TikTok Video Classification Project

Task 1: Project proposal

Overview

Objective:

The primary objective of this project is to develop a machine learning model capable of reliably classifying user interaction data on TikTok as either a claim or an opinion. This model will help streamline the moderation process by efficiently prioritizing user reports, thereby reducing the backlog and enhancing the speed and accuracy of content review.

Impact:

By successfully implementing this predictive model, TikTok will significantly improve its content moderation efficiency. This will lead to faster resolution of user reports, better user experience, and an overall enhancement in platform safety and trustworthiness. Additionally, the project will enable TikTok to handle the increasing volume of user interactions more effectively, supporting the platform's growth and user satisfaction.

Primary audience: TikTok data team and cross-functional team members.

Milestone	Tasks	Deliverables/Reports	Relevant Stakeholder	Deadlines
1	Establish structure for project workflow (PACE) Plan	Global-level project document	Rosie Mae Bradshaw — DS Manager	1-2 days
1a	Write a project proposal		Mary Joanna Rodgers — PM Officer	

	Plan			
2	Compile summary information about the data Analyze	Data files ready for EDA	Orion Rainier — DS	1-2 weeks
2a	Begin exploring the data Analyze		Orion Rainier — DS	
3	Data exploration and cleaning Plan **and** Analyze	EDA report	Orion Rainier — DS	3-4 days
3a	Visualization building Construct **and** Analyze	Tableau dashboard/visualizations	Orion Rainier — DS	
4	Compute descriptive statistics	Analysis of testing results between two important variables	Willow Jaffey — DS Lead	4-5 days

	Analyze			
4a	Conduct hypothesis testing Analyze **and** Construct		Orion Rainier — DS	
5	Build a regression model Analyze **and** Construct		Willow Jaffey — DS Lead	2-3 days
5a	Evaluate the model Execute	Determine the success of the model	Rosie Mae Bradshaw — DS Manager	
6	Build a machine learning model Construct	Final model	Willow Jaffey — DS Lead	1-2 weeks
6a	Communicate final insights with stakeholders Execute	Report to all stakeholders		

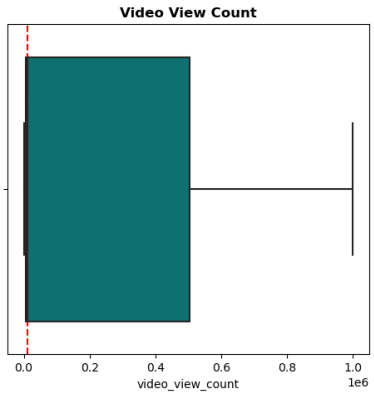
Task 2: Data preparation

Overview

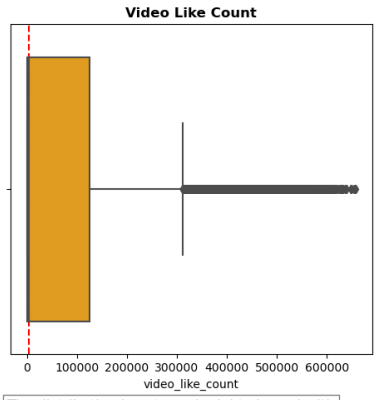
The TikTok data team seeks to develop a machine learning model to assist in the classification of claims for user submissions. To begin, the data team needs to organize the raw dataset and prepare it for future exploratory data analysis. In this preliminary investigation, I aim to summarize column Data types, data value nonnull counts, relevant and irrelevant columns and understand the data distribution.

Understand the Data

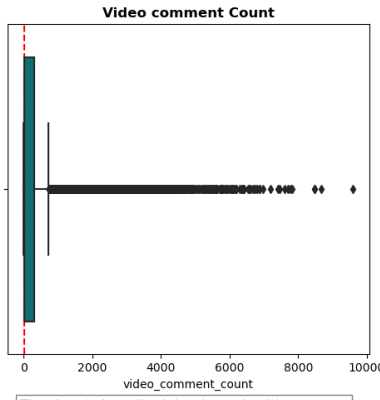
- The provided dataset is in long format, with **19382 rows and 12 columns**
- Most of the numeric columns are in **integer or float datatypes**; however, claim_status, verified_status, and author_ban_status are currently in object types and they should be boolean masked in order to build our predictive model.
- The dataset contains **298** rows with null values. I consider to drop these as the null values appeared simultaneously to many key columns and our sample size is large enough to yield an accurate result. However, further investigation is necessary to find the reasons for these nulls.
- There are **no duplicated values** found in the dataset.
- The counts of each claim status are quite balanced. There are **9,608 claims and 9,476 opinions**.
- View, like, comment, share count are heavily right skewed, suggesting there are some significant outlier videos with significant audience engagement. Therefore, the data distribution is **right skewed**, which will inform the types of models that we will build.



