Executive Summary

Module 2 of the TikTok Claims Classification Project

Overview

This report presents the findings from the exploratory data analysis (EDA) of the TikTok dataset, which contains information about videos classified as either claims or opinions, along with various engagement metrics and author details.

Objective

The primary objective of this EDA was to gain insights into the provided dataset, including the distribution of claims and opinions, factors correlating with a video's claim status, and factors influencing its engagement level. These insights will inform the development of a claims classification model for TikTok videos.

Results

1. Distribution of Claims and Opinions: The dataset comprises 19,382 samples, with 49.6% (9,608) classified as claims and 50.4% (9,774) as opinions.

2. Factors Correlating with Claim Status: Videos classified as claims tend to receive significantly higher views compared to opinion videos. The engagement rates (likes, comments, and shares per view) for claim videos are generally higher than those for opinion videos, indicating more positive audience response.

3. Factors Correlating with Engagement Level: Author ban status strongly correlates with engagement level. Videos from banned authors or authors under review receive substantially more views, likes, and shares compared to those from active authors. For claim videos specifically, banned authors have slightly higher rates of likes and shares per view compared to active authors or those under review. For opinion videos, active authors and those under review receive higher engagement rates across all categories (likes, comments, shares) than banned authors.

Next Steps

1. Conduct further exploratory analysis.
2. Perform data preprocessing and feature engineering to prepare the dataset for machine learning modeling.
3. Develop and evaluate supervised learning models for claims classification.