Predicting Controversiality of Social Media Posts

Ellora Devulapally and **Riana Doctor**
{eldevulapally,ridoctor}@davidson.edu
Davidson College
Davidson, NC 28035
U.S.A.

1 Introduction

Social media platforms thrive on rapid interaction, yet not all posts generate the same level of response. Some attract little attention, while others spark intense discussion, disagreement, or sustained engagement. Understanding why certain posts draw disproportionate attention is important because high-engagement or controversial content can shape community norms, amplify conflict, and influence how users interact online.

Our goal was to make predictions on the popularity and engagement metrics of posts on Davidson College's YikYak API. At Davidson, YikYak is the predominant form of anonymous forum social media engagement. We planned to measure what specific metrics informed a post's virality, and explored how to measure this through a variety of different methods.

In this project, we initially set out to predict controversiality, defined as whether a post would provoke polarized reactions. However, early experiments revealed that controversiality was difficult to predict reliably using textual and lightweight engagement features alone. Across multiple models, including Logistic Regression, Support Vector Machines, and Random Forests, performance remained average in terms of prediction. Furthermore, it was difficult to determine whether a post was deemed controversial in YikYak, as there is no "controversial" flag.

These results motivated a reframing of our research question toward a more tractable and behaviorally grounded task: predicting high engagement. Engagement, defined through observable user interactions such as comments and upvotes, provides a clearer signal of community response. When applied to this task, the same modeling framework yielded substantially stronger results. So, we focused our analysis on engagement prediction.

2 Data Collection

Source Dataset: YikYak

Our primary dataset consists of posts scraped from Sidechat/YikYak using a reverse-engineered API wrapper implemented in Node.js.

We recreated the approach by Matthew Esposito to scrape the YikYak data (Esposito 2023). We collected YikYak posts using a reverse engineered API developed by Micah Lindley (Lindley 2023b; 2023a).

Posts were collected by repeatedly calling getGroupPosts() and retrieving data in batches up to a defined cutoff date, which we set as a calendar year from the current date. The data is stored in JSONL format, where each line represents a single post. Below are the metadata elements collected for each YikYak post:

- **id:** anonymized identifier
- **text:** full post content
- **created_at:** timestamp of submission
- **vote_total:** net up-votes
- **comment_count:** number of comments
- **group_id:** numeric group identifier
- **index_code:** internal platform code

Unlike Reddit, YikYak does not provide an explicit controversiality label. Although we considered manually a subset of the YikYak posts as controverisal/not controversial, we decided to train our models on Reddit data in order to determine what made a post controversial.

Source Dataset: Reddit

Our original approach leveraged Reddit, where posts are explicitly labeled as controversial by the platform, to train supervised models that could later be applied to YikYak. Reddit comments were also selected as the training dataset due to their structural and linguistic similarities to YikYak posts, as well as the ease of scraping the Reddit API.

Reddit data were collected using the platform's official public API (Reddit, Inc. 2024). We collected data from a wide range of subreddits in order to capture a broad range of engagement behaviors. For each subreddit, we retrieved posts from several standard Reddit listings, including `controversial`, `top`, `hot`, and `new`. Thus, we were able to list a binary label of marked controversial or not marked controversial for each post in the dataset without requiring manual annotation.

To ensure comparability with YikYak posts, we restricted the metadata collected for each Reddit post to match the YikYak JSON file. All posts were transformed into a unified

schema aligned with the YikYak dataset to support cross-platform modeling.

We ran into limitations with Reddit's rate limits and usage policies, encountering HTTP 429. Though we attempted to use the Reddit Data API, our request for access was denied. Thus, we had to collected data incrementally across multiple runs and different computers. We later merged the data collected into a single dataset and removed any duplicate posts prior to modeling.

3 Preprocessing and Feature Engineering

Reddit Data Preprocessing

Reddit data required the most extensive preprocessing due to being collected across multiple API calls and stored in separate JSON files. These files were first merged into a single JSON object.

After loading the merged Reddit dataset, we removed any posts missing essential fields such as text content or labels. Text preprocessing then standardized all content by converting text to lowercase, removing URLs, stripping user mentions, and normalizing whitespace. This step was important given the informal and noisy nature of Reddit comments.

As mentioned earlier, to ensure compatibility between the training and target domains, we transformed the Reddit dataset to match the structure of the YikYak dataset as closely as possible.

Specifically, we:

- Renamed and standardized feature fields to match YikYak conventions
- Retained only features that are observable or computable in both datasets
- Normalized engagement-related attributes (e.g., comment counts and upvotes)

This transformation reduced the risk of domain mismatch.

When we structured our models around controversiality prediction, we used Reddit's platform-provided controversial flag as the target label. When we transitioned to an engagement prediction model, we defined high engagement as posts whose comment-to-upvote ratio fell within the top quartile of the distribution.

For both focuses, we engineered several categories of features:

- **Engagement-derived features:** comment-to-upvote ratios, later used to define high-engagement labels.
- **Sentiment features:** polarity scores from TextBlob and sentiment components (positive, neutral, negative, compound) from VADER
- **Linguistic cues:** post length, first- and second-person pronoun counts and ratios, and disagreement markers based on curated lexicons.
- **Temporal features:** hour of posting, day of week, and weekend indicators.

Many of these engineered features were developed consistent with prior findings that controversy is strongly influenced by discussion structure, not only post content (Hessel and Lee 2019).

YikYak Data Preprocessing

Similar cleaning steps were applied to our YikYak data, including removal of empty or malformed posts and normalization of text. In addition, emojis were converted to text descriptors to preserve emotional content rather than discarding it. We engineered linguistic, sentiment, and temporal features to maintain alignment with the Reddit dataset and to bolster our findings further than textual data alone could.