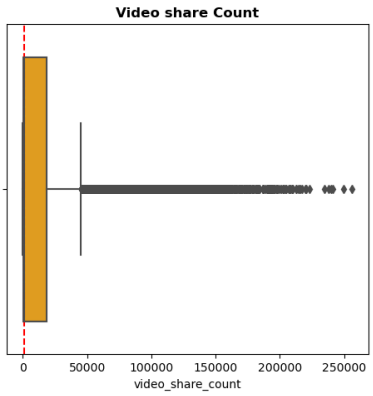
Distribution: The data is highly right-skewed with a concentration of values at the lower end. (The red line is median)



The distribution is extremely right-skewed with most values clustered near the lower end. There are significant outliers, indicating a few videos with exceptionally high like counts.



The data is heavily right-skewed, with most values near the lower end. Many outliers are present, showing a few videos with very high comment counts.



The distribution is highly right-skewed, similar to the others. There are several outliers indicating videos with a high number of shares.

Preliminary findings

This section lays the foundation understanding of the relationship between variables.

1. Claim videos have more views: By grouping video claim status on their mean and median video view count, I discovered that the mean and median number of views for 'claim' is higher than 'opinion'. In other words, people watch more videos identified as 'claim' rather than 'opinion'.

2. Variables like author ban status, verified status, and video view count are important for our predictive model

3. I created three extra columns to calculate proportion of like, comment, share per view. These new columns may help to further derive deeper insights.

like_%	share_%	comment_%
0.056584	0.000702	0.000000
0.549096	0.135111	0.004855
0.108282	0.003168	0.000365

The next step

In summary, I believe that video view count, author ban status, share count may be some key independent variables for our predictive model. However, we might discover more relevant variables as we progression to later stages of the project. With the key variables identified and the initial investigation of the claims classification dataset, the process of exploratory data analysis can begin.

Task 3: EDA process

Overview

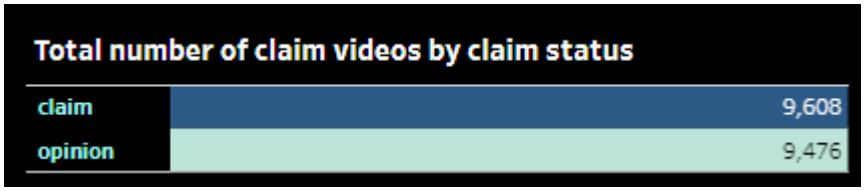
The data team seeks to develop a machine learning model to assist in the classification of claims for user submissions. In this part of the project, I need to analyze, explore, clean, structure prior to any model building.

The purpose of the exploratory data analysis (EDA) is to understand how different video engagement metrics vary based on video claim status and author ban status.

