Machine Learning Claims Classification Executive Summary

Milestone 2

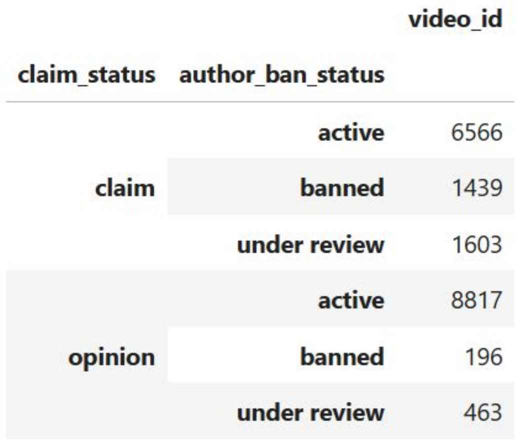
OVERVIEW

The purpose of this project is to classify TikTok videos as “claims” or “opinion” using a machine learning model. This requires data to be organized and summarized to prepare it for exploratory data analysis (EDA).
The data team summarized the characteristics of the data as well as identified potential relationships between variables.

PROJECT STATUS

In this phase of the project, 4 main tasks were completed.

1. Imported and reviewed the data
2. Summarized the data (types, null counts, average values, etc.)
3. Investigated relationships between variables
4. Created meaningful variables for engagement trends (likes per view, shares per view, comments per view)



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

Figure 2: Claim Status and Author Ban Status Breakdown

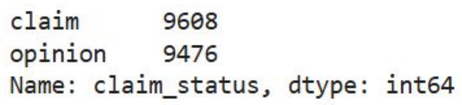
NEXT STEPS

The next step is to conduct EDA to further understand the data.

KEY INSIGHTS

Data breakdown:

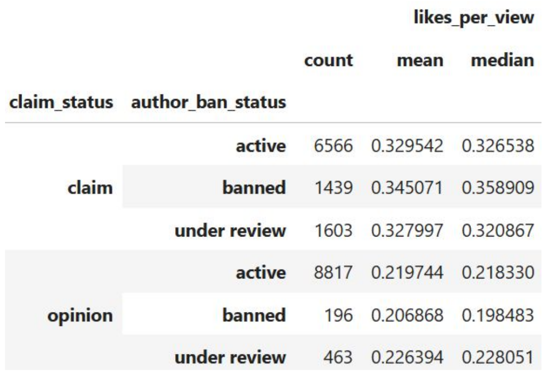
- Claim status: Almost equal about of claim and opinion videos (Figure 1)



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

Figure 1: Claim Status Counts

- Author ban status: claim status videos had a significantly higher banned and under review count than opinion-based videos (Figure 2)
- Engagement trend variables are more affected by claim status than author ban status, shown by the means and medians varying drastically between claim statuses but not author ban statuses (Figure 3)



		likes_per_view		
		count	mean	median
claim_status	author_ban_status			
claim	active	6566	0.329542	0.326538
	banned	1439	0.345071	0.358909
	under review	1603	0.327997	0.320867
opinion	active	8817	0.219744	0.218330
	banned	196	0.206868	0.198483
	under review	463	0.226394	0.228051

Figure 3: Engagement Trends