TikTok project: detect claims and opinions

Milestone #3: Hypothesis test

Overview

The objetive of this project is to implement a machine learning model to classify the videos uploaded to TikTok into claims (unsourced statements) and opinions (personal thoughts) in order to reduce moderating times and reduce economical costs.

Objective

The objective of this milestone was to investigate the differences between verified and not verified users (*verification_status*) in terms of video view counts (*video_view_count*). To do so, we derived a descriptive statistic (the mean), then stated the null and alternative hypothesis and finally performed an independent two-sample t-test.

Results

	mean_video_view_count	
	count	mean
verified_status		
not verified	17884	265663.785339
verified	1200	91439.164167

- First, we grouped the data by *verified_status* and calculated the count and mean values per group. The results indicate:
  - There are far more not verified (~18k) than verified users (1.2k).
  - Not verified users show more (mean) views (~265k) than verified (~91k).
- Then, we conducted a hypothesis test to evaluate the hypothesis suggested by the descriptive statistic (latter bullet point above). The hypothesis were:
  - H_0: $\mu_{verified} == \mu_{not\ verified}$.
  - H_A: $\mu_{verified} \sim= \mu_{not\ verified}$.
- We found a statistically significant difference between the two distributions (p-value < 0.05). This implies that there is a difference in the population means of the two groups: not verified users generate more views than verified ones.
- Further analyses need to investigate the reasons behind this:
  - Do not verified users use spam bots to improve their engagement?
  - Do not verified users click-bait videos?

Next Steps

The next step in the project could be the implementation of a logistic regression analysis to further investigate the behavior of the different types of verification users in terms of engagement. For the rest of the analyses, we need to keep in mind the distribution of the engagement variables are skewed to the right and there is a statistically significant difference between the two user groups in terms of engagement.