TikTok project: detect claims and opinions

Milestone #2: Exploratory data analysis (EDA)

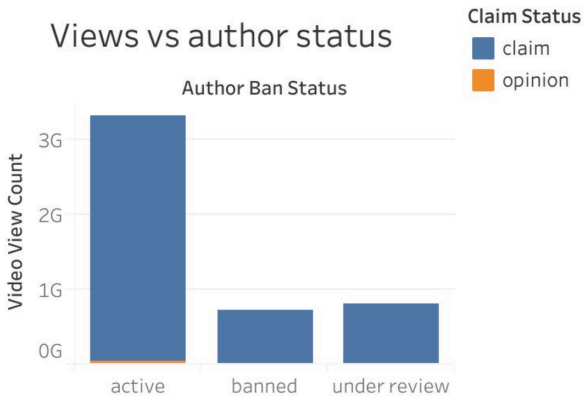
Overview

The objetive of this project is to implement a machine learning model to classify the videos uploaded to TikTok into claims (unsourced statements) and opinions (personal thoughts) in order to reduce moderating times and reduce economical costs.

Objective

- In this stage of the project, we performed EDA to explore, clean, structure and analyze the dataset containing different metrics (columns) for each video (rows).
- The results from this analysis will provide insights which will be considered during the development of the machine learning model.
- During the analysis, we focused on examining the engagement variables, especially the counts and central tendency measures for the views, shares and likes.

Results



- The analysis of the distribution of the engagement variables reveals most of the videos in the platform have reduced engagement, whilst a smaller group of videos (viral videos), achieve much higher engagement. This is, the distributions are all skewed to the right instead of normal.
- There are many more not verified than verified users, but when the users are verified, they tend to post many more opinion videos than claim videos. Users banned and under review seem to post more claim videos than opinions, while active users seem to post both claims and opinions. The view count may be indicative of the claim status.
- We found approximately ~20% of the samples corresponding to engagement variables represent outliers. This sums up to the missing/null values detected in the previous deliverable, which focused on the Discovering phase of the EDA (or data inspection).

Next Steps

According to the evidence gathered via EDA, the machine learning model will have to account for the null values and for the imbalance between claims and opinions with respect to the user status to produce accurate predictions. The view count may also be useful as an indicator of the claim status.