Hashtag Trend Forecasting

Summer Internship

Bachelor of Computer Application

Submitted by:

Name: Devansh Baluni	**Sap_ID: 500119907**
Name: Anjali Danu	**Sap_ID: 500119189**
Name: Rohit Saini	**Sap_ID: 500125218**
Name: Daksh Kumar	**Sap_ID: 500125415**
Name: Shubhi Raj Saxena	**Sap_ID: 500119792**

Submitted to:

Dr. Shahina Anwarul

University of Petroleum & Energy Studies

Bidholi, Via Prem Nagar, Dehradun, Uttarakhand

June - July – 2025

Abstract

This project aims to identify trending hashtags across social media platforms, forecast their future popularity using predictive modeling, and visualize engagement insights through interactive dashboards in Power BI. Data was collected from social media datasets containing post-level engagement metrics and hashtag-level popularity statistics. The project combined data cleaning, exploratory data analysis (EDA), forecasting with Facebook Prophet, and visualization techniques to help identify the most effective hashtags for audience engagement.

Introduction

In today's fast-paced digital landscape, social media trends evolve rapidly, influencing public opinion, marketing strategies, and brand visibility. Hashtags serve as powerful tools for categorizing content and amplifying reach, making them a crucial element in online engagement. Understanding which hashtags are currently popular—and predicting future trends—can provide valuable insights for businesses, influencers, and content creators.

The relevance of this topic lies in its real-world application: brands can tailor their campaigns to ride on trending hashtags, while influencers can maximize engagement by aligning their content with predicted trends. With billions of posts generated daily, identifying patterns manually is impractical; therefore, data-driven approaches are essential.

In this project, we collected hashtag and engagement data from **Kaggle** datasets containing information such as postdate, platform, hashtag, likes, shares, and mentions. We pre-processed the data, performed exploratory analysis to uncover patterns, and implemented forecasting model using Facebook Prophet to predict future hashtag popularity. Visual dashboards were created using Power BI to present trends, regional engagement patterns, and forecasted results in an interactive manner.

Literature Review

Research in hashtag trend forecasting spans multiple approaches, from real-time big data analytics to advanced machine learning models.

Rodrigues et al. (2021) demonstrated real-time Twitter trend analysis using big data frameworks such as Apache Spark Streaming. They applied methods including hashtag counting, noun counting, cosine similarity, Jaccard similarity, Latent Dirichlet Allocation (LDA), and K-means clustering to identify trending topics quickly across domains such as politics, sports, and entertainment. Their results showed that big data tools can process high-volume social streams efficiently and improve real-time trend detection accuracy.

Shams et al. (2020) explored time series forecasting of Twitter hashtags using the Long Short-Term Memory (LSTM) neural network model. Focusing on the gaming community, they collected 24 days of hashtag data, accounted for both trend and seasonality, and demonstrated that LSTM could outperform traditional ARIMA for non-stationary social media data. This work highlighted the benefit of sequence-based deep learning for community-specific trend prediction.

Bansal et al. (2025) provided a comprehensive review of hashtag recommendation systems, tracing developments from simple frequency-based models to transformer-based deep learning architectures and graph neural networks (GNNs). They emphasized the challenges of semantic ambiguity, low adoption of hashtags, and multilingual complexity, as well as the importance of real-time, context-aware systems for improving content discoverability and engagement. The review also identified the lack of lightweight, resource-efficient methods as a persistent gap in the field.

These studies collectively show that while advanced models and big data infrastructure can significantly enhance trend forecasting, they require substantial computational power, specialized expertise, and extensive datasets — resources not always available to smaller teams or projects. This gap leaves space for simpler, more accessible forecasting pipelines that still deliver actionable insights for decision-making.

Problem Statement

Hashtags are integral to organizing content, improving visibility, and tracking social media trends. However, trends on platforms like Twitter and Instagram can emerge and decline within hours, creating a challenge for individuals, brands, and organizations that need to respond quickly. Existing research offers high-performance forecasting systems using big data pipelines or deep learning models such as LSTM and transformers, but these solutions often demand high-end hardware, large-scale datasets, and specialized technical skills.

During this project, several practical challenges shaped the design and execution:

- Data quality issues — missing records, inconsistent hashtag formatting, duplicates, and noise in engagement metrics.
- Tool limitations — inability to deploy large-scale ML models due to hardware constraints and team skill levels.
- Time constraints — the need to coordinate tasks across five team members, each focusing on different parts of the workflow (data collection, pre-processing, forecasting, Power BI visualization).
- Modelling restrictions — reliance on accessible methods rather than computationally heavy approaches.

These constraints underscored the need for a lightweight yet effective forecasting pipeline capable of identifying emerging hashtag trends from historical data and visualizing results in an intuitive way.

The approach in this project emphasizes accessibility, reproducibility, and adaptability, enabling similar teams to implement trend forecasting without advanced infrastructure.