Because YikYak does not provide an explicit controversiality flag comparable to Reddit's, we initially attempted to define controversiality ourselves using a comment-to-upvote ratio. However, we found that this distinction was too ephemeral, subjective and unstable to support a reliable binary classification task. As a result, we transitioned to a more data-driven formulation focused on predicting engagement. Engagement labels were derived from vote totals, with high engagement defined as posts in the top decile of upvotes.

Handling Class Imbalance (YikYak)

Our YikYak dataset exhibited severe class imbalance, with high-engagement posts (top decile of vote totals) representing only 10 percent of the data. To address this challenge, we employed multiple strategies: First, we performed undersampling of the majority class during training, reducing the training set to a balanced distribution while preserving the test set's original proportions. Second, we implemented class-weighted loss functions across all models, assigning higher penalty weights to misclassified minority-class instances. Third, we experimented with balanced variants of ensemble methods, such as BalancedRandomForestClassifier, which incorporates built-in balancing mechanisms. These approaches were essential to prevent models from defaulting to majority-class predictions while still learning meaningful signals for high-engagement detection.

Advanced Feature Engineering (YikYak)

Beyond basic text and temporal features, we engineered several feature categories specifically for YikYak's community dynamics. We made sure that we calculated each of these features using only past data in order to prevent data leakage when making predications on unposted text:

- Burstiness Features: We calculated the number of posts in the previous 2 hours both globally (postsprev2hall) and within each group (postsprev2hgroup), allowing models to capture temporal patterns in community activity.
- Relative Activity Metrics: The ratio relpostsprev2h measured whether a group was unusually active compared to the entire platform.
- Linguistic Engagement Signals: Beyond sentiment, we counted exclamation points, question marks, and conflict words (e.g., "but", "however", "actually") that might indicate provocative or discussion-worthy content.

Duplicate Removal and Leakage Prevention (YikYak)

During experimentation, we observed unrealistically high model performance, which led us to identify duplicate posts

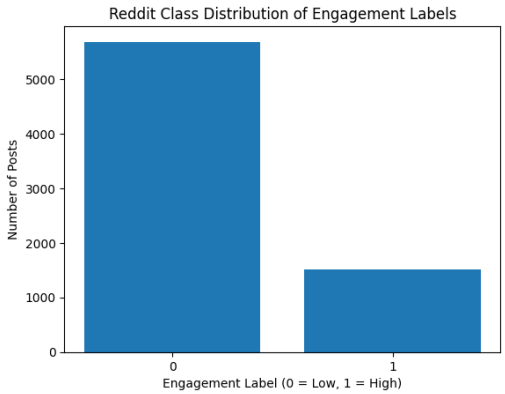

Figure 1: Reddit Class Distribution of Engagement Labels

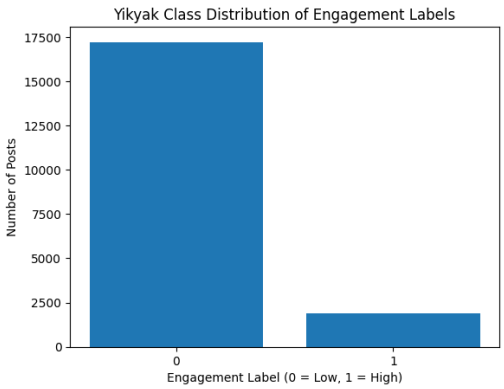

Figure 2: Yikyak Class Distribution of Engagement Labels

leaking across train and test splits. We removed duplicate entries based on cleaned text before splitting the data, resulting in more realistic and reliable evaluation metrics.

Exploratory Data Analysis

Figures 1 and 2 show the class distribution of engagement labels for the Reddit and YikYak datasets, respectively. In both cases, high-engagement posts constitute a clear minority of the data. This imbalance is expected by design, as high engagement is defined using a percentile-based threshold (top quartile for Reddit and top decile for YikYak). While this labeling strategy allows us to isolate posts that receive disproportionately high attention, it also introduces substantial class imbalance into the learning problem.

This imbalance has important implications for model training and evaluation. In particular, standard accuracy becomes a misleading metric, as a model can achieve high accuracy by consistently predicting the majority class while failing to identify high-engagement posts. As a result, we prioritize precision, recall, and F1-score for the high-engagement class when evaluating model performance. Additionally, the observed imbalance motivates our use of stratified train–test splits, class-weighted loss functions, and undersampling of the majority class to ensure that models are encouraged to learn minority-class signals rather than defaulting to majority predictions.

We found generally that YikYak engagement is temporally structured, highly skewed and weakly linear. Figure 4 shows generally weak linear correlations between most numeric features and engagement-related variables. While vote totals and comment counts are correlated with each other, no single scalar feature exhibits a strong linear relationship with high engagement.

The pair plot (Figure 3) further illustrates this pattern, showing substantial overlap between low- and high-engagement posts across numeric feature dimensions and highly skewed feature distributions. Together, these observations suggest that engagement is not driven by any single numeric attribute, motivating the use of high-dimensional textual representations and regularized models that can combine multiple weak signals. This generally corresponds to

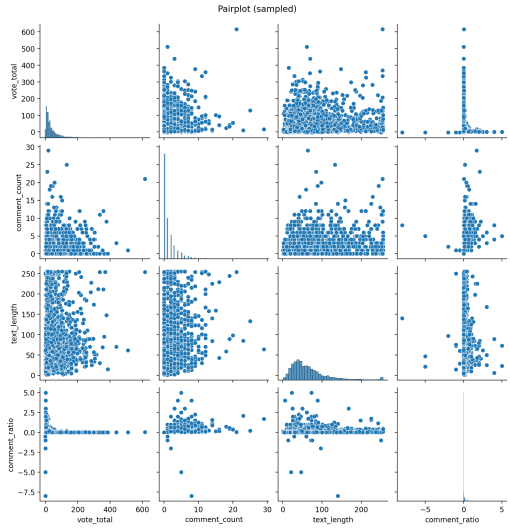

Figure 3: YikYak Pair Plot

the findings reported by Hessel and Lee (Hessel and Lee 2019).

