# IML Project

## Explore page and discoverability

# Table of Content


1 Problem statement

2 Introduction

3 Instagram’s explore page?

4 Data preprocessing

5 EDA Analysis

6 Building the Model

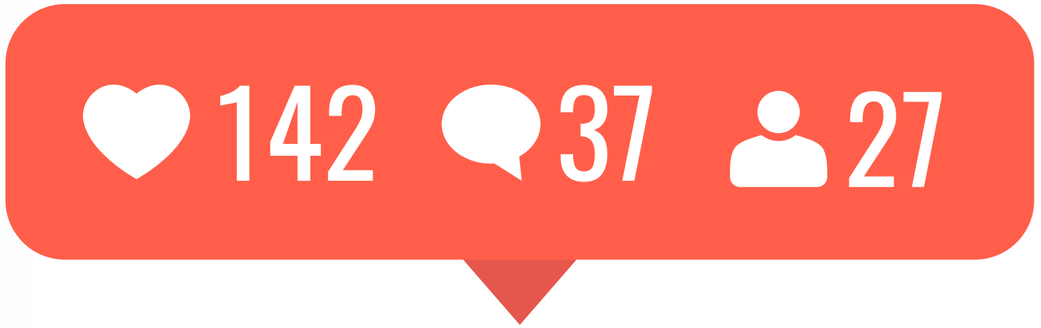
7 Limitations

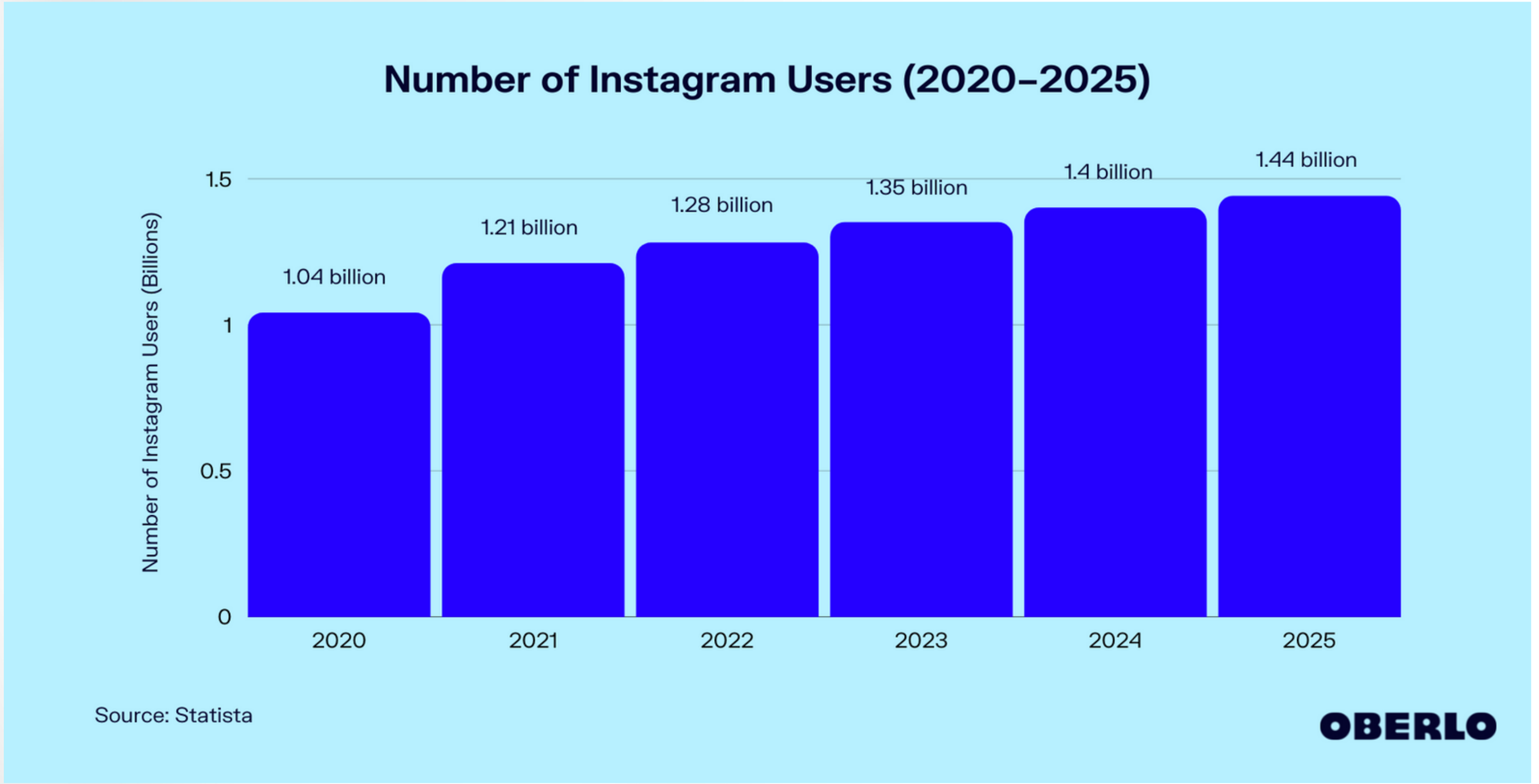
8 Our team

9 Endnotes

# Problem Statement

1. **Study the factors that influence content appearing on users' Explore pages.**
2. **Analyze the impact of Instagram's algorithms on content discoverability.**






# Introduction

Analyzing user behavior's on websites is essential to know your audience better, to create relevant content, to time your posts, use effective marketing strategies and maximize the impact of your advertising campaign.