Tableau visualization: click here

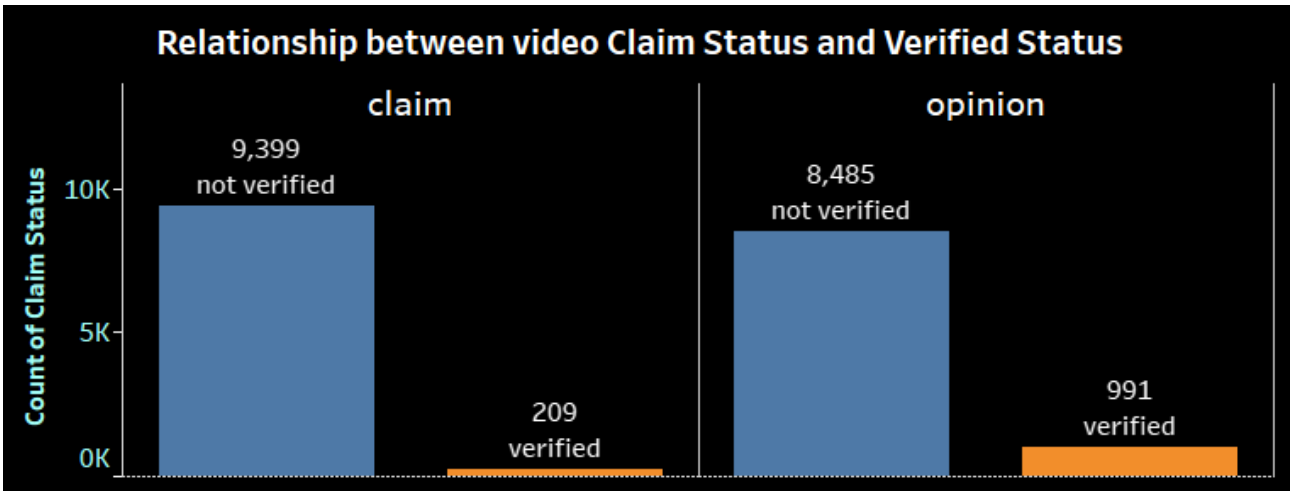
Key findings

1. **There are more claim related videos than opinion related videos**



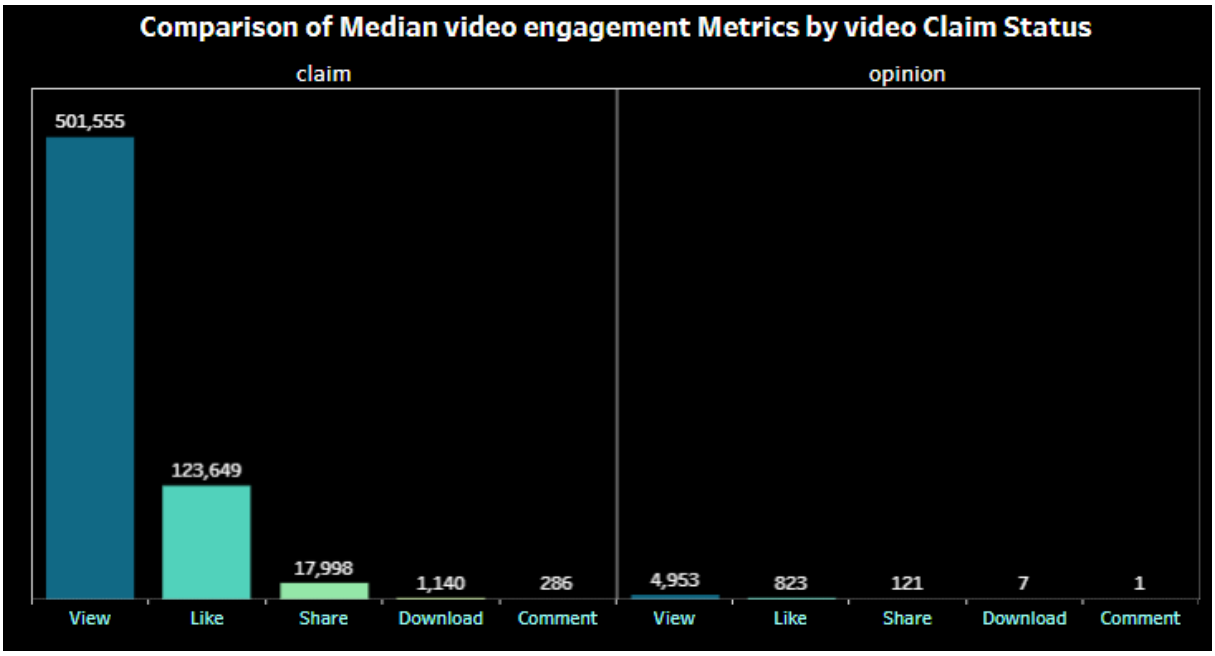
Impact: The future model will need to account for the imbalance in video counts as well as the null values.

2. **Claim related videos tend to have less verified authors than opinion related videos**



Impact: Most of the video authors not verified in both categories. However, claim related videos tend to have less verified authors than opinion related. This insight may help to determine our independent variables for the predictive model later on.

3. **Claim related videos have higher video engagement metrics than opinion related**

