TikTok Claims Classification Project

Hypothesis Testing

Project Overview

The data team is developing a machine learning algorithm to precisely detect whether a video contains claims or opinions. After conducting Exploratory Data Analysis we performed hypothesis testing on all variables to determine which variables are statistically significant for future predicting model.

Key Insights

1. Following variables/features are statistically significant and will be used for further steps: views, likes, shares, downloads, comments, author ban status and author verified status

2. The video duration and the video transcription text are not used for further steps because they have no statistically significant difference between claims and opinions

3. For further steps the original data will be used and not the log-transformed data, because it will not be necessary anymore.

Next Steps:
The Data Team will build and compare XGBoost and Random Forest models, evaluating which model predicts the claim status best.

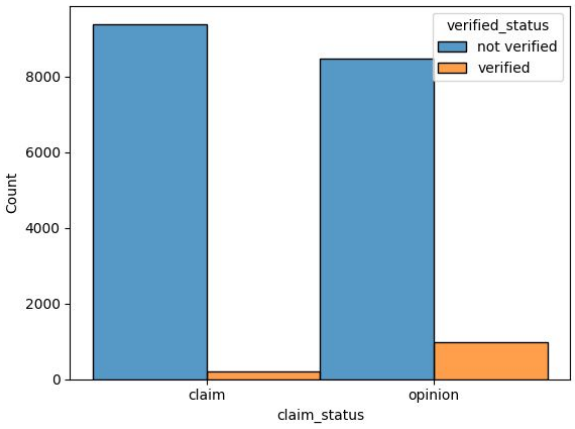
Details

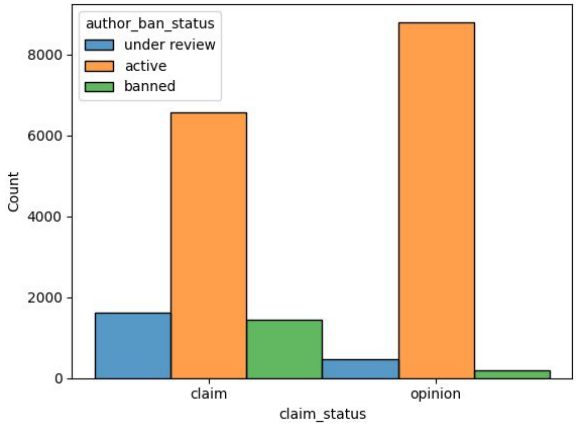
Rates have p-value close to 0

```
for rate in rates_list:
  p_list = []
  for i in range(1000):
    p_value = stats.ttest_ind(a=claim_data[rate].sample(30), b=opinion_data[rate].sample(30), equal_var=False).pvalue
    p_list.append(p_value)

  p_value = np.array(p_list).mean()
  decision = ("definitely statistically significant" if p_value * 1000 < significance_level
                                                     else "statistically significant" if p_value < significance_level else "by chance")
  print(f"The difference in {rate} between claims and opinions is {decision} with a p-value of {p_value}.")
```

```
The difference in video_duration_sec between claims and opinions is by chance with a p-value of 0.5035878414226542.
The difference in video_view_count_log between claims and opinions is definitely statistically significant with a p-value of 8.619851150615286e-18.
The difference in video_like_count_log between claims and opinions is definitely statistically significant with a p-value of 5.3139185797998696e-15.
The difference in video_share_count_log between claims and opinions is definitely statistically significant with a p-value of 8.628720335285769e-11.
The difference in video_download_count_log between claims and opinions is definitely statistically significant with a p-value of 7.533111223124776e-11.
The difference in video_comment_count_log between claims and opinions is definitely statistically significant with a p-value of 3.920553287654959e-11.
```





The difference in author verified/ban status between claims and opinions is statistically significant.