4 Model Implementation

Initial Controversy Measurement Models (Reddit)

We first implemented three baseline classifiers to predict Reddit's platform-defined controversial flag: a Support Vector Machine (SVM), Logistic Regression, and a Random Forest. All models used a shared feature representation consisting of TF–IDF vectors extracted from cleaned post text (unigrams and bigrams, capped at 8,000 features) concatenated with a small set of numeric features, including comment-to-upvote ratio and sentiment polarity. To account for class imbalance, all models were trained using stratified train–test splits and class-weighted loss functions. Numeric features were scaled using a sparse-compatible standardization procedure before being concatenated with text features.

Logistic Regression Model: The Logistic Regression model showed similar behavior, achieving 0.52 accuracy.

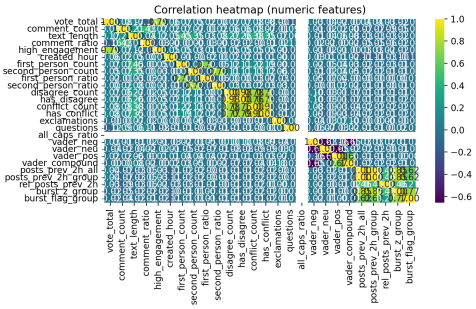

Figure 4: Correlation heatmap of numeric features in the YikYak dataset.

Classification Report:

	precision	recall	f1-score	support
0	0.74	0.61	0.67	1137
1	0.13	0.22	0.16	304
accuracy			0.52	1441
macro avg	0.44	0.41	0.41	1441
weighted avg	0.61	0.52	0.56	1441

Figure 5: Initial Logistic Regression Model Classification Report

Although recall for controversial posts was slightly higher than the SVM, overall performance remained close to chance, reflecting the difficulty of learning controversy signals from text and lightweight engagement features alone. Figure 5 shows the classification report.

SVM Model: The Linear SVM achieved an overall accuracy of 0.50. While it performed moderately well on non-controversial posts, it struggled to correctly identify controversial ones, yielding very low recall and F1 scores for the positive class. Figure 6 shows the classification report.

Random Forest Model: The Random Forest model achieved the highest accuracy among the three, with an accuracy of 0.61. However, this improvement was largely driven by strong performance on the majority class. Precision and recall for controversial posts remained extremely low, indicating that the model effectively defaulted to majority-class predictions. Figure blank shows the classification report. Figure 7 shows the classification report.

Classification Report:

	precision	recall	f1-score	support
0	0.72	0.60	0.65	1137
1	0.08	0.12	0.09	304
accuracy			0.50	1441
macro avg	0.40	0.36	0.37	1441
weighted avg	0.58	0.50	0.54	1441

Figure 6: Initial SVM Model Classification Report

Classification Report:

	precision	recall	f1-score	support
0	0.749	0.776	0.762	1137
1	0.034	0.030	0.032	304
accuracy			0.618	1441
macro avg	0.392	0.403	0.397	1441
weighted avg	0.598	0.618	0.608	1441

Figure 7: Initial SVM Model Classification Report

These results demonstrate that controversiality is difficult to predict reliably. Even models capable of capturing nonlinear feature interactions failed to generalize meaningful signals for the minority class.

Interpretation: Controversy Measurement Models (Reddit) Across all three models, performance on controversiality prediction remained near chance, with particularly poor recall for controversial posts. Although the Random Forest achieved higher overall accuracy, this gain was driven almost entirely by correct classification of the majority non-controversial class, rather than meaningful detection of controversy. Both linear models similarly struggled to identify controversial posts, indicating that controversy signals are not easily separable using text features and lightweight engagement metadata alone.

These results suggest that Reddit's platform-defined controversiality label captures complex, context-dependent dynamics that are not well reflected in surface-level linguistic patterns or simple interaction statistics. Unlike engagement, which reflects aggregate user attention, controversy depends on polarization within the voting population and may be influenced by factors such as community norms, thread context, and user identity. Consequently, even models capable of learning nonlinear interactions fail to generalize reliable signals for the minority class. This limitation motivated our shift toward engagement prediction, a more behaviorally grounded and tractable modeling target.

High Engagement Measurement Models (YikYak)

Given our results showing that controversy is highly context-dependent, we reframed the task to predicting high engagement. Engagement provides a clearer and more consistent behavioral signal than controversiality, and it is observable across both Reddit and YikYak.

To predict high engagement on YikYak posts, we implemented and compared three supervised classification models. Across all models, textual content was represented using a TF–IDF vectorization scheme. TF–IDF is a widely used weighting scheme for sparse text representations (GeeksforGeeks 2023). This representation allowed each model to capture both individual word usage and short phrase patterns associated with engagement.

Logistic Regression Model: First, we implemented a logistic regression classifier and scaled the features using StandardScaler.This was done to maintain compatibility with the sparse TF–IDF matrix, and concatenated with text

Classification Report:

	precision	recall	f1-score	support
0	0.92	0.88	0.90	3443
1	0.24	0.33	0.28	382
accuracy			0.83	3825
macro avg	0.58	0.61	0.59	3825
weighted avg	0.85	0.83	0.84	3825

Figure 8: Logistic Model Classification Report

features using hstack.The logistic regression classifier was trained with the liblinear solver, max_iter=2000, C=1.0, and class_weight="balanced" to account for the minority high-engagement class. After fitting, predicted probabilities for the positive class were thresholded at 0.6 to improve detection of high-engagement posts, rather than using the default 0.5 cutoff. The model classification report is shown in Figure 8:

The classification report indicated:

- Low Engagement (Class 0): Precision = 0.92, Recall = 0.85, F1-score = 0.88
- High Engagement (Class 1): Precision = 0.20, Recall = 0.33, F1-score = 0.25

These results reflect that the model is highly precise in identifying low-engagement posts, meaning that when it predicts a post as low-engagement, it is usually correct. Recall for low-engagement posts is also high, indicating that most low-engagement posts are correctly captured. In contrast, the model has low precision and moderate recall for high-engagement posts. This means that when it predicts high engagement, it is often wrong, and it misses a substantial number of actual high-engagement posts. The F1-score of 0.25 for high-engagement posts summarizes this poor balance, highlighting the difficulty of correctly identifying minority-class instances in an imbalanced dataset.

