Claims Classification Project | Milestone Two

Executive summary prepared for TikTok leadership by the TikTok data team

Overview

TikTok leadership asked the TikTok data team to develop a ML model that will accurately predict whether a video contains a claim or offers an opinion. A powerful model can help us better understand user submissions, reduce the backlog of user reports and prioritize them more efficiently, and, as a result, improve users experience.

Problem

Increased user submissions is a major concern for TikTok because effective prioritization and management of user reports are crucial for maintaining user satisfaction, improving product quality, and ensuring overall organizational success. To build a machine learning model that can streamline the claims process, the data teams needs to inspect the data, organize it, and prepare it for exploratory data analysis phase.

Solution

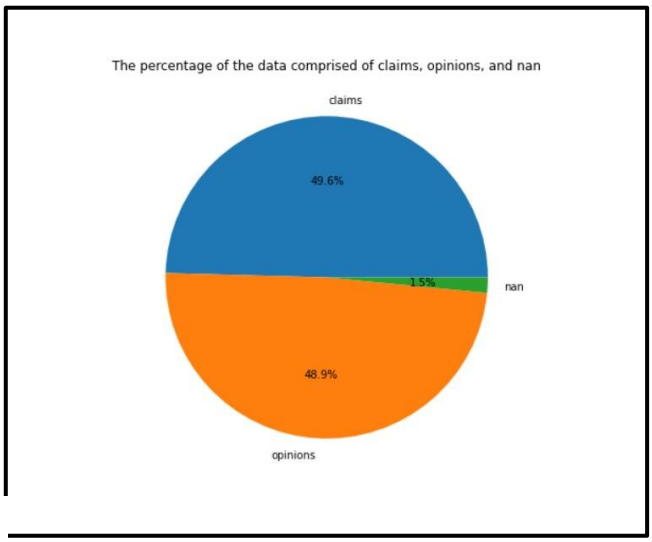
To obtain a model with the highest predictive power, the TikTok data team in the preliminary data analysis phase prepared the data needed for the claims classification project. To get clear insights, the data team framed the problem, . imported the data, built a dataframe, and not only examined the data type of each column but gathered descriptive statistics of the dataset as well in order to understand the data.

Details

- **Using the Python methods describe(), we learned about the dataset.** We examine the `claim_status` and `author_ban_status` variables and determine how many videos there are for each combination of categories of claim status and author ban status.
- **Filtering the data according to claim status and calculating the mean and median view counts for each claim status,** we see that although the mean and median for each different claim status are similar to each other there is huge discrepancy between the view counts of claims and opinions. The following screenshot indicates this:

```
average_view_count [claim]: 501029.4527477102 median_view_count [claim]: 501555.0
average_view_count [opinion]: 4956.43224989447 median_view_count [opinion]: 4953.0
```

- **There appears to be an equal distribution between the view counts of claims and opinions** as indicative in the pie chart.



Next Steps

- **The data team recommends performing further data cleaning and data analysis steps to understand unusual variables and outliers.** We need to decide whether to remove or cap outliers and whether to impute or drop rows with missing values based on their impact on analysis.
- **The data team will use descriptive statistics to gain a better grasp of the data and conduct a complete exploratory data analysis**. We need to dive deeper into specific variables or subsets of the data.