TikTok Claims Classification Project - Preliminary Data Summary

Executive Summary Report I
Milestone 2

ISSUE / PROBLEM

The TikTok data team aims to develop a machine learning model to help classify user-submitted claims. As a first step, they need to organize the raw dataset to prepare it for exploratory data analysis (EDA).

IMPACT

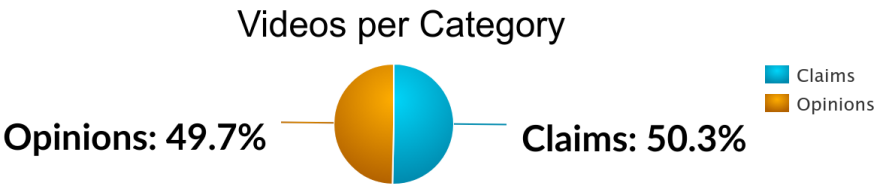
The findings from this initial analysis will inform the next stages of the project. After reviewing the dataset, the variable for the claim status emerged as particularly valuable. Two other critical variables for predicting video outcomes were also identified: the number of views the video received and the duration of the video (in seconds). These are essential factors to consider in the development of future prediction models.

RESPONSE

The data team conducted a preliminary investigation of the claims classification dataset to uncover key relationships between variables. Given the need for a classification of user claims, the team analyzed the distribution of claims and opinions to better understand the proportion of each type of video content.

KEY INSIGHTS

- The dataset has a nearly equal number of opinion and claim videos. This balance allows the team to proceed with confidence in the analysis, knowing that claims and opinions are evenly represented (9,608 for videos labeled as claims and 9,476 for videos labeled as opinions).
- With the identification of key variables and the completion of the initial investigation, the team is now ready to begin exploratory data analysis.



	Mean View Count	Median View Count
Claim	501029.45	501555.00
Opinion	4956.43	4953.00