Executive summary

Milestone #2 of the TikTok claim detection project

Overview

The objective of the project is to implement a machine learning classifier to detect the claims and opinions corresponding to the videos shared in the platform.

Problem

In this milestone, the objective is to load the raw dataset, perform a preliminary inspection and prepare the data for further analysis.

Solution

The data team loaded and inspected the dataset. During the analysis, we focused on examining the engagement metrics, especially the counts and central tendency measures for the views, shares and likes.

Details

```
claim      9608
opinion    9476
Name: claim_status, dtype: int64
```

The claim_status column holds the target variable for this project. It represents the status of a video (claim or opinion). From the figure to the left, it is evident that the dataset is very balanced.

```
claim_status  author_ban_status
claim         active               6566
              banned               1439
              under review         1603
opinion       active               8817
              banned                196
              under review          463
Name: video_id, dtype: int64
```

The group analysis of the claim_status and the author_ban_status reveals that, in relative terms, many more users were banned when sharing claim videos compared to opinion videos.

Further analysis showed that the claim videos produced better engagement metrics (likes, shares and comments per view). This evidences a trend regarding the distribution of unsourced information in the platform.

Next Steps

After checking the dataset is balanced and providing preliminary information about the data, the exploratory data analysis (EDA) can begin. With this analysis, the data team will fill the missing values, examine the relationship between the dataset columns and perform statistical comparisons.