TikTok Claims Classification Project

Exploratory Data Analysis (EDA)

Project Overview

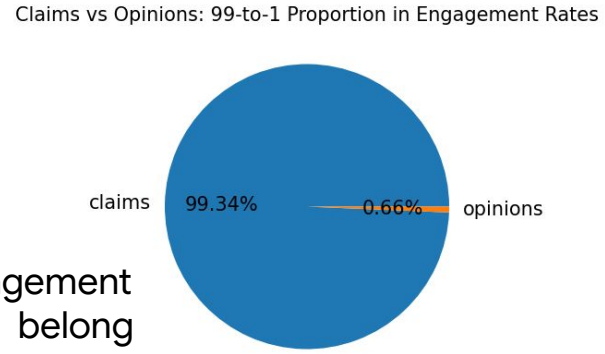
The data team is developing a machine learning algorithm to precisely detect whether a video contains claims or opinions. Therefore we conducted an Exploratory Data Analysis to gain further knowledge about the data. It is an essential step before building predictive models.

Key Insights

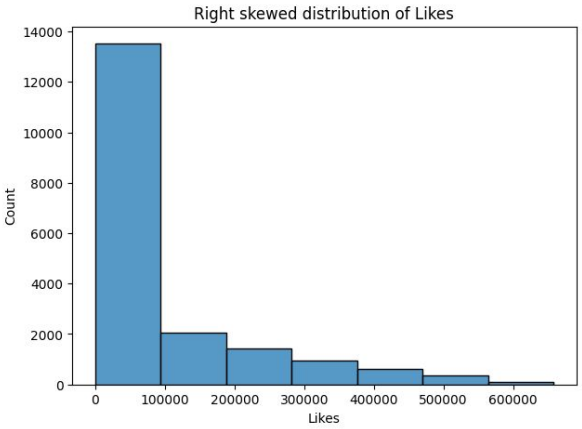
1. In the dataset, 2% of all reports are not yet finished and have been filtered out for a clearer analysis .
2. 99% of total engagement rates come from videos containing claims.
3. Claims have engagement rates that are 100 to 250 times higher than opinions. This is why the distribution of engagement rates skews towards higher values.
4. Outliers should be determined after seperating claims from opinions because all considered outliers would be claims and no opinions. That would infiltrate bias for further analysis and modeling.

Next Steps:
The Data Team will conduct hypothesis testing and statistical analysis to find the variables that are important for model building.

Details



99% of all engagement rates like views belong to videos with claims.



More than 70% of the videos have <100k Likes.

Claims have 100 to 250 times higher engagement rates than opinions.