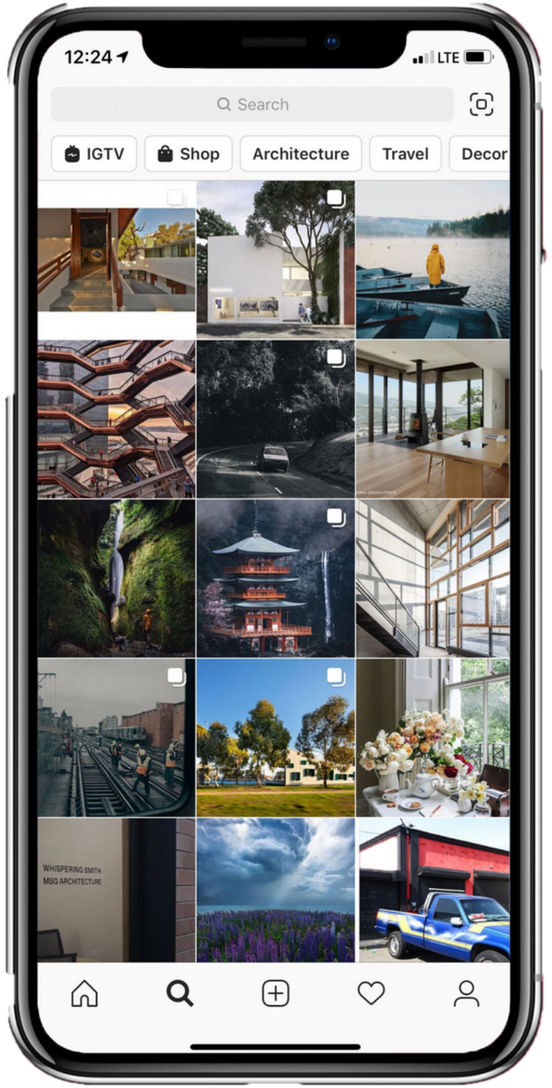
# Instagram’s explore page

Instagram's Explore page is where users discover content. It offers:

Benefits for Users:
- Personalized content recommendations.
- Content variety and inspiration.
- Discovery of new accounts and interests.

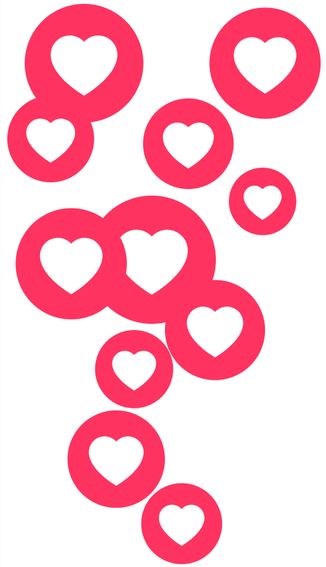
Benefits for Content Creators:
- Exposure to a wider audience.
- Increased engagement.
- Growth in followers and recognition.

In summary, Explore enriches the Instagram experience for users and provides a platform for content creators to reach a broader audience.



# Factors affecting content discoverability

Some of the factors that affect content discoverability are- Comments, Emojis used, Hashtags count, previous content history and many more.

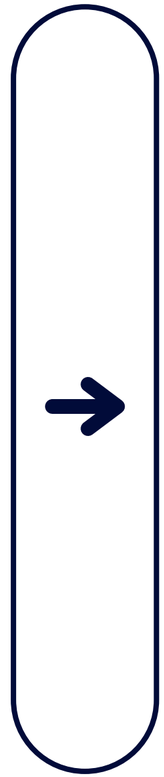
# Dataset

- For this project, we used the Kaggle notebook datasheet provided along with the problem statement for our analysis.
- The datasheet consists of 7488 rows and 9 columns ( the important ones being user id, photo id, emoji, hashtags used count and comments.

| | id | comment | User id | Photo id | created Timestamp | posted date | emoji used | Hashtags used count |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | unde at dolorem | 2 | 1 | 13-04-2023 08:04 | Apr-14 | yes | 1 |
| 1 | 2 | quae ea ducimus | 3 | 1 | 13-04-2023 08:04 | Apr-14 | no | 2 |
| 2 | 3 | alias a voluptatum | 5 | 1 | 13-04-2023 08:04 | Apr-14 | no | 4 |
| 3 | 4 | facere suscipit sunt | 14 | 1 | 13-04-2023 08:04 | Apr-14 | yes | 2 |
| 4 | 5 | totam eligendi quaerat | 17 | 1 | 13-04-2023 08:04 | Apr-14 | yes | 1 |
| 5 | 6 | vitae quia aliquam | 21 | 1 | 13-04-2023 08:04 | Apr-14 | no | 2 |
| 6 | 7 | exercitationem occaeca | 24 | 1 | 13-04-2023 08:04 | Apr-14 | yes | 0 |
| 7 | 8 | sint ad fugiat | 31 | 1 | 13-04-2023 08:04 | Apr-14 | no | 5 |
| 8 | 9 | nesciunt aut nesciunt | 36 | 1 | 13-04-2023 08:04 | Apr-14 | no | 6 |
| 9 | 10 | laudantium ut nostrum | 41 | 1 | 13-04-2023 08:04 | Apr-14 | yes | 2 |

