TikTok project: detect claims and opinions

Milestone #1: Data inspection

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

For this milestone, the objetive was to load the dataset and perform a preliminary data inspection: explore the columns, data types, target variable, create new variables and prepare the data for further analysis in subsequent milestones.

Results

claim_status	author_ban_status	count
claim	active	6566
claim	banned	1439
claim	under review	1603
opinion	active	8817
opinion	banned	196
opinion	under review	463

- The dataset contains 19382 videos (rows) and 11 features (columns).
- The dataset is balanced since contains approximately the same number of claim videos and opinion videos (9608 and 9476, respectively).
- The dataset present null or missing values in the following columns: *claim_status, video_transcription_text, views, likes, shares, downloads and comments.*
- Claim videos seem to have higher engagement compared to opinion videos, as suggested by higher mean like, comment and shares per view.
- Group analysis of the *claim_status* and the *author_ban_status* suggests banned users mainly share claim videos, while active users evenly share claim and opinion videos.

Next Steps

After performing the preliminary data inspection of the TikTok dataset which was presented in this executive summary, exploratory data analysis (EDA) can begin. Through EDA, the data team will:

- Fill the missing values.
- Examine the relationships between the columns.
- Perform statistical comparisons.