Objectives

Core Objective: To forecast the future popularity of trending hashtags using historical social media engagement data and present the results through an interactive Power BI dashboard for actionable marketing insights.

Specific Objectives:

1. **Collect and compile** historical hashtag usage and engagement datasets from Kaggle for analysis.
2. **Clean and pre-process** the data by removing duplicates, handling missing values, and standardizing hashtag formats for consistency.
3. **Perform exploratory data analysis (EDA)** to identify top-performing hashtags and study engagement trends over time, across platforms, and by region.
4. **Implement the Facebook Prophet model** to predict future popularity trends for selected hashtags.
5. **Validate model performance** by comparing predicted values with historical trends to assess forecasting accuracy.
6. **Create an interactive Power BI dashboard** that visualizes engagement metrics, predicted popularity, and regional hashtag distribution.
7. **Provide data-driven insights** to help stakeholders optimize marketing strategies and content planning based on forecasted trends.

Methodology

Step 1- Dataset Identification

The dataset used in this project was sourced from **Kaggle** and contains detailed social media engagement and trending hashtag information. This data forms the foundation of our analysis, enabling us to track engagement patterns, identify top-performing hashtags, and forecast their future popularity.

- **social media.csv**

  Fields: Post_ID, Post_Date, Platform, Hashtag, Content_Type, Region, Views, Likes, Shares, Comments, Engagement_Level.
  Purpose: Engagement metrics per post.

- **TRENDING#.csv**
  Fields: Date, Hashtag, Mentions, Reach, Sentiment_Score, Top_Country.
  Purpose: Hashtag-level popularity and sentiment.

Step 2 – Data Preprocessing

The first phase involved preparing the datasets for analysis. We removed duplicate records and handled any null or missing values to ensure data quality. Hashtags were standardized by converting variations like #AI and #ai into a consistent lowercase format. Date fields were cleaned and formatted properly to maintain uniformity across the dataset. In addition, engagement metrics such as likes, shares, and comments were aggregated where necessary to create a consolidated view of audience interaction.

```python
import pandas as pd

# Load original data
df1 = pd.read_csv("Cleaned_Viral_Social_Media_Trends.csv")

# Preprocessing
df1['Post_Date'] = pd.to_datetime(df1['Post_Date'])
numeric_cols = ['Views', 'Likes', 'Shares', 'Comments']
df1[numeric_cols] = df1[numeric_cols].apply(pd.to_numeric, errors='coerce')
df1.dropna(subset=numeric_cols, inplace=True)
df1.drop_duplicates(inplace=True)
df1.dropna(inplace=True)
df1['Hashtag'] = df1['Hashtag'].str.strip().str.lower()
df1['Platform'] = df1['Platform'].str.strip().str.title()

#Save file
df1.to_csv("Step-2.csv", index=False)
```

Figure 1 - Cleaning the Dataset

```python
import pandas as pd

# Load your cleaned hashtag data
df = pd.read_csv("cleaned_hashtag_trends.csv")

# Define related hashtag groups (you can expand this)
hashtag_groups = {
    'ai': ['ai', 'chatgpt', 'machinelearning', 'deeplearning', 'artificialintelligence'],
    'fitness': ['fitness', 'workout', 'gym', 'fitlife'],
    'dance': ['dance', 'dancer', 'dancing', 'hiphop', 'choreography'],
    'fashion': ['fashion', 'ootd', 'style', 'trendy'],
    'food': ['foodie', 'food', 'cooking', 'yum', 'recipes'],
}

# Create an empty list to store grouped DataFrames
grouped_rows = []

# Go through each topic group and extract matching rows
for keywords in hashtag_groups.values():
    pattern = '|'.join(keywords)
    matches = df[df['hashtag'].str.contains(pattern, case=False, na=False)]
    grouped_rows.append(matches)

# Concatenate all groups into a single DataFrame
final_df = pd.concat(grouped_rows).drop_duplicates()

# Save the grouped data into one CSV
final_df.to_csv("grouped_hashtag_trends.csv", index=False)

print("Done! Saved all related hashtags in one file: grouped_hashtag_trends.csv")
```

Figure 2 - Grouped related Hashtag

Step 3 – Exploratory Data Analysis (EDA)

Using Python and Power BI, we performed exploratory data analysis to uncover key insights from the datasets.