The confusion matrix (Figure 9) illustrates that the model is highly accurate for low-engagement posts but underdetects high-engagement posts, reflecting the challenge of predicting a minority class even when using additional numerical and burstiness features. It shows:

- True Negatives (2923): Low-engagement posts correctly predicted
- False Positives (520): Low-engagement posts misclassified as high-engagement
- False Negatives (255): High-engagement posts misclassified as low-engagement
- True Positives (127): High-engagement posts correctly predicted

The model's ROC AUC was 0.71 (Figure 10), indicating moderate discriminative power. These results demonstrate that combining textual, numerical, and burstiness features allows logistic regression to capture meaningful signals associated with post engagement. Adjusting the probability threshold improved sensitivity to high-engagement

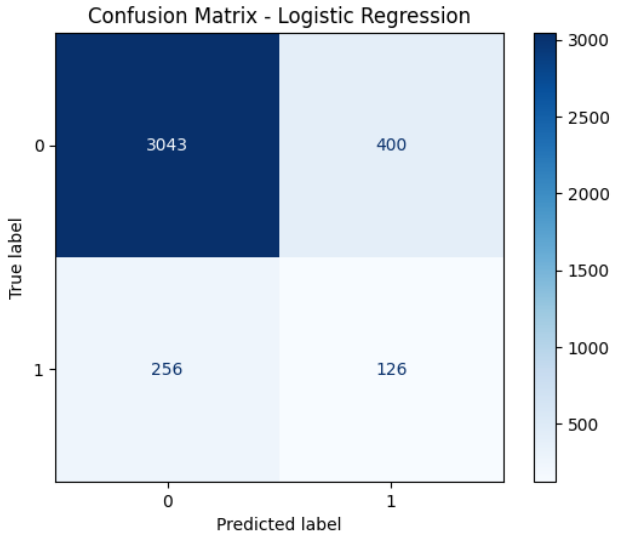

Figure 9: Logistic Model Confusion Matrix

posts, though recall remains limited. Overall, this approach provides a strong, interpretable baseline for engagement prediction on short-form social media content.

SVM Model: In addition to logistic regression, a Linear Support Vector Machine (SVM) was trained to predict high engagement in YikYak posts. The classifier used LinearSVC with a regularization parameter of C = 2.5, a maximum of 1,000 iterations, and class-weight balancing to mitigate the strong class imbalance between high- and low-engagement posts. Unlike logistic regression, this SVM model does not produce calibrated probabilities, but instead learns a decision boundary that maximizes the margin between the two classes. The model achieved an overall accuracy of 0.80, comparable to the logistic regression model. The model classification report is shown in Figure 11:

The classification report indicated:

- Low-engagement posts (class 0): precision = 0.92, recall = 0.85, F1-score = 0.88
- High-engagement posts (class 1): precision = 0.18, recall = 0.30, F1-score = 0.23

As with logistic regression, the model performs very well on the majority class. High precision for low-engagement posts indicates that predictions of low engagement are usually correct, and high recall suggests that most low-engagement posts are successfully identified. However, performance on high-engagement posts is substantially weaker. The low precision (0.18) indicates that many posts predicted as high engagement are actually low engagement, while the recall of 0.30 shows that the model fails to capture a majority of truly high-engagement posts. The resulting F1-score of 0.23 reflects this imbalance and difficulty in modeling the minority class.

The confusion matrix (Figure 12) illustrates, similar to logistic regression, the majority of errors arise from misclassification of high-engagement posts, either by missing

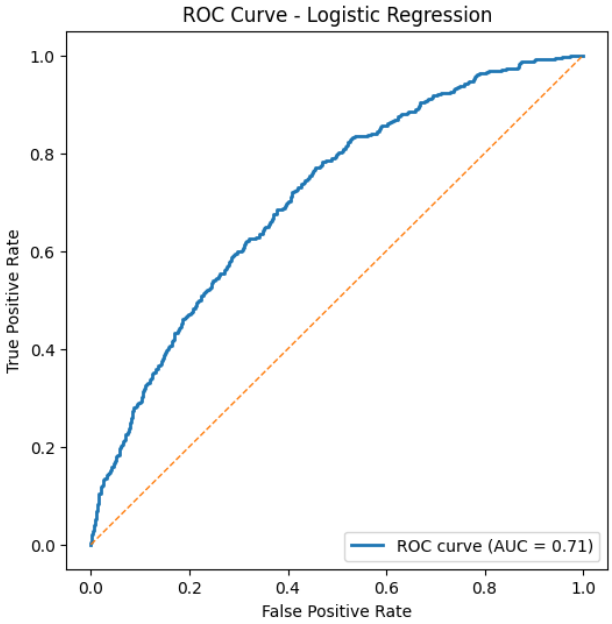

Figure 10: Logistic Model ROC CURVE: YikYak Dataset

	precision	recall	f1-score	support
0	0.92	0.85	0.88	3443
1	0.18	0.30	0.23	382
accuracy			0.80	3825
macro avg	0.55	0.58	0.55	3825
weighted avg	0.84	0.80	0.82	3825

