TikTok Claim Classification: Exploring Engagement Patterns

Executive Summary of Key Insights from Exploratory Data Analysis

ISSUE / PROBLEM

The TikTok data team seeks to develop a machine learning model to classify user-submitted videos as either claims or opinions. In this phase of the project, the dataset must be explored, cleaned, and structured to prepare for model development.

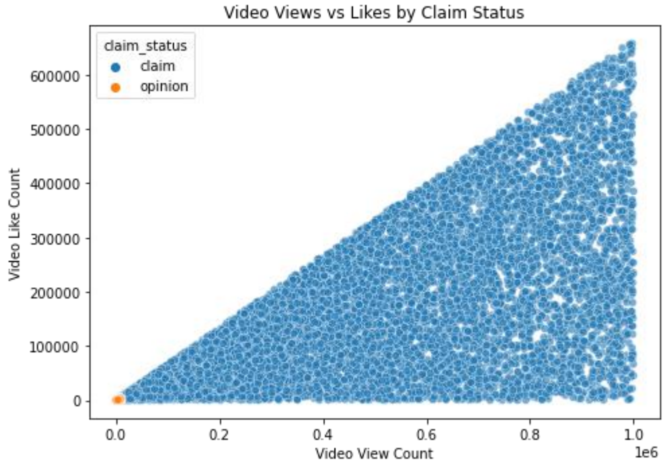
RESPONSE

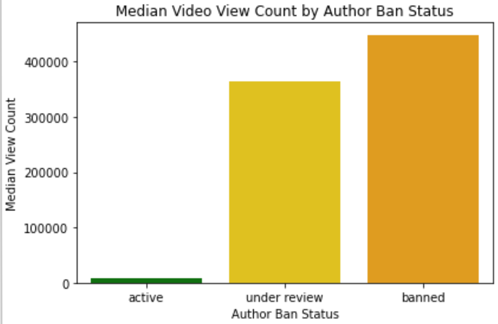
A preliminary exploratory data analysis (EDA) was conducted to assess the dataset's structure, quality, and key variables. Engagement metrics were examined, new ratios were created (likes/views, shares/views, comments/views), and patterns between claim status and author characteristics were identified to support future model development.

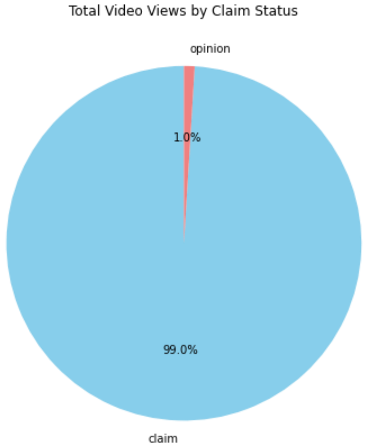
IMPACT

The exploratory analysis revealed a balanced dataset between claims and opinions, but highlighted a strong association between claim videos and higher engagement, particularly from banned or under-review authors. These patterns suggest the predictive model should incorporate engagement metrics and author status, while also accounting for missing values to ensure robustness.

A key component of this project's exploratory data analysis involves visualizing relationships between engagement metrics. The scatterplot below shows a strong positive correlation between views and likes, especially for claim videos, which dominate the upper range of both variables. This highlights a potential predictive relationship between engagement and claim classification.







This indicates that videos from non-active authors tend to receive much more attention and visibility.

The overall view count for claim videos is overwhelmingly higher than that of opinion videos. Despite the number of claim and opinion videos being nearly equal, claim videos account for the vast majority of total views. This suggests that claim videos are more likely to go viral or attract greater attention from viewers.

KEY INSIGHTS

Balanced Dataset: The dataset contains 9,608 claim videos and 9,476 opinion videos, ensuring near-equal representation (50.3% vs. 49.7%) for classification modeling.
Higher Engagement for Claims: Claim videos consistently show higher views, likes, and shares compared to opinions—especially among banned and under-review authors.
Skewed Distributions: Variables such as video views, likes, and comments are heavily right-skewed, with most videos receiving low engagement and a few viral outliers.
Moderation Linked to Virality: Authors who are banned or under review tend to have higher median engagement, suggesting a connection between moderation and viral content.
Missing Data: Approximately 300 entries (1.5%) have null values in key fields. These should be considered in future modeling steps to avoid biased predictions.