# Data preprocessing

- We checked for any null value by using the function data.info() and found out that all values are non null.
- Emoji used (yes or no) has been converted to binary values ( 1 and 0).
- Comment length has been calculated by using data['comment'].apply(len)

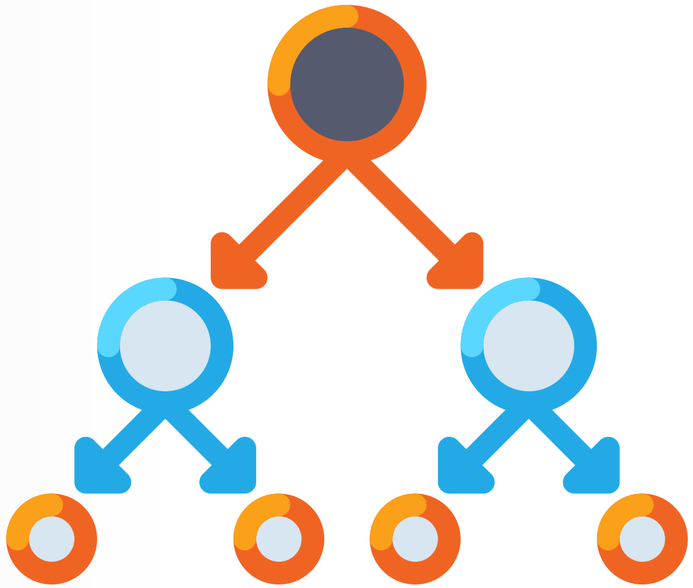


# Analyze impact of instagram’s algorithms on content discoverability.

Instagram's algorithms, governing content in Reels and Explore pages, consider diverse factors. User engagement, such as liking specific posts or reels, influences content recommendations. Building models based on features like emoji usage or hashtag count in comments can enhance algorithmic predictions.

# Model selection

We have used RandomForest Regressor for building our model for the following reasons.



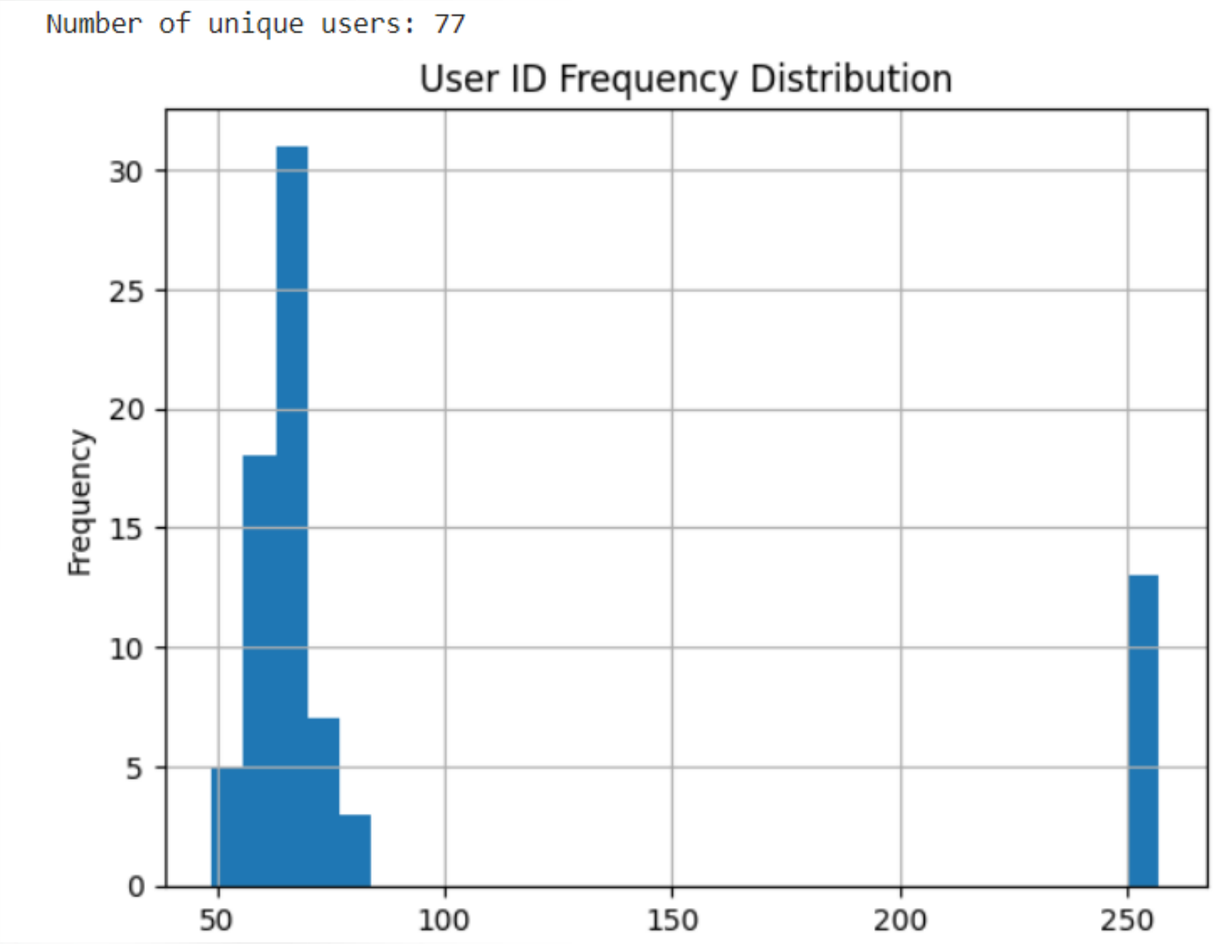
- Works well for comples datasets with non linear relationships.
- Is resistant to outliers
- Consists of many decision trees and the final output considered based on Majority Voting or Averaging for Classification and regression, respectively.

