2025

Comparison of Machine Learning Models for X Classification

CISC684-010: INTRO TO MACHINES LEARNING

COREY ZHANG

This study focuses four machine learning models. We use SVM, Logistic Regression, Random Forest, and Decision Tree to classify tweets using three different feature sets: Full, Reduced (first five features), and Expanded (interaction terms). The dataset was balanced using SMOTE to ensure fair representation of both classes, and models were assessed based on accuracy, precision, and recall. The goal was to determine the optimal model and feature set for accurate classification.

Feature Selection

Feature selection plays a critical role in model performance. Eight features were selected to represent tweet characteristics. These features were chosen to capture tweet length, user engagement, and metadata properties that influence classification performance.

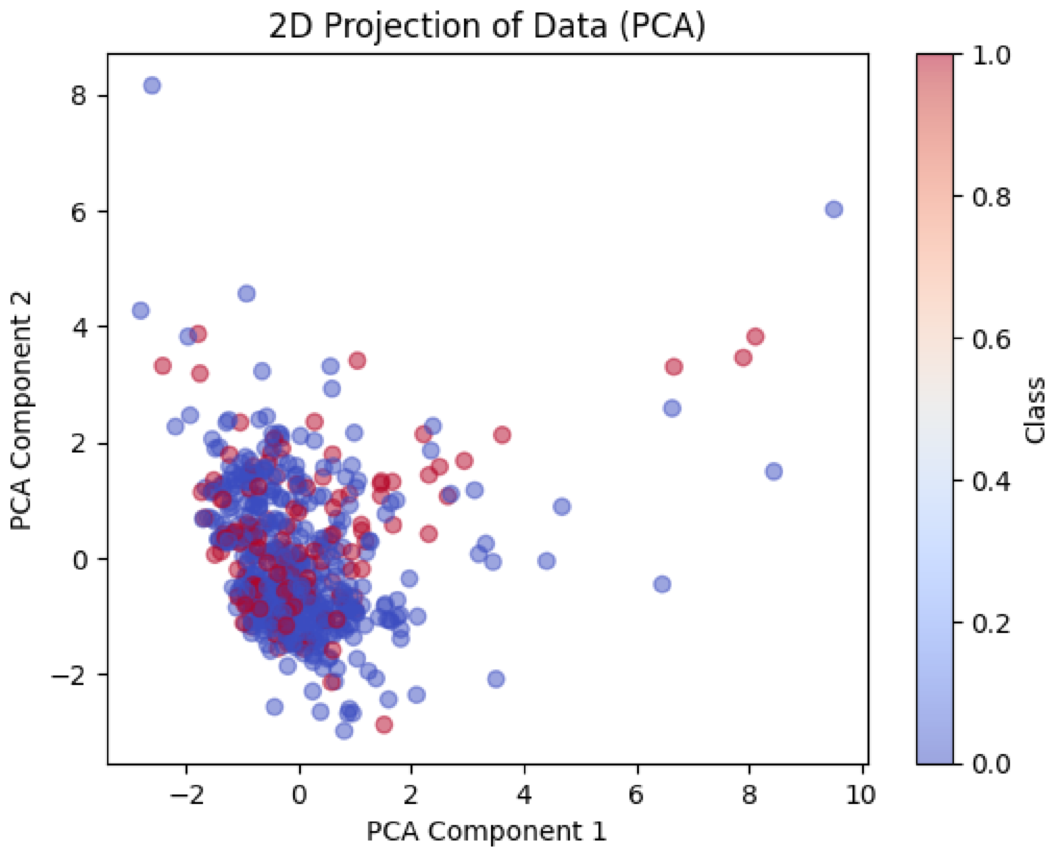
Feature Name	Description	Reason for Selection
Username Length	Number of characters in username	Longer usernames may indicate bot accounts.
Has Profile Image	1 if user has a profile picture, 0 otherwise	Profile images may indicate real accounts.
Has URL Bio	Binary: 1 if user has a URL in their profile bio, 0 otherwise	Bots often include links in their bio.
Follower-Friend Ratio	Ratio of followers to friends	Helps detect bot-like or suspicious accounts.
Tweet Length	Number of characters in the tweet	Longer tweets may have different engagement.
Has Mentions	1 if tweet contains mentions (@user), 0 otherwise	Mentions indicate interaction with other users.
Has Hashtags	1 if tweet contains hashtags (#topic), 0 otherwise	Hashtags reflect engagement in trends.
Day of the Week	0 = Monday, ..., 6 = Sunday	Some tweets have different engagement patterns based on the day.

