Executive Summary – Claim Detection Model for TikTok Videos

Using Machine Learning to Support Content Moderation and Ensure Information Integrity

Overview

With the rapid dissemination of short-form video content on platforms like TikTok, distinguishing between personal opinions and potentially misleading claims has become increasingly important. Machine learning provides a scalable solution to support content moderation efforts and safeguard users from misinformation.

Problem

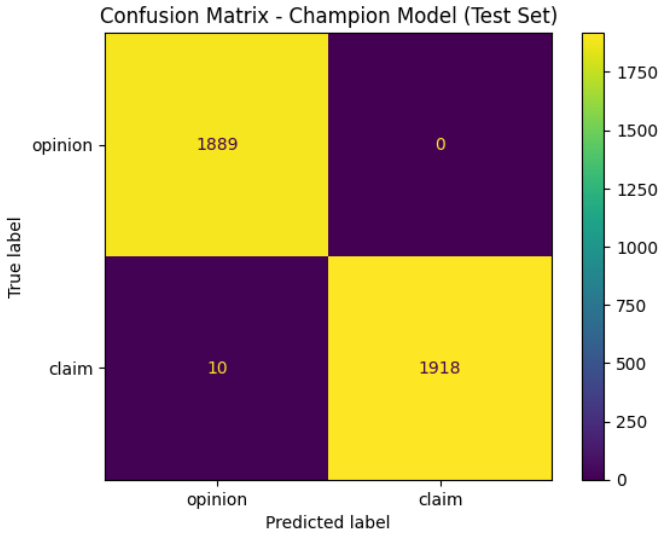
Manual moderation of millions of videos is time-consuming and prone to error. TikTok needed an efficient and accurate way to identify videos that present unverified claims, as these can influence public opinion and spread misinformation rapidly.

Solution

We developed and evaluated two tree-based classification models using metadata and video transcript features to predict whether a TikTok video is an “opinion” or a “claim.” The Random Forest classifier emerged as the champion model, delivering near-perfect results with high recall and precision. This model can help automate claim detection and prioritize content review workflows.

Details

Two models were developed and evaluated: Random Forest and XGBoost. Both achieved outstanding performance on the validation set, with near-perfect accuracy, precision, recall, and F1 scores. However, Random Forest slightly outperformed XGBoost in identifying claims — the target class of interest — with fewer false negatives. The final Random Forest model (champion model) was tested on unseen data. The confusion matrix shows only 10 misclassified claims out of 3,807 total observations. This means the model correctly flagged 1,918 out of 1,928 claims (recall ≈ 99.5%) and did not misclassify any opinions as claims, which is essential to avoid false accusations. These results confirm the model’s robustness and reliability in real-world application.Feature importance analysis revealed that engagement metrics — particularly video views, likes, and shares — were the strongest predictors, while profile-related variables such as verification status contributed little to the model's decisions.



Next Steps

Before deployment, the model should be validated on additional user segments and monitored for shifts in video engagement patterns. Future improvements could include text-based features (e.g., sentiment, source detection) to enhance robustness. Regular bias checks are recommended to ensure ethical use.