# EDA Analysis

Most of the user id’s lie in 50 to 100 range and only few lies in range of more than 250 whereas none in between 100 to 250 range.

- Content diversity- More unique User IDs often correlate with a diverse range of content.

- Understanding the unique User IDs allows platforms to customize explore page recommendations for individual users.



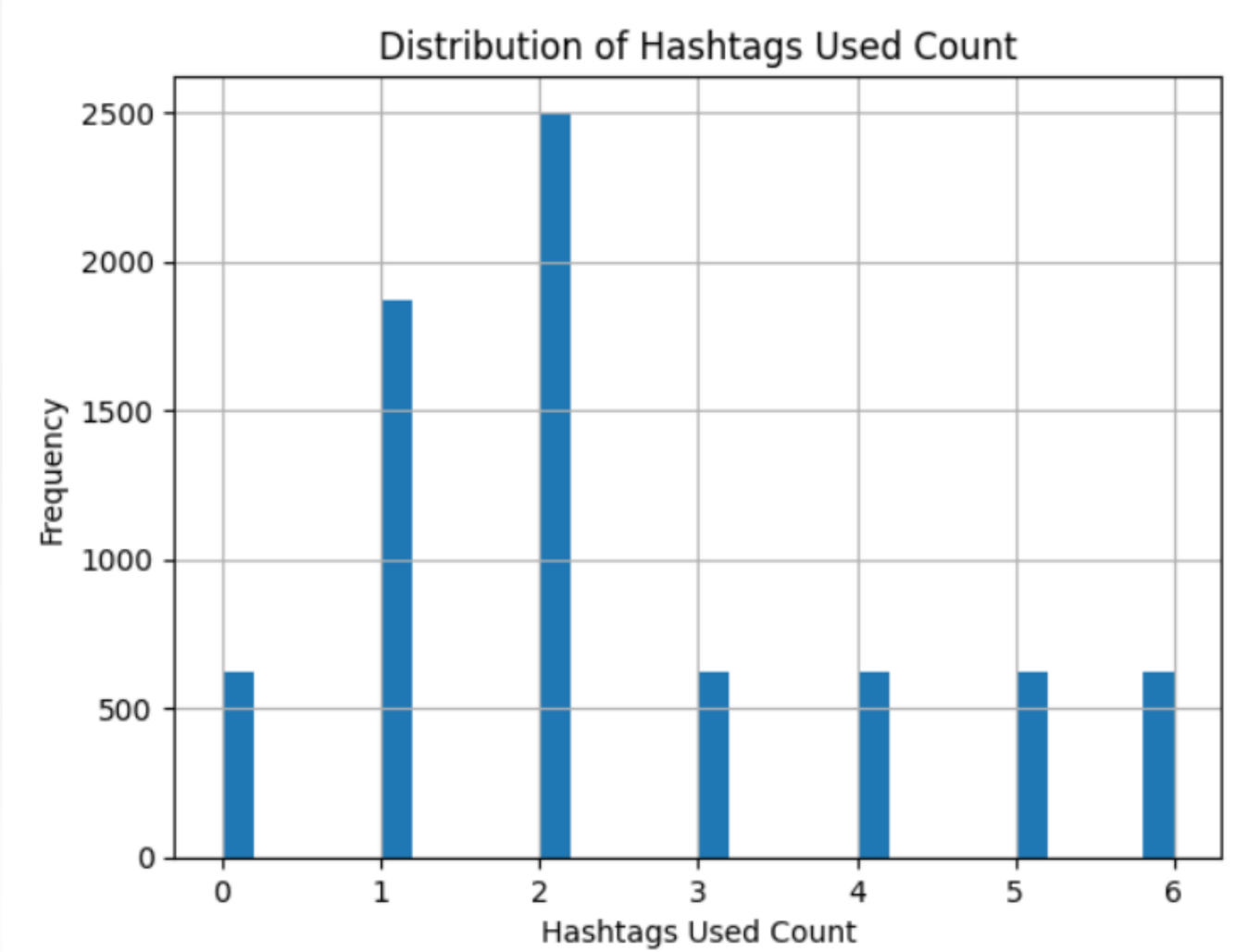
# No of unique photo id’s

Popular and engaging content is crucial for content discoverability. Content with a high number of unique Photo IDs may be featured in trending or popular sections of the platform, contributing to its visibility.

```
Number of unique photos: 257
13     39
157    39
247    39
8      38
146    37
       ..
144    23
199    23
16     22
230    22
179    21
Name: Photo id, Length: 257, dtype: int64
```