These features were tested across three different feature sets:

- Full Feature Set: Includes all eight features.
- Reduced Feature Set: First five features only.
- Expanded Feature Set: Includes all features plus interaction terms between key features (e.g., Follower-Friend Ratio × Has URL in Bio).

Visualization of Data using PCA

To understand the structure of the dataset, Principal Component Analysis (PCA) was applied to reduce the feature space to two principal components. The resulting 2D projection of the dataset is shown below:



The red points represent one class, while the blue points represent the other. The visualization suggests significant overlap between the classes, indicating that classification may be challenging. However, noticeable clusters suggest that machine learning models can still identify patterns to distinguish between the two categories.

Results Overview

Model	Feature Set	Accuracy	Precision	Recall
SVM	Full	0.6325	0.3696	0.5484
Logistic Regression	Full	0.6239	0.3617	0.5484
Random Forest	Full	0.7265	0.4878	0.6452
Decision Tree	Full	0.6239	0.3774	0.6452
SVM	Reduced	0.5043	0.3099	0.7097
Logistic Regression	Reduced	0.5897	0.3455	0.6129
Random Forest	Reduced	0.7436	0.5128	0.6452
Decision Tree	Reduced	0.6068	0.3585	0.6129
SVM	Expanded	0.5641	0.3387	0.6774
Logistic Regression	Expanded	0.6154	0.3704	0.6452
Random Forest	Expanded	0.7521	0.5250	0.6774
Decision Tree	Expanded	0.6068	0.3684	0.6774

Model Comparison

Random Forest consistently outperformed other models, achieving the highest accuracy (75.21%) and precision (52.50%) with the expanded feature set. It also maintained strong recall (64.52%–67.74%) across all feature sets, making it the most reliable model.

SVM prioritized recall (70.97%) when using the reduced feature set, making it useful when detecting positive cases is more important than minimizing false positives. However, its precision was low (30.99%), meaning it misclassified many tweets as positive when they were not.

Logistic Regression performed moderately well, balancing precision and recall but still falling behind Random Forest. Decision Trees showed similar recall to Random Forest but had lower precision and accuracy, making them less reliable in comparison.

Feature Set Impact

Using the full feature set resulted in the highest accuracy for most models, while reducing features increased recall but lowered precision. Expanding the feature set provided minor improvements in recall but did not significantly enhance accuracy.

The best feature set depends on the goal:

- Full features → Best for balanced accuracy.
- Reduced features → Best for high recall (catching more positive tweets).
- Expanded features → Slight recall improvements but no major accuracy gains.

The following table summarizes SVM’s performance across feature sets:

Feature Set	Accuracy	Precision	Recall
Full Feature Set	0.6325	0.3696	0.5484
Reduced Feature Set	0.5043	0.3099	0.7097
Expanded Feature Set	0.5641	0.3387	0.6774

These results indicate that reducing features increases recall at the cost of lower precision and accuracy, while expanding features has minimal impact on overall accuracy.

Conclusion

Random Forest is the best model, delivering the highest accuracy and maintaining a strong balance between precision and recall. SVM is valuable for high-recall tasks, but not with precision, making it more likely to get false positives. Expanding features does not guarantee performance improvements, while reducing features may boost recall with the cost of accuracy.

Recommendations

1. Use Random Forest for best overall accuracy and performance across all.
2. Using SVM with reduced features if recall is the priority, such as detecting as many positive tweets as possible.
3. Avoid overly complex feature expansions, as they do not significantly improve accuracy.
4. Further improvements could be achieved with hyperparameter tuning.

This study highlights the importance of selecting the right model and feature set. The Random Forest model with the full feature set is the best choice, achieving the highest accuracy and best balance of precision and recall.

Citation

ChatGPT. (2025). *Comparison of Machine Learning Models for Tweet Classification* (with assistance from ChatGPT). OpenAI.