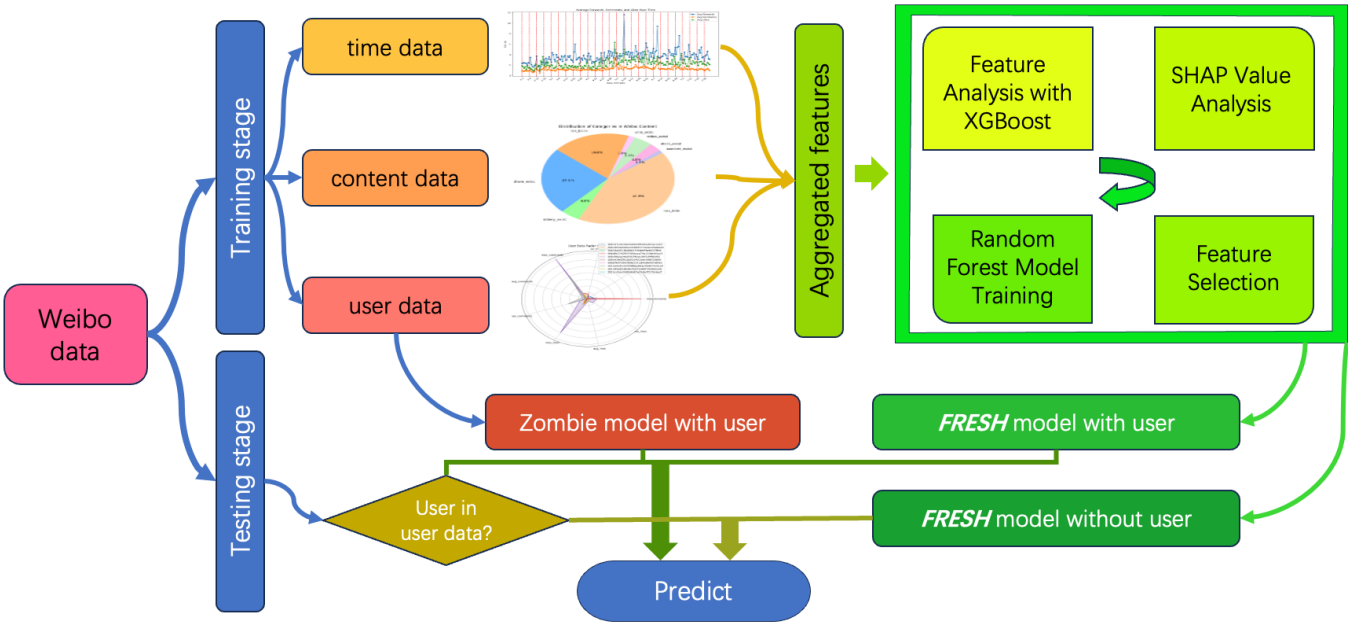

Fig. 16. pipeline of FRESH

from 31.21% to 31.51% ranking seventh among all participants in the engagement prediction challenge. This enhancement demonstrates the effectiveness of our approach in optimizing the feature set, thereby improving the overall predictive accuracy of the model.

VIII. Challenges with Neural Networks on Tabular Data

After implementing our random forest baseline which achieved strong performance, we explored deep neural networks with the hypothesis that they could discover latent patterns in the data that traditional models might miss. However, despite extensive hyperparameter tuning of hidden layers and careful epoch selection with validation data, our neural network implementation achieved only 19.00% accuracy, significantly underperforming the random forest baseline.

This underperformance aligns with the findings from Borisov et al. [1], who identify several fundamental challenges when applying deep neural networks to tabular data:

1) Missing Spatial Dependencies:
   Unlike image or text data where nearby features have strong correlations, our Weibo engagement features lack inherent spatial or sequential relationships. As noted in the survey, there is often no spatial correlation between the variables in tabular data sets , making it difficult for neural networks to learn meaningful feature hierarchies.
2) Feature Importance Sensitivity:
   Individual features in our dataset, such as user history metrics and temporal patterns, can have outsized importance on the prediction. Neural networks struggle with this characteristic of tabular data, where the smallest possible change of a categorical (or binary) feature can entirely flip a prediction .
3) Mixed Feature Types:
   Our data contains a heterogeneous mix of categorical variables (e.g., user IDs), numerical engagement metrics, and temporal features. This aligns with the survey's observation that tabular data's mixed structure of discrete and continuous features along with their different value distributions still poses a significant challenge for neural networks.

This experience validates the survey's conclusion that despite significant research efforts, deep learning approaches still generally underperform traditional tree-based methods on typical tabular datasets. The structural assumptions inherent in neural network architectures appear fundamentally mismatched with the statistical properties of tabular data like our Weibo engagement prediction task.