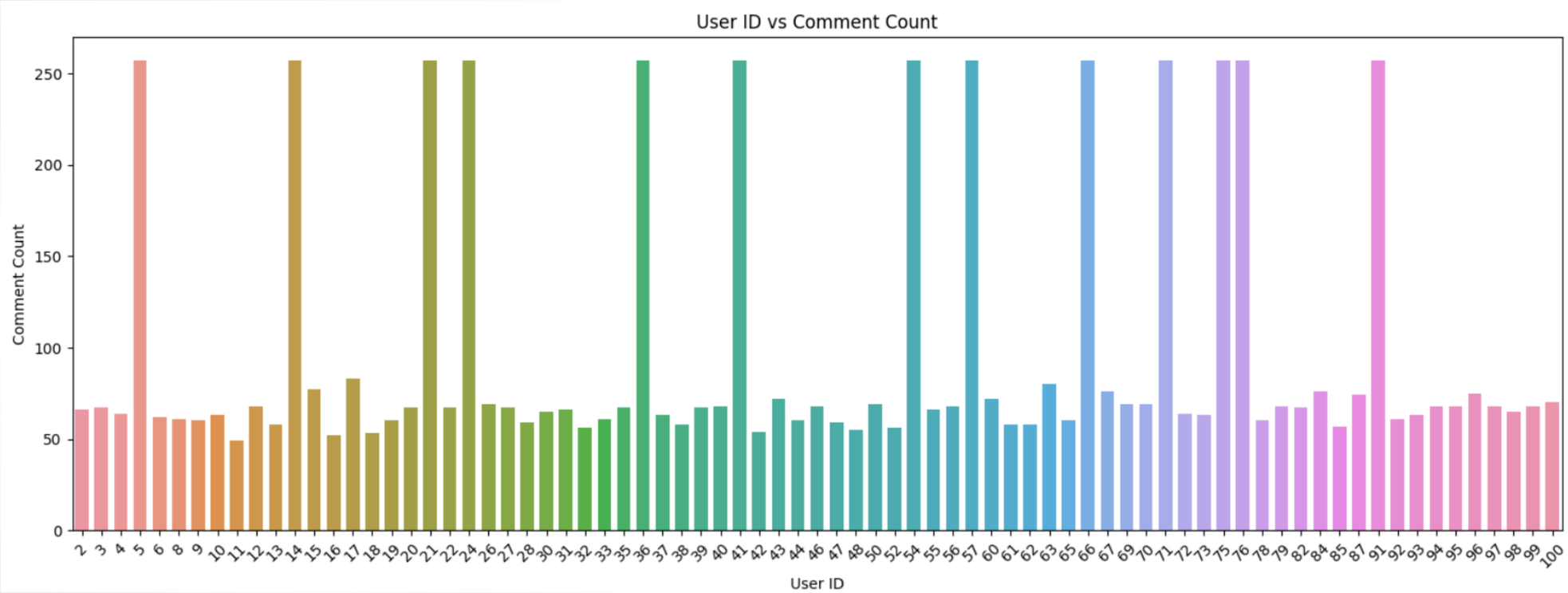
# Hashtags

- We can draw the insights from the graph that most probable no of hashtags that will be used with a comment will either be 2 or it will be 1.
- Categorization: Users can search or follow specific hashtags to explore related content.
- If users engage more with content featuring specific hashtags, the algorithm may prioritize such content on explore pages