Figure 11: SVM Model Classification Report

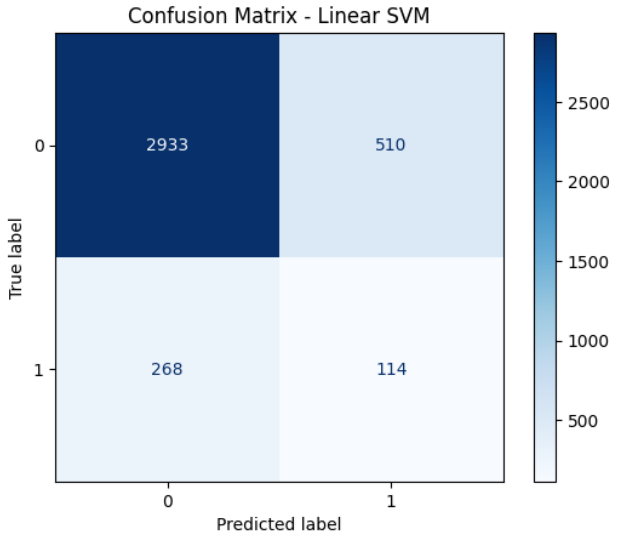

Figure 12: SVM Model Confusion Matrix

them entirely or by over-predicting high engagement for low-engagement content. It shows:

- True Negatives (2933): Low-engagement posts correctly classified
- False Positives (510): Low-engagement posts misclassified as high engagement
- False Negatives (268): High-engagement posts misclassified as low engagement
- True Positives (114): High-engagement posts correctly identified

The model's ROC AUC was 0.66 (Figure 13). This distribution shows that, similar to logistic regression, the majority of errors arise from misclassification of high-engagement posts, either by missing them entirely or by over-predicting high engagement for low-engagement content. The result suggests that while the SVM captures some separation between engagement levels, it does not substantially improve minority-class detection.

Random Forest Model: We implemented both standard and balanced Random Forest classifiers to predict high engagement on YikYak posts. The balanced variant used the BalancedRandomForestClassifier from the imblearn library with 500 estimators and built-in class balancing mechanisms. Both models were trained on the same feature representation combining TF-IDF text features with 14 numeric features including post length, pronoun ratios, sentiment scores, and burstiness metrics.

The Balanced Random Forest achieved an overall accuracy of 85.5 percent, with precision and recall for high-engagement posts at 0.249 and 0.225 respectively (F1-score = 0.237). The confusion matrix (Figure 15) revealed that the model correctly identified 3,184 low-engagement posts but misclassified 259, while detecting only 86 out of 382 high-engagement posts. Overall, our Standard Random

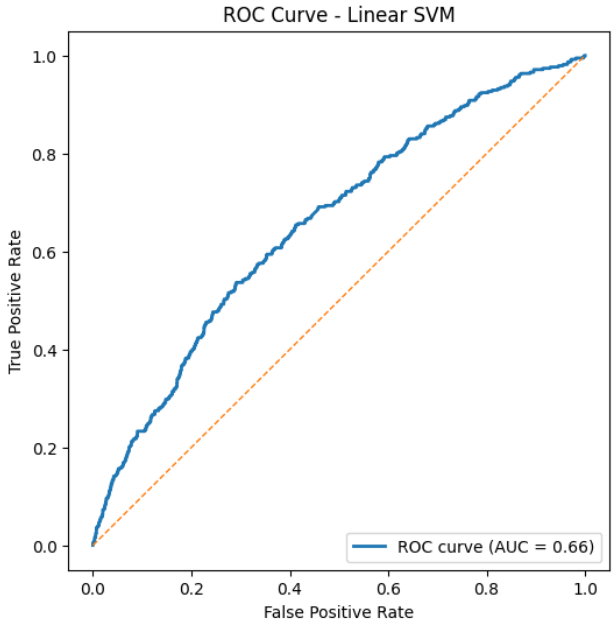

Figure 13: SVM Model ROC CURVE: YikYak Dataset

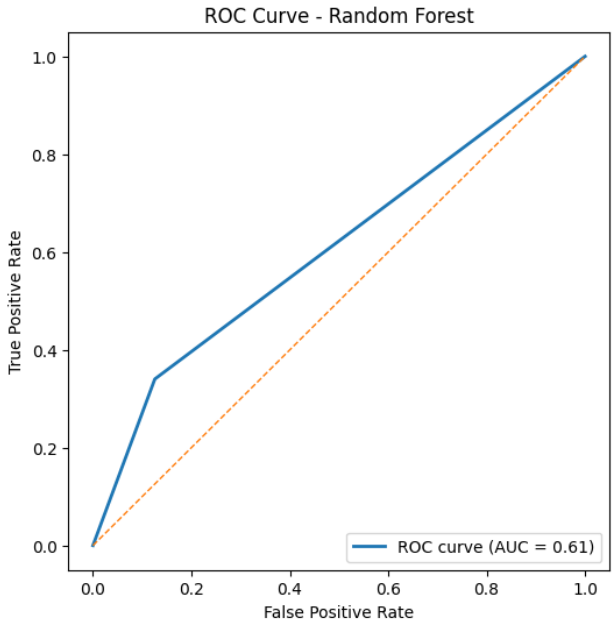

Figure 14: Balanced Random Forest Model ROC CURVE: YikYak Dataset

Forest Model continued to struggle with identifying high-engagement posts. We often found ourselves sacrificing recall for higher accuracy.

Interestingly, when we trained a standard Random Forest without explicit balancing, it achieved 90 percent overall accuracy but with virtually no recall for high-engagement posts (0.003). This demonstrates the critical importance of addressing class imbalance: unbalanced models tend to default to majority-class predictions, achieving high accuracy at the expense of minority-class detection. The balanced version showed significantly improved recall for high-engagement posts (0.225 vs. 0.003), though still far from optimal. However, a significant more number of true negatives were identified, as visible in the Confusion Matrix (Figure 16).

