TikTok claims classification project

Executive summary report for TikTok prepared by the TikTok data team

Overview

The project aims to develop a machine learning model to classify reports as either claims or opinions. Previous analysis of the available data identified video engagement levels as strong indicators of claim status. The team is confident that the final model meets all performance requirements, marking the successful completion of the project's objectives.

Problem

TikTok receives a high volume of user reports, making it impossible for moderators to review every video. Claim-based videos are more likely to violate terms of service. To address this, TikTok seeks to identify and prioritize them. The growing backlog allows harmful claim videos to remain unchecked and shared.

Solution

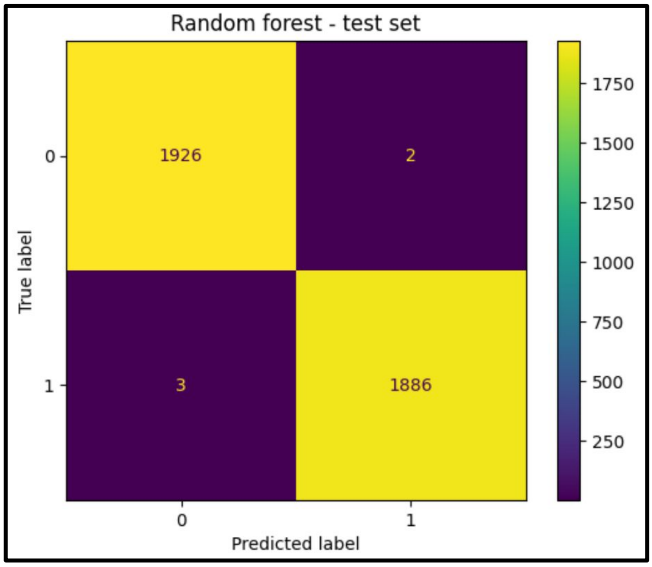
The data team developed two tree-based models, selecting the final one by recall score. The model was tested on a separate dataset to estimate performance. This classification model distinguishes claims from opinions, helping moderators identify harmful content. Prioritizing high-risk reports improves review efficiency and reduces workload.

Details

The data team developed two tree-based models, selecting the final one based on recall score. Tested on a separate dataset, the model classifies claims vs. opinions, helping moderators prioritize harmful content and reduce workload. The test set showed near-perfect performance, misclassifying only five out of 3,817 samples.

Video engagement metrics—view count, likes, shares, and downloads—were the strongest predictors, with high engagement linked to claims. No opinion video exceeded 10,000 views. Random Forest outperformed XGBoost, achieving a 0.9948 recall and 99.92% test accuracy. Given its strong performance, the model is ready for deployment.

Confusion matrix for the champion RF model on test holdout data shows only five misclassified samples out of 3,817.



The confusion matrix shows the model's predictions vs. actual labels. The top-left represents correctly classified class "0" instances, while the bottom-right shows correctly classified class "1" instances. Misclassifications appear in the top-right (false positives) and bottom-left (false negatives). Lighter colors indicate higher values.

Next Steps

The model performed well on test data, but the team recommends further evaluation on more user data and monitoring engagement distributions for robustness. While some opinions may be misclassified, the focus is on identifying harmful content. Misclassifications can be corrected through further processing, ensuring accuracy while limiting malicious claims.