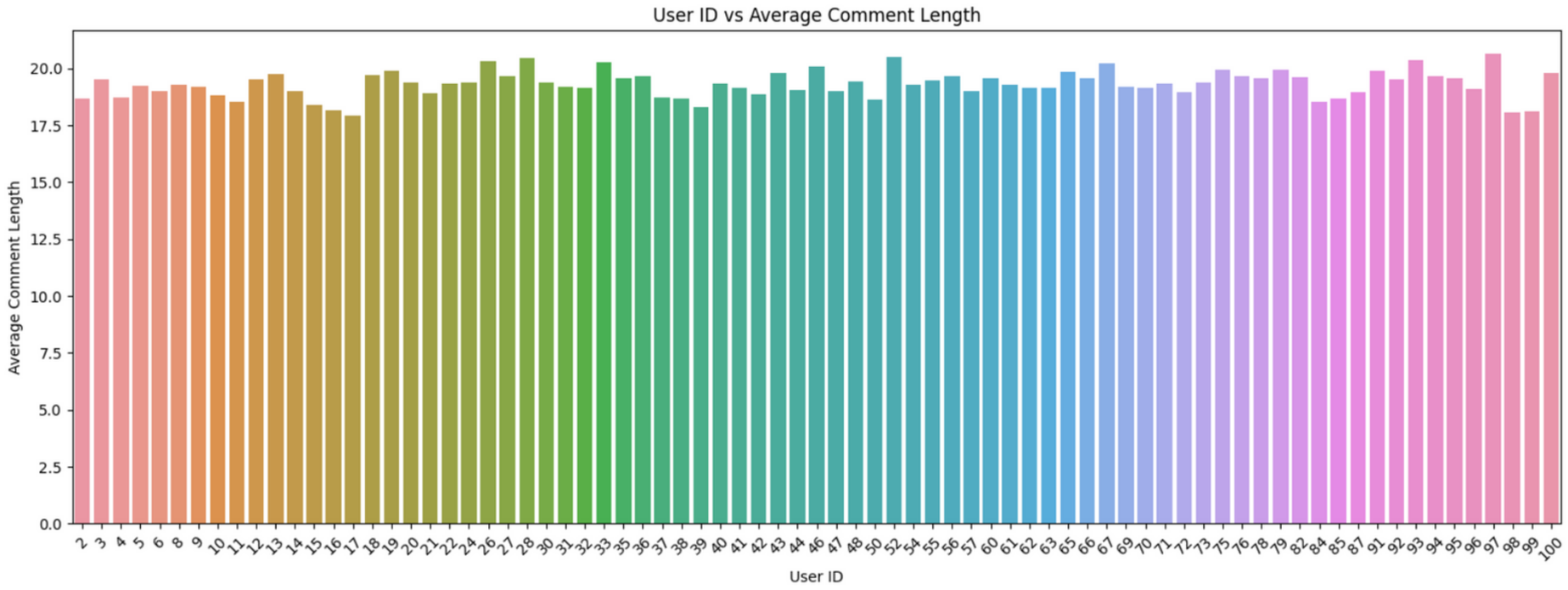
# User id vs comment count

- From the graph, we draw the conclusion that there is stark contrast between no of comments per user. While most of the users have comparable comments, some users like 5,14 comment a lot.



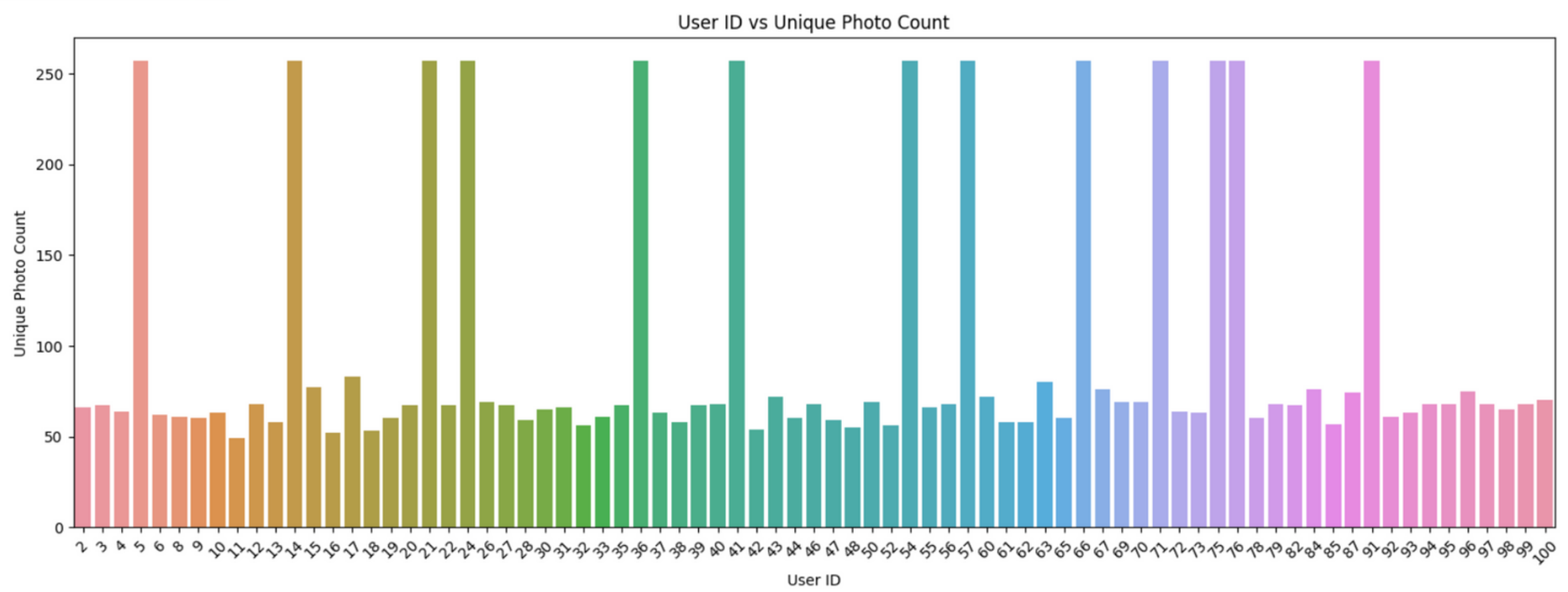
# User id vs Avg comment  length

- Users who consistently provide longer comments may contribute more substantial content to the platform
- Content Depth: Longer comments may indicate a more in-depth discussion or expression of thoughts.