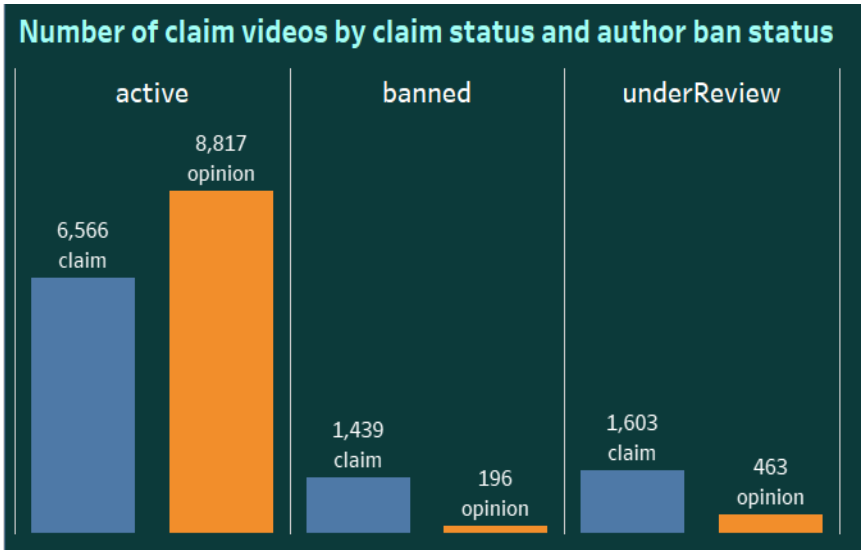


Impact: Claim related videos have higher median view, like, share, download, comment count than videos about author's opinion. This insight may accurately consider higher views, likes, shares, comments associate with claim related videos.

About the author ban status

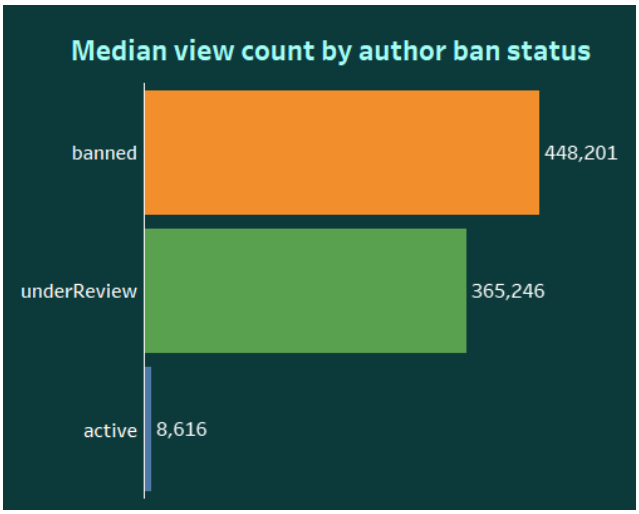
After investigating the dataset, I find author ban status is one of the key independent variables to determine the video classification. By comparing various metrics grouping on author ban status, I found these insights:

4. **There are more claim related videos authors are being banned or under review than opinion related authors**



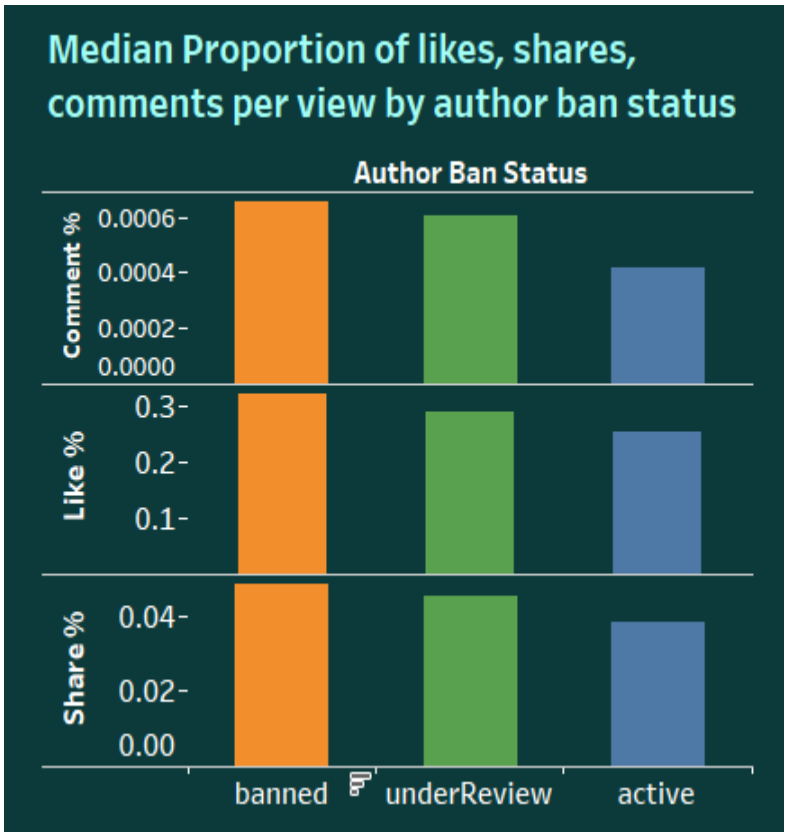
Impact: There are more claim related videos authors are being under review or banned than opinions authors. Surprisingly, there are more opinion related video authors are active than the claim category. The reasons I believe are because claim related videos can have consequences of user's real life. Claims related to health, safety, or other critical issues are often under stricter scrutiny because they can have real-world consequences if they mislead users.