The Confusion Matrix for both the Standard (Figure 15) and Balanced Random Forests (Figure 16) indicated frequent errors in identifying high-engagement posts, similar to the errors found in our logistic regression model and our SVM model. The similarity in performance across different models indicates that the primary challenge is not the standard of model itself but rather feature representation and class imbalance.

Interpretation: High Engagement Measurement Models (YikYak) Across all three models, engagement prediction on YikYak remains challenging, with consistent difficulty in identifying high-engagement posts. Logistic regression achieved the strongest overall balance among the models, with the highest recall for the high-engagement class and the best ROC AUC (0.71), indicating moderate discriminative ability. The linear SVM performed similarly but exhibited slightly weaker separation, while both standard and balanced Random Forest models failed to substantially improve

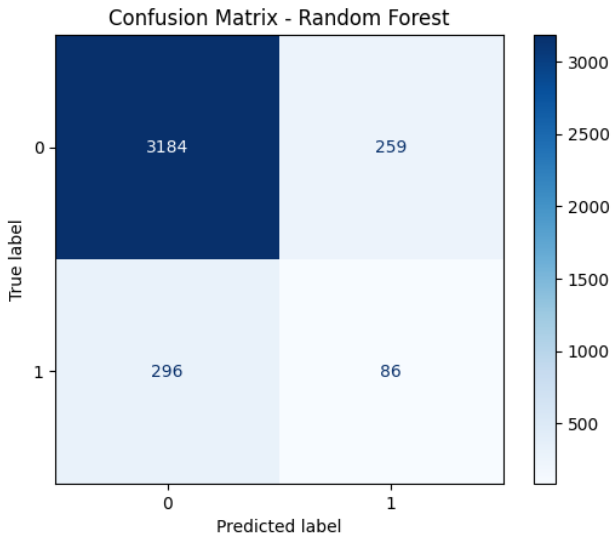

Figure 15: Confusion Matrix: Standard Random Forest

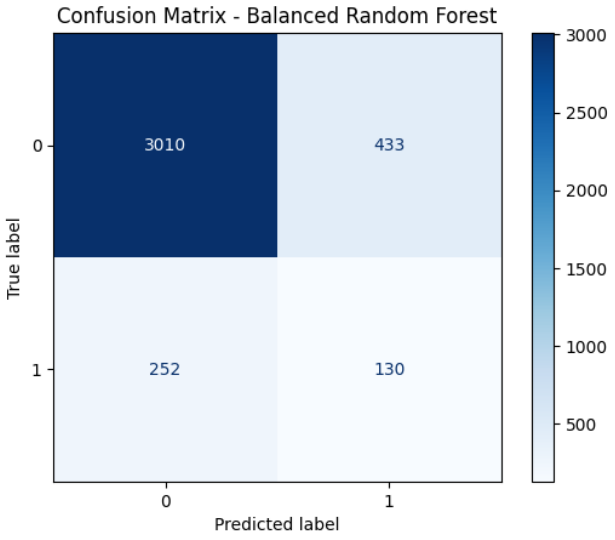

Figure 16: Confusion Matrix: Balanced Random Forest

minority-class detection despite higher overall accuracy.

The comparable performance of linear and tree-based models suggests that the primary limitation is not model expressiveness but the underlying feature representation and extreme class imbalance. Although burstiness and numerical features provide some signal, high-engagement posts remain difficult to distinguish from low-engagement posts using surface-level text and metadata alone. These results contrast with the stronger performance observed on Reddit and highlight that engagement on YikYak is more sparse, volatile, and dependent on short-term community dynamics. Overall, these findings indicate that while linear models provide a strong baseline, more expressive contextual or temporal features may be necessary to reliably capture high engagement on anonymous platforms like YikYak.

High Engagement Measurement Models (Reddit)

Although our initial focus was YikYak data, we then tested our models on the existing Reddit dataset in order to determine its classification effectiveness on other social media forums. So, we evaluated three supervised learning models for engagement prediction on Reddit data: Logistic Regression, a linear Support Vector Machine (SVM), and Random Forest–based classifiers. All models were trained and tested using the same stratified data splits and preprocessed feature representations.

Logistic Regression Model: Logistic Regression achieved strong performance, reaching 0.90 accuracy with balanced precision and recall for the high-engagement class (F1 = 0.80). These results indicate that a combination of TF–IDF text features and simple numeric metadata, such as post length, captures meaningful signals associated with user engagement. Figure 17 shows the classification report.

SVM Model: The linear SVM model produced nearly identical results, also achieving 0.90 accuracy and an F1-score of 0.80 for high-engagement posts. The consistency between these two linear models suggests that engagement-related linguistic patterns are largely linearly separable in the feature space. Figure 18 shows the classification report.

Classification Report:

	precision	recall	f1-score	support
0	0.93	0.94	0.93	1073
1	0.81	0.78	0.80	368
accuracy			0.90	1441
macro avg	0.87	0.86	0.86	1441
weighted avg	0.90	0.90	0.90	1441

Figure 17: Logistic Regression Model Classification Report

Classification Report:

	precision	recall	f1-score	support
0	0.93	0.94	0.93	1073
1	0.81	0.79	0.80	368
accuracy			0.90	1441
macro avg	0.87	0.86	0.87	1441
weighted avg	0.90	0.90	0.90	1441

Figure 18: SVM Model Classification Report

Random Forest Model: In contrast, the Balanced Random Forest model underperformed relative to the linear models. While it achieved an overall accuracy of 0.825, performance on the high-engagement class was substantially weaker, with a precision of 0.234, recall of 0.330, and an F1-score of 0.274. This indicates that although class balancing improved the model's ability to identify high-engagement posts compared to an unbalanced Random Forest, it still struggled to reliably distinguish minority-class instances. This outcome aligns with prior findings that tree-based models struggle with sparse, high-dimensional text representations. Figure 19 shows the classification report.