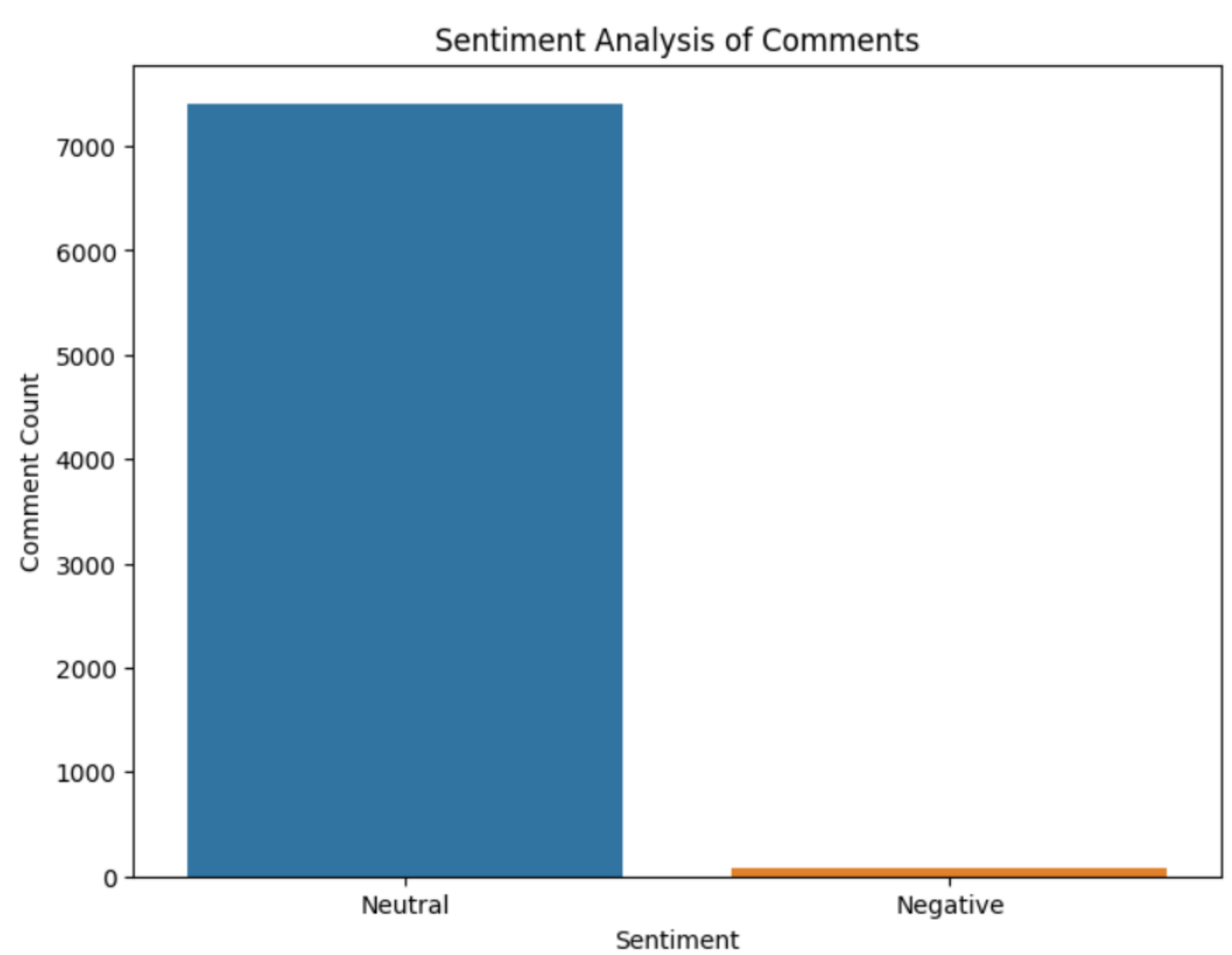
# User id vs Unique Photo id count

- The graph shows us that almost all users have comparable unique photo id’s
- Some users like 5,14 have large no of unique photo id’s. This means either they have created various accounts for different purposes or most of their accounts are junk.



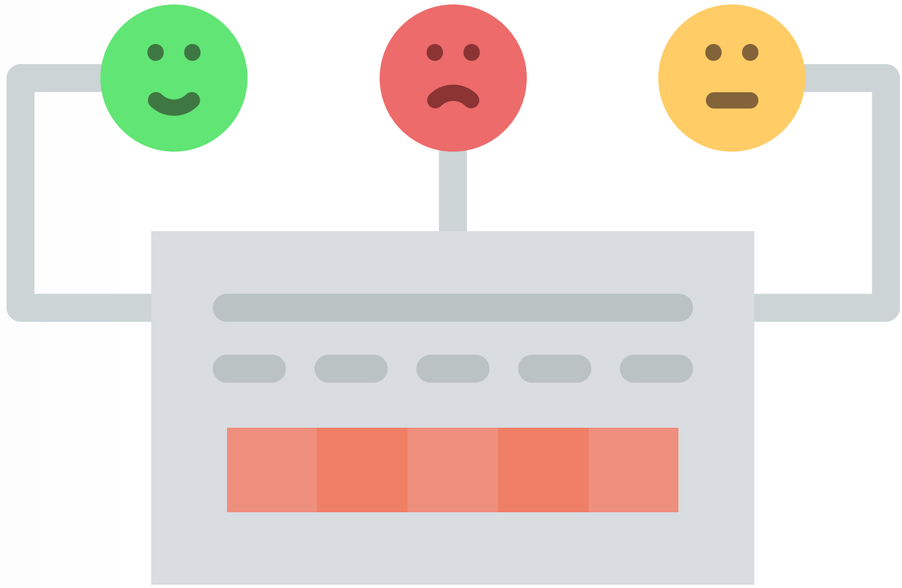
# Comment Sentiment Analysis

- The graph of sentiment analysis shows that more than 99% of the comments were neutral and a handful of negative comments are received.
- Positive comments may indicate a happy and engaged community.
- Diversity of Content: Sentiment analysis contributes to the diversity of content on explore pages. Platforms may balance positive and diverse content to cater to a wide range of user interests.



# Comment Sentiment Analysis

- We have used inbuilt TextBlob library in python.
- TextBlob is a Python library for processing textual data.
- It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more.