5. **Banned and under review authors have higher median view counts**



Impact: We can see that banned and under review authors both gain a significant number of views than active authors. I believe Banned or under review authors might have produced more controversial or sensational content that attracts a higher number of views. Controversial content tends to garner more attention, leading to higher view counts.

6. **Banned and under review authors have higher median proportion of likes, shares, comments per view**



Impact: Contents from banned and under review authors have higher users engagement than active authors. The rationales I believe are authors who are banned or under review may discuss polarizing topics that tend to divide opinions. This polarization can drive higher engagement as users like and share the content to express their views. This insight lays the foundation for content analysis, themes of the content posted by banned and under review authors. Identify any common traits that might explain the higher engagement rates.

Task 4: Statistical model

Overview

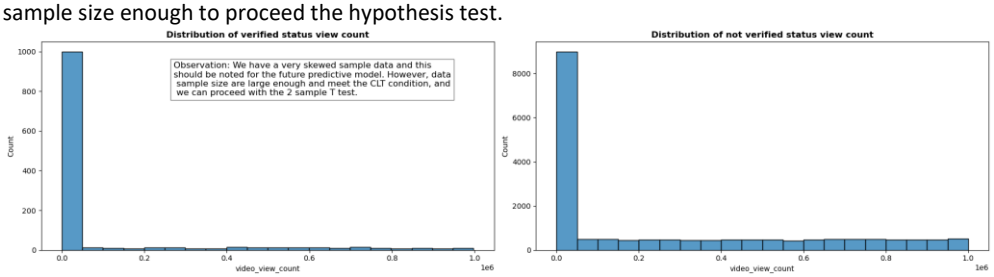
The TikTok data team seeks to develop a machine learning model for content classification. In this part of the project, the clients are interested in whether there is a difference between verified and not verified accounts. To tackle this, I conduct a 2 sample T hypothesis test to verify whether not verified accounts have more mean view count than verified accounts.

Key Insights

- The hypothesis test results reveal that not verified accounts have much higher mean views than verified accounts.
- The results are statistically significant (P value < 0.05)
- I believe there are might be fundamental behavioral differences between the two groups and it would be interesting to investigate the roots of the differences:
  1. What make unverified videos more engaging than verified ones?
  2. Why verified accounts have lower views?

Details

- Distribution of views for verified and not verified accounts are highly skewed, yet we have large sample size enough to proceed the hypothesis test.



- Not verified accounts have a mean view of **265.6k**
  Verified accounts have a mean view of **91.4k**

- Null (HO): Mean view count (not verified) <= Mean view count (verified)
- Alternative (H1): Mean view count (not verified) > Mean view count (verified)

- 2 sample Welch's T test results: T statistics (25.49) > T critical value (1.65), P value <0.05

- Conclusion: reject the null hypothesis (Ho)

The next steps

The team can move forward and build a predictive regression model on verified status. A technical note is that data sample is highly skewed and should be considered to choose appropriate models.

Task 5: Logistics regression model

Overview

The TikTok data team seeks to develop a machine learning model to assist in the classification of claims for user submissions. Earlier, the data team observed that if a user is verified, they are much more likely to post opinions. Since the end goal is to classify claims and opinions, it's important to build a model that shows how to predict the behavior of the account type (verified) that tend to post more opinions. So, in this part of the project, the data team built a logistic regression model that predicts verified_status (verified or not verified.)

Key Insights

- **Accuracy (65%):** the model correctly classified 65% of all predictions, whether "verified" or "not verified".
- **Precision (61%):** the model is able to correctly predict 61% of the cases labeled as 'not verified'.
- **Recall (84%):** the model successfully identified 84% of the actual 'not verified' cases.
- **F1-Score (71%):** indicating better performance in detecting not verified accounts.

What this means:

-The model is better at predicting "not verified" outcomes than it is at predicting "verified" ones.

-We might need to refine the model further if we want to improve the accuracy for "verified" predictions, depending on the importance of these classifications.

