Predictive Modeling for Claims Classification

Final Report and Recommendations

Overview

This project was initiated to address the challenge of manually reviewing large volumes of user-reported videos. We successfully developed a machine learning model to automatically distinguish between “claims” and “opinion”, enabling our moderation teams to work more efficiently and effectively. The final model is exceptionally accurate and ready for deployment.

Problem

TikTok's content moderators face a significant backlog of user-reported videos. Manually reviewing each report to identify videos containing unverified "claims" is time-consuming and inefficient. This leads to delays in addressing potentially harmful content and creates a significant operational burden.

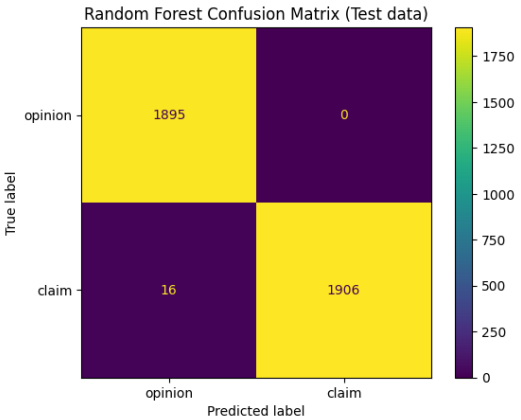
Solution

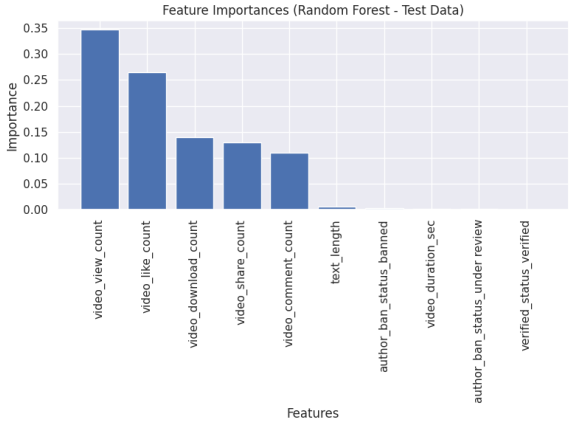
We built and trained a **Random Forest classification model** that analyzes video metadata and user engagement patterns to predict whether a video contains a “claim”. By flagging videos that are highly likely to be claims, the model allows our moderation team to prioritize their queue, focusing their efforts where they are needed most.

Details

The final model's performance on unseen test data exceeded all expectations:

- **100% Precision:** Every single video the model identified as a "claim" was correct. This means **zero moderator time will be wasted** on false alarms.
- **99.2% Recall:** The model successfully found 99.2% of all "claims" in the test set, ensuring very few go unreviewed.
- **Key Insight:** The model's most predictive features were **user engagement metrics** (views, likes, shares, etc.). This reveals that the community's reaction to a video is a more reliable signal of a "claim" than the author's status.





Next Steps

- **Deploy the Model:** Immediately integrate the *Random Forest* model into the content moderation workflow to prioritize the review queue. We should begin with an A/B test to precisely quantify the gains in moderator efficiency.
- **Invest in Model V2 (with NLP):** Charter a follow-up project to enhance the model by incorporating Natural Language Processing (NLP) to analyze the actual text of video transcriptions.
- **Review Internal Heuristics:** Re-evaluate any existing moderation rules that rely heavily on an author's verification or ban status, as our model has demonstrated these are weak predictive indicators compared to engagement patterns.