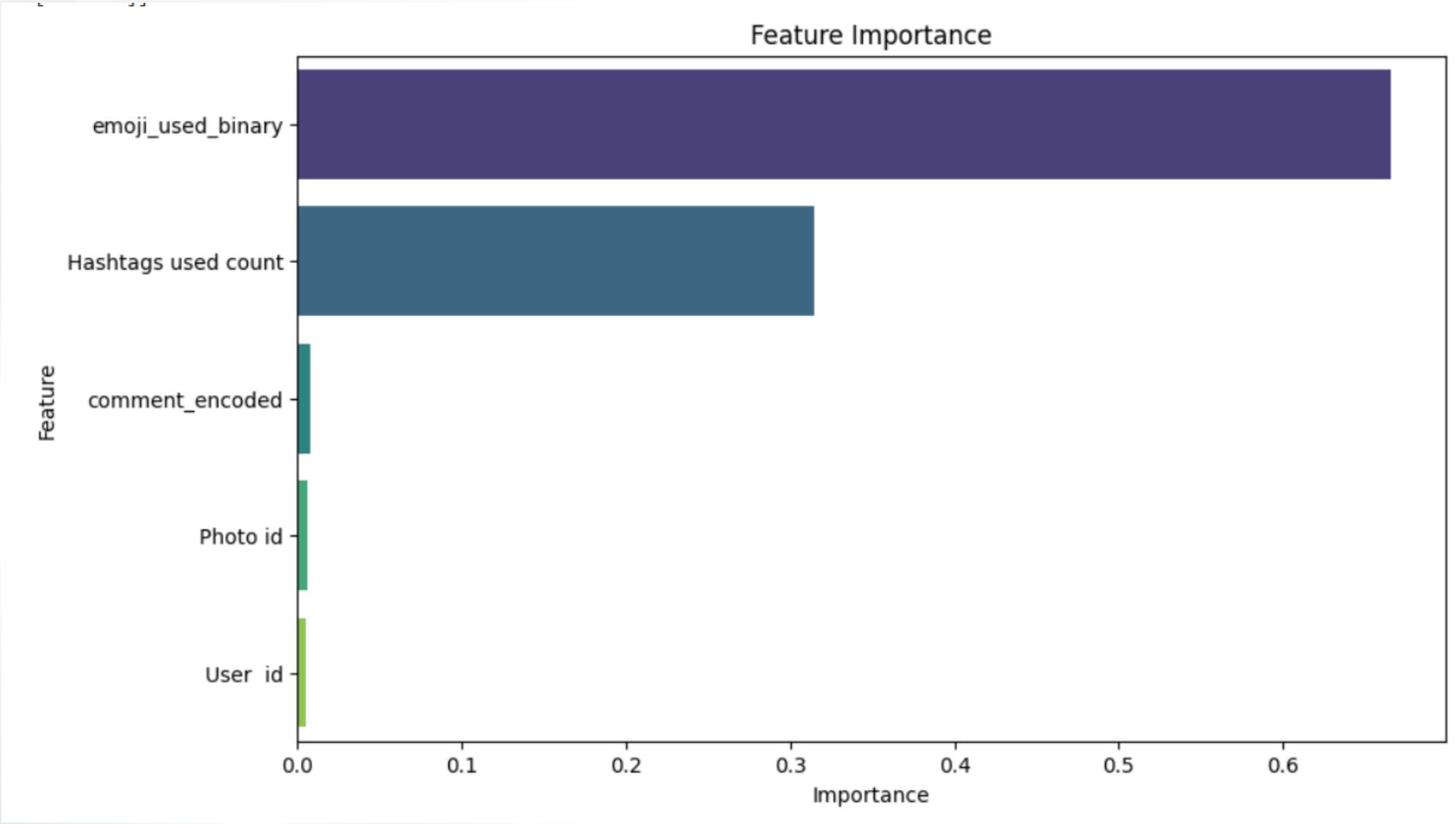


# Building the model

1. There are many ways to decide which content will appear on the explore pages. In our case, we decided that the final content that appears will depend on whether emoji has been used or not. For this, we took emoji as the target variable and comment, user id, photo id and hashtags used count as the input variables.
2. Now, if emoji is used, then a particular type of comment will appear on the explore pages and if emoji is not used, then another type of content will appear.
3. The content that appears may also depend on whether the emoji used is positive or negative.
4. We built the model using RandomForestRegressor and for comparison, also built another model with a single decision tree.
5. We have also analysed feature importance as which features have more say in whether the emoji will be used or not.

# Feature Importance

1. It is clear that only two features, namely emoji used and hashtags used count have a say whether emoji will be used or not.
2. Strikingly, the comment has very little say in deciding whether emoji will be used or not.
3. This is a very important insight that users use emoji on the basis of hashtags and not on the basis of comment which goes against our popular perception.



# Limitations

Analyzing page discoverability is a complex task and RandomLogisticRegressor may not be the best model for it.

Our dataset is quite small compared to actual user interactions on instagram.

Instagram’s algorithms are continuously changing and our dataset might not capture the most up-to-date algorithmic behaviours.

Instagram takes into account thousands of variables before deciding the content for your explore page, whereas we have very few variables present in our dataset like comments, emoji used, hashtags used count. So our models are bound to be less efficient in comparison to actual instagram’s algorithms for content discoverability.

# Our Team

Aditya Mundhada

Aaditya Kamble

Vivek Saroj

# Thank You

Do You Have Any Question?