IX. Conclusion and Future Work

A. Summary of Contributions

This research presents several significant contributions to the field of social media engagement prediction:

1) **FRESH Framework Development**: We introduced the novel FRESH (Feature Random Forest Explained by SHAP Hyperboost) framework, which successfully combines the strengths of Random Forest and XGBoost with SHAP interpretation, achieving a 31.51% accuracy rate in engagement prediction.
2) **Zombie User Detection System**: Our implementation of a zombie user detection system significantly improved model performance from 20.08% to 25.81%,

demonstrating the importance of user categorization in engagement prediction.

3) **Comprehensive Feature Engineering**: We developed a sophisticated feature engineering pipeline that incorporates content analysis, temporal patterns, and user behavior profiles, providing a holistic approach to engagement prediction.
4) **Model Architecture Comparison**: Through extensive experimentation with different model architectures, including Random Forest, XGBoost, and Neural Networks, we identified the most effective approaches for social media engagement prediction on tabular data.

B. Key Findings

Our research revealed several important insights about social media engagement prediction:

- User behavior patterns emerged as the strongest predictors of engagement success, outperforming both content and temporal features in importance.
- The identification and separate handling of zombie users significantly improved model accuracy, suggesting the importance of user segmentation in engagement prediction.
- Traditional machine learning approaches (Random Forest and XGBoost) outperformed deep learning methods for this specific tabular data prediction task, aligning with findings from recent literature.
- The integration of SHAP analysis provided valuable insights into feature importance, enabling more effective feature selection and model interpretation.

C. Future Research Directions

Based on our findings, we identify several promising areas for future research:

1) Enhanced User Profiling: Development of more sophisticated user categorization systems that go beyond the current zombie user detection, potentially incorporating behavioral clustering and temporal activity patterns.
2) Real-time Prediction: Extension of the current framework to support real-time engagement prediction, incorporating streaming data and dynamic feature updating.
3) Temporal Dynamics: Further exploration of temporal effects on engagement patterns, including the development of specialized models for different time periods and seasonal effects.
4) Content Analysis Enhancement: Integration of more advanced natural language processing techniques, including sentiment analysis and topic modeling, to better capture content-based engagement factors. Actually we try to use the bert tokenizer for the model to better understand the deeper meaning of the Weibo post, but considering the cost we choose not to take it as a main method to handle with the content.

In conclusion, this research advances our understanding of social media engagement prediction through the development of novel methodologies and comprehensive analysis of engagement factors. The success of our approach, particularly the FRESH framework and zombie user detection system, provides a strong foundation for future research in this rapidly evolving field.

X. Ethics Statement

This research was conducted with careful consideration of ethical principles and data privacy concerns. All data used in this study was collected from publicly available sources on the Sina Weibo platform, in accordance with the platform's terms of service and applicable regulations. To ensure user privacy and data security, we implemented the following measures:

1) Data Anonymization: All user identifiers (UIDs) were encrypted using secure hashing algorithms to prevent any possibility of reverse identification.
2) Content Protection: Post identifiers and content were similarly encrypted to protect user privacy while maintaining the integrity of the analysis.
3) Aggregate Analysis: Our research focused on aggregate patterns and trends rather than individual user behavior, further protecting user privacy.
4) Secure Data Handling: All data processing and analysis were conducted in a secure computing environment with appropriate access controls.

We acknowledge our responsibility to protect user privacy while advancing social media research, and we believe our methodology strikes an appropriate balance between these competing concerns. Our approach to data handling and privacy protection can serve as a model for future research in social media analytics.

References

[1] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, Deep Neural Networks and Tabular Data: A Survey, *IEEE Trans. Neural Networks Learn. Syst.*, vol. 35, no. 6, pp. 7499–7519, 2024.

[2] Meng, Y., Yang, N., Qian, Z., and Zhang, G. (2021). What Makes an Online Review More Helpful: An Interpretation Framework Using XGBoost and SHAP Values. Journal of Theoretical and Applied Electronic Commerce Research, 16(3), 466-490.

[3] Chen, Tianqi, and Carlos Guestrin. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

[4] Ahmed, M., Seraj, R. and Islam, S.M.S., 2020. The k-means algorithm: A comprehensive survey and performance evaluation. Electronics, 9(8), p.1295.

[5] Sina Weibo Data Center, 2015 Weibo User Development Report, Sina Weibo Data Center, Beijing, China, Tech. Rep., Dec. 2015. [Online]. Available: https://data.weibo.com/report/reportDetail?id=297