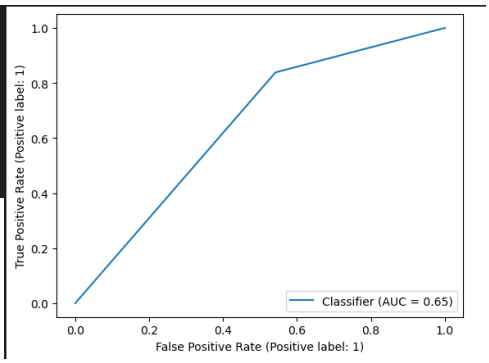
Details

Evaluation Metrics report

	precision	recall	f1-score	support
verified	0.74	0.46	0.57	4459
not verified	0.61	0.84	0.71	4483
accuracy			0.65	8942
macro avg	0.67	0.65	0.64	8942
weighted avg	0.67	0.65	0.64	8942

*The model is better than random guessing but still has room for improvement in terms of distinguishing between "verified" and "not verified" cases. The AUC score of 0.65 shows it's somewhat effective but not perfect.

ROC curve



Coefficients

video_duration_sec	0.00860789
video_view_count	-0.00000213
video_share_count	0.00000593
video_download_count	-0.00001100
video_comment_count	-0.00064042
('claim_status_opinion',)	0.00039084
('author_ban_status_banned',)	-0.00001782
('author_ban_status_under review',)	-0.00000097

*Based on the logistic regression model, video duration has the highest coefficient. Each additional second of the video is associated with 0.009 increase in the log-odds of the user having a verified status. In other words, if we exponentiate e^0.009 = 1.009, then for each additional second of video duration, the odds of being verified increase by about 0.9%.

The next steps

The next step is to construct a classification model that will predict the status of claims made by users. That is the final project and original expectation from the TikTok team. Now, there is enough information to analyze the results of that model with helpful context around user behavior.

Task 6: Final model

Overview

TikTok wants to build a machine learning model to help identify claims and opinions. Videos that are labeled opinions will be less likely to go on to be reviewed by a human moderator. Videos that are labeled as claims will be further sorted by a downstream process to determine whether they should get prioritized for review. For example, perhaps videos that are classified as claims would then be ranked by how many times they were reported, then the top x% would be reviewed by a human each day.

A machine learning model would greatly assist in the effort to present human moderators with videos that are most likely to be in violation of TikTok's terms of service.

Key Insights

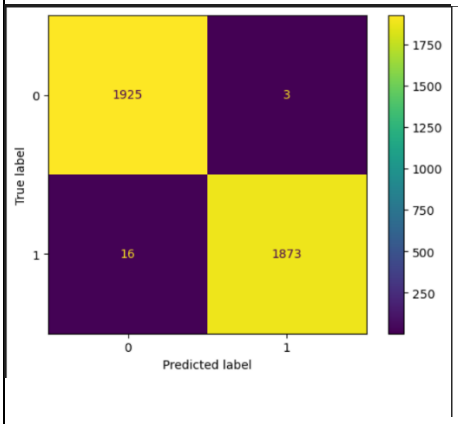
- **Accuracy (99.5%):** the model correctly classified 99.5% of all predictions, whether "claim" or "opinion".
- **Precision (99.84%):** the model is able to correctly predict 99.84% of the cases labeled as 'claim'.
- **Recall (99.15%):** the model successfully identified 99.15% of the actual 'claim' cases.

What this means:

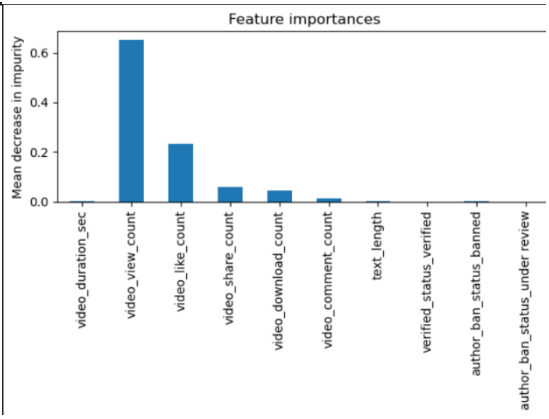
Yes, one can recommend this model because it performed well on both the validation and test holdout data with consistent high metrics.

Details

Confusion Matrix



Feature importances



*I believe the **Random Forest** is the most suitable for our predictive model. One can use this model to predict whether the video is about a claim or opinion. The model very successfully classified claims and opinions with high Recall score. The model's most predictive features were all related to the user engagement levels associated with each video, specifically **views, likes, shares, and downloads.**

The next steps

As noted, the model performs exceptionally well on the data. Before deploying the model, the team considers further evaluation with more user data.