Executive Summary: TikTok Claims Classification Project

Milestone 2 - Understanding the Data and Initial Findings

Project Overview

This project aims to develop a machine learning model that can accurately classify user-submitted TikTok videos as either "claim" or "opinion," contributing to TikTok's efforts in improving content moderation and combating misinformation.

Key Insights

Balanced Dataset: The dataset contains a near-equal distribution of "claim" and "opinion" videos, which is beneficial for training a robust and unbiased machine learning model.

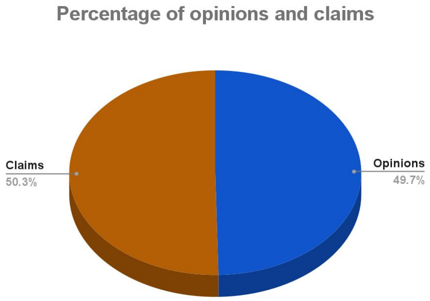
Higher Engagement for Claims: Claim videos generally exhibit significantly higher engagement levels (views, likes, shares, comments) compared to opinion videos, suggesting that engagement metrics could be strong predictors for classification.

Association with Banned Authors: A notable portion of claim videos are from authors who have been banned from TikTok, indicating a potential correlation between claims and policy violations.

Details

Distribution of Claims and Opinions:

- Claims: 9,608 videos (approximately 50%)
- Opinions: 9,476 videos (approximately 50%)
- This balance is visualized in the pie chart below.



Engagement Metrics:

- Claim videos have substantially higher average view counts (501,029.45) compared to opinion videos (4,956.43).
- Other engagement metrics (likes, shares, comments) also tend to be higher for claim videos.

Next Steps

- **Refine Data Understanding:** Perform in-depth Exploratory Data Analysis (EDA) on key variables.
- **Prepare Data for Modeling:** Conduct feature engineering and selection to optimize model performance.
- **Develop and Evaluate Models:** Train and assess machine learning models for accurate claim classification.
- **Deploy and Monitor:** Implement the chosen model and track its performance over time.