Behavior Simulation and Content Simulation

Yash Kumar, Vivek Yadav, Vishal Das
Department of Computer Science and Technology
Delhi Technological University, Delhi, India
{yashkumar_co21a8_34, vivekyadav_co21a8_30, vishaldas_co21a8_24}@dtu.ac.in

Abstract

This document provides an overview of the data science challenge on "Behavior Simulation and Content Simulation." The challenge focuses on estimating user engagement on social media content and creating content that aligns with key performance indicators (KPIs).

1 Introduction

The goal of the challenge is to address the complexities of user behavior and engagement on social media platforms, particularly Twitter. The dataset comprises sampled tweets from enterprise accounts over the past five years, including metrics such as likes, retweets, and comments.

2 Problem Description

The problem statement outlines the communication process and user engagement on social media. Our task is to analyze the provided dataset and develop models that can predict user engagement based on various features.

3 Phase 1: Exploratory Data Analysis

During this phase, we will clean and preprocess the dataset. Exploratory Data Analysis (EDA) techniques will be applied to understand data characteristics, identify outliers, and generate insights for subsequent modeling. Cleaning methods will include text normalization, punctuation removal, and other necessary steps.

3.1 Data Cleaning

We will normalize text data, remove unwanted characters, and prepare the dataset for further analysis. We will also be extracting hashtags present in the tweets and use it for fruther analysis.

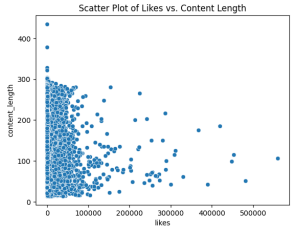

Figure 1: Like vs Content Length Scatterplot

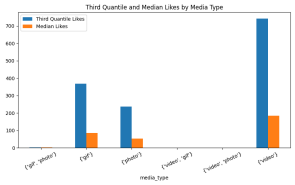

Figure 2: Enter Likes by media type

3.2 Exploratory Data Analysis

The analysis of the graph in Figure 1 indicates that the content_length alone does not serve as a reliable predictor for likes. This is evident from the substantial variance in likes observed at content_length = 0. However, the presence of other favorable factors can enhance the predictability. In such cases, having a content_length around the median value tends to be beneficial.

From the analysis presented in Figure 2, it is evident that the 'video' media type garners approximately twice the number of likes compared to the 'photo' media type. This suggests a higher level of user engagement with video content on the platform.

An analysis of bar graph 3 reveals a discernible correlation between video duration and their corresponding median likes. Most bars are of moderate height, indicating a moderate number of median likes for those respective video durations. However, there is a noticeable spike in the last bar on the right, indicating a significantly higher number of median

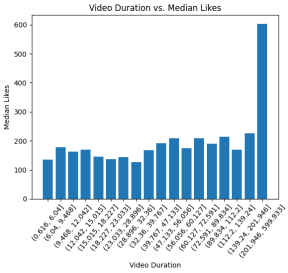

Figure 3: Video duration and median likes

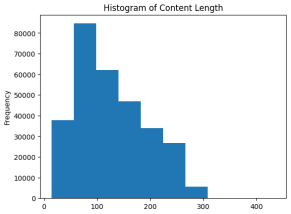

Figure 4: Frequency vs Content-Length

likes for videos within that particular duration segment. This pattern suggests a potential relationship between the two variables.

In bar graph in figure 5, our analysis underscores a significant correlation between the popularity and contemporaneity of hashtags and the median likes received by tweets incorporating those hashtags. The dataset, comprising hashtags utilized by at least four distinct brands in their tweets, reveals a discernible pattern: an augmentation in median likes is concomitant with the increased prevalence of the hashtag. Furthermore, temporal relevance emerges as a pivotal factor; hashtags echoing ongoing events or prevailing trends tend to amass a higher number of likes. This observation is instrumental in delineating the dynamics of user engagement on social media platforms and can be integral in strategising content dissemination to optimize user interaction.

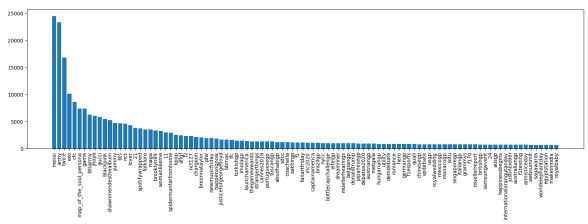

Figure 5: Median Likes vs Popular Hashtags