Interpretation: High Engagement Measurement Models (Reddit) Across all three models, engagement prediction on Reddit is substantially more successful than on YikYak. Both logistic regression and the linear SVM achieve strong and nearly identical performance, indicating that engagement-related signals in Reddit data are largely

Classification Report:

	precision	recall	f1-score	support
0	0.922	0.880	0.901	3443
1	0.234	0.330	0.274	382
accuracy			0.825	3825
macro avg	0.578	0.605	0.587	3825
weighted avg	0.853	0.825	0.838	3825

Figure 19: Random Forest Model Classification Report

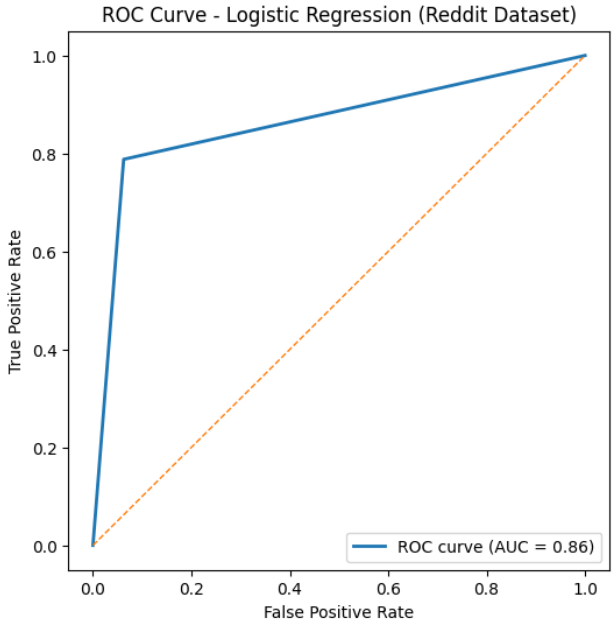

Figure 20: Logistic Regression ROC CURVE: Reddit Dataset

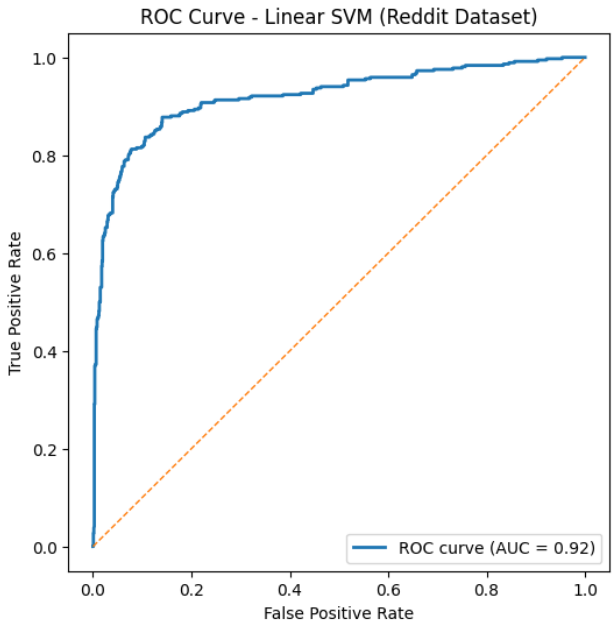

Figure 21: Linear SVM ROC CURVE: Reddit Dataset

linearly separable when represented using TF–IDF text features and simple numeric metadata. This consistency suggests that linguistic patterns associated with high engagement on Reddit are stable and well captured by sparse linear representations.

In contrast, the Random Forest model underperforms despite explicit class balancing, highlighting the limitations of tree-based methods on high-dimensional, sparse text features. These results reinforce that model complexity alone does not guarantee improved performance and that representation choice plays a central role. Overall, the strong performance of linear models demonstrates that engagement prediction is a tractable task on Reddit and provides a useful upper bound for comparison with the more challenging YikYak setting.

Interpreting Reddit vs. Yikyak Engagement Measurement Models

Figures 20–22 illustrate a clear performance gap between engagement prediction on Reddit and YikYak. On Reddit, all models achieve strong discriminative performance, with linear models in particular showing high AUC values (logistic regression: 0.86; linear SVM: 0.92). This suggests that engagement-related signals in Reddit data are relatively stable and linearly separable when represented using TF–IDF features.

In contrast, ROC curves for the YikYak dataset show substantially weaker performance across all models, with AUC values closer to chance. This difference highlights the increased difficulty of engagement prediction on YikYak, where engagement is sparser, more imbalanced, and more sensitive to short-term community dynamics. Together,

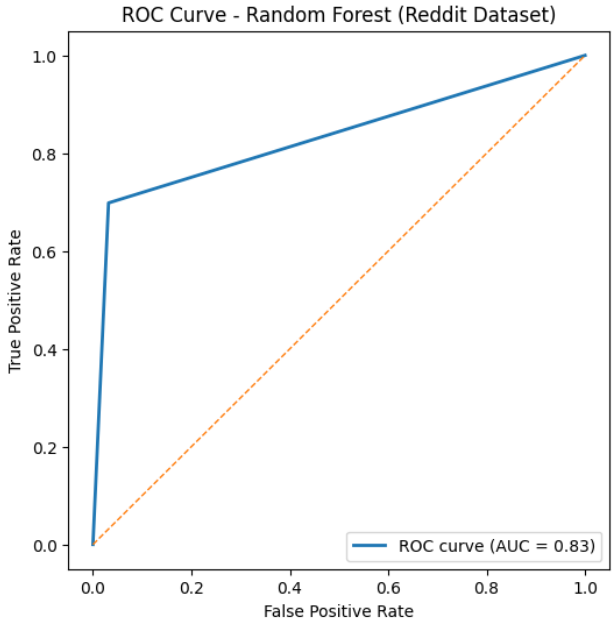

Figure 22: Random Forest ROC CURVE: Reddit Dataset

these results indicate that while text-based linear models generalize well on Reddit, additional contextual and temporal signals are necessary to model engagement effectively on YikYak.

