Claims Classification Project | Milestone-2-Pre-Processing Results

Executive Summary Prepared for the Tiktok Leadership and Team By the Tiktok Data Team

Overview

TikTok leadership has tasked the data team with developing a machine learning model to classify user content as either a claim or an opinion. This initiative aims to reduce the backlog of user reports and enhance the platform's ability to filter misinformation. This executive summary outlines the initial data preprocessing conducted by Saswat Seth, laying the groundwork for future exploratory data analysis (EDA) and the subsequent machine learning model development.

Objective

The increasing volume of user reports against TikTok content has led to significant backlogs, which poses a challenge for content moderation. The objective of this project is to efficiently classify user submissions as either claims or opinions, reducing the backlog and ensuring the content complies with platform policies. In this initial phase, the data team focused on preprocessing the dataset to ensure that it is ready for deeper analysis and machine learning model development.

Results

The dataset consists of 12 columns with key observations:

- **Claim Status**: The `claim_status` column shows a near-equal distribution of claims (9,608) and opinions (9,476), indicating balanced categories.
- **Outliers**: Significant differences between the 75th percentile and maximum values suggest potential outliers, which will need attention during EDA.
- **Null Values**: The `claim_status` column has 298 null values, consistent across related attributes, requiring further handling.
- **Banned Users**: More banned users are associated with the claim category, indicating that users posting claims are more likely to be flagged for policy violations.
- **Engagement**: Banned users generate higher engagement per post (mean and median) despite fewer total views, likes, and shares, suggesting controversial claims attract more attention.

Next Steps

The next phase will involve conducting exploratory data analysis (EDA) on the organized and pre-processed dataset. This will help uncover deeper insights into the data and identify important features for the machine learning model. Additionally, the data will be further processed to handle potential outliers and null values, ensuring a clean dataset for model development.