5 Broader Impacts

Controversy vs. Engagement

Unlike prior approaches that focus primarily on predicting platform-specific controversiality labels (e.g., Reddit's controversial flag), our work emphasizes engagement prediction as a more stable and transferable modeling target. While controversiality is often defined through opaque or platform-dependent mechanisms, engagement reflects observable user behavior—such as commenting and voting—that is present across many social media platforms. This makes engagement prediction more suitable for cross-platform analysis and deployment, particularly in settings where explicit controversy labels are unavailable or inconsistently defined.

Prior work has shown that controversy prediction often depends heavily on discussion structure and community-specific norms, limiting generalization across platforms (Hessel and Lee 2019). By reframing the task around engagement, we aim to capture a broader signal of collective attention that is less tightly coupled to a single platform's labeling system.

Bias and Community Skew

Our YikYak dataset reflects a single, geographically bounded college community, which introduces inherent demographic and cultural skew. Language use, humor, and what constitutes engaging or provocative content at Davidson College may differ substantially from other campuses or regions. As a result, the learned engagement signals may encode local norms rather than universal patterns of online interaction. Because content visibility and participation are constrained by geographic location, engagement patterns are shaped not only by textual content but also by who is physically nearby at a given moment. Events such as campus activities, weather, or academic stress cycles may significantly influence engagement in ways that are difficult to capture using text and lightweight metadata alone. It's likely that our model could not be applicable to YikYak data on other campuses.

Additionally, we ran the risk of domain transfer risk with our original strategy of training on Reddit data and transferring the model to YikYak. Training on Reddit data introduces its own biases, as Reddit communities tend to overrepresent particular demographics and discourse styles. Additionally, Reddit's definition of controversiality is derived from voting behavior and may not align with how controversy manifests on YikYak. When models trained on Reddit are applied to YikYak (our original strategy), these biases may manifest as systematic misclassification of posts.

Ethical Use

Developing a model that predicts online controversy involves significant ethical considerations. Although we used a reverse-engineered API wrapper, we were careful and made sure to restrict our scraping to publicly visible content and avoided collecting any identifying information. We wanted to align our work with responsible data-collection practices. At the same time, we recognize that any model trained on community behavior risks learning and reinforcing biases already present within that community. Posts written in certain dialects or certain identities might receive more negative or polarized responses, and a model could mistakenly interpret that as controversial.

Furthermore, while such a system could help moderators anticipate conflicts or help researches study online discourse, it could also contribute to over-moderation or discourage users from expressing their opinion, which could be socially important, but uncomfortable. Our project highlights these concerns so that any future use of a similar system can be done with transparency and care.

Model Improvements

Prior work has demonstrated that engagement and controversy are often driven by discussion dynamics and temporal structure rather than post text alone ((Hessel and Lee 2019), (Fan et al. 2021). Our findings align with this observation: linear text-based models perform well on Reddit, where engagement patterns are relatively stable, but struggle on YikYak, where engagement is sparse, volatile, and highly time-dependent.

Our feature set focuses primarily on surface-level linguistic cues, sentiment, and short-term activity patterns. While these features capture some engagement signals, they may overlook specific triggers such as campus-specific buzzwords, slang, or references to local events. Incorporating dynamically learned topic or keyword features, particularly those grounded in local context, could improve performance.

6 Conclusions

Although our initial goal was to predict controversiality, empirical results revealed that controversy was difficult to model reliably using text and lightweight engagement features alone. This motivated a shift toward engagement prediction, a more observable and behaviorally grounded target.

Our results show that engagement prediction is substantially more tractable on Reddit than on YikYak, highlighting the role of platform structure and community dynamics in shaping online interaction.

The comparatively weaker performance on YikYak underscores the limits of surface-level text modeling in hyperlocal, anonymous environments.

Future work may benefit from incorporating richer temporal, contextual, and location-aware features, as well as dynamically learned representations of local discourse. Despite these limitations, our work demonstrates that reframing predictive tasks around transferable behavioral signals can yield more robust insights across platforms.

7 Acknowledgements

We would like to acknowledge the contributions of the SideChat API wrapper developers, whose open-source tools enabled the data collection necessary for this study...

8 Model Card

Model Cards (especially the “Limitations” or “Warnings” section)

They’re used in papers and reports to highlight:assumptions you made, when the model shouldn’t be used, risks or biases, data constraints, reproducibility notes.

References

Esposito, M. 2023. Reverse engineering the yikyak api. https://matthew.science/posts/yikyak/. Accessed 2025-11-25.

Fan, J.; Han, J.; Li, B.; Tian, X.; and Xin, Z. 2021. Predicting scores and controversialities of reddit posts using machine learning. *National University of Singapore*.

GeeksforGeeks. 2023. Understanding tf-idf (term frequency–inverse document frequency). https://www.geeksforgeeks.org/machine-learning/understanding-tf-idf-term-frequency-inverse-document-frequency/. Accessed 2025-03-XX.

Hessel, J., and Lee, L. 2019. Something’s brewing! early prediction of controversy-causing posts from discussion features. *arXiv preprint arXiv:1904.07372*.

Lindley, M. 2023a. micahlt/sidechat.js. https://github.com/micahlt/sidechat.js. GitHub repository, accessed 2025-11-12.

Lindley, M. 2023b. sidechat.js: Unofficial sidechat/yikyak api wrapper. https://micahlindley.com/sidechat.js/. Accessed 2025-11-30.

Reddit, Inc. 2024. Reddit api documentation. https://www.reddit.com/dev/api/. Accessed 2025-12-01.