We identified:

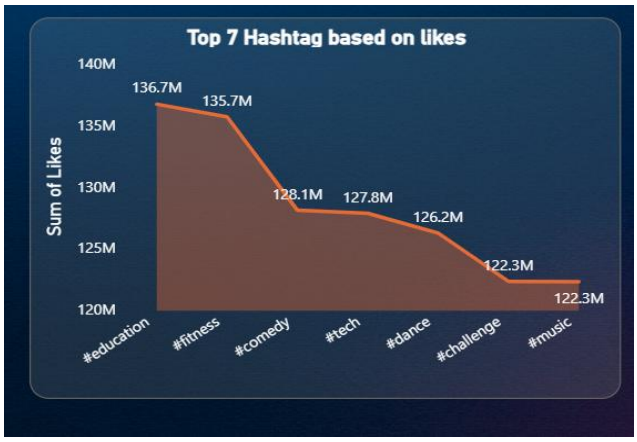


Figure 3 - Top 7 Hashtag based on Likes

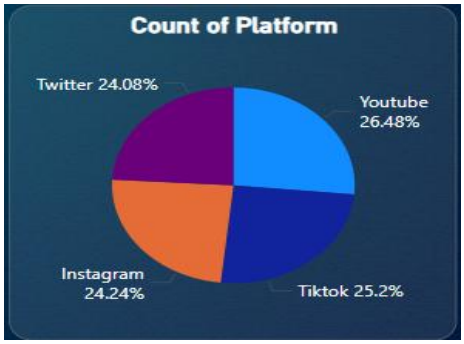


Figure 4 - Count of Platform

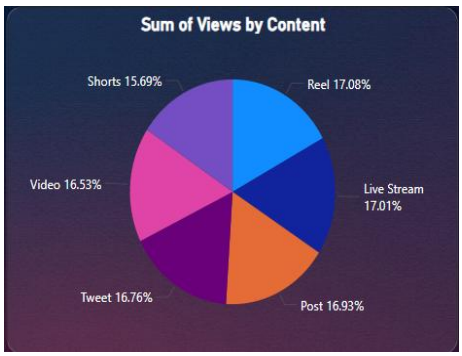


Figure 5 - Sum of Views by Content

Figure 6 - Count of Sentiment Score

Figure 7 - Sum of Comments

Figure 8 - Sum of Trend

Figure 9 - Sum of Shares

Figure 10 - Sum of Likes

Figure 11 - Sum of Views

Figure 12 - Average Estimated

Step 4 – Forecasting using Prophet Model

To predict future trends, we implemented the Facebook Prophet forecasting model. Before training, we applied a 7-day rolling average to smooth out short-term fluctuations and reduce noise in the data. We also generated an actual vs predicted comparison chart to evaluate how closely the model followed historical engagement patterns.

```
import pandas as pd
from prophet import Prophet
from sklearn.metrics import mean_absolute_error, mean_squared_error

# Load CSV
df = pd.read_csv('TRENDING#.csv')

# Choose which metrics to test
metrics_to_test = ['estimated_reach', 'sentiment_score']

# How many days to forecast
test_days = 30

# Loop through each metric
for METRIC in metrics_to_test:
    print("============================")
    print(f"Checking forecast for: {METRIC}\\n")

    agg = (
        df
        .groupby('date', as_index=False)[METRIC]
        .sum()
        .rename(columns={'date':'ds', METRIC:'y'})
    )
    agg['ds'] = pd.to_datetime(agg['ds'])

    # Sort by date
    agg = agg.sort_values('ds')

    train_df = agg.iloc[:-test_days]
    test_df  = agg.iloc[-test_days:]

    print(f"Training samples: {len(train_df)}")
    print(f"Testing samples: {len(test_df)}\\n")
```

Figure 13 - Accuracy Code - 1

```python
    m = Prophet()
    m.fit(train_df)

    future = m.make_future_dataframe(periods=test_days)
    forecast = m.predict(future)

    # Keep only forecast rows matching test dates
    forecast_test = forecast[forecast['ds'].isin(test_df['ds'])]

    comparison = test_df.merge(
        forecast_test[['ds','yhat']],
        on='ds',
        how='left'
    )

    # Compute errors
    mae  = mean_absolute_error(comparison['y'], comparison['yhat'])
    rmse = mean_squared_error(comparison['y'], comparison['yhat'])
    rmse = rmse ** 0.5
    mape = (abs((comparison['y'] - comparison['yhat']) / comparison['y'])).mean() * 100

    print("Model Accuracy Metrics:")
    print(f"MAE:  {mae:.2f}")
    print(f"RMSE: {rmse:.2f}")
    print(f"MAPE: {mape:.2f}%\n")

    # Print comparison preview
    print("First few rows of actual vs. forecast:")
    print(comparison.head(), "\n")
```

Figure 14 - Accuracy Code - 2

```
Checking forecast for: estimated_reach\n
Training samples: 335
Testing samples: 30\n
00:57:36 - cmdstanpy - INFO - Chain [1] start processing
00:57:37 - cmdstanpy - INFO - Chain [1] done processing
Model Accuracy Metrics:
MAE:  4040904.46
RMSE: 5366156.88
MAPE: 8.78%

First few rows of actual vs. forecast:
          ds         y          yhat
0 2025-04-28  51708848  4.884989e+07
1 2025-04-29  38622717  4.948216e+07
2 2025-04-30  50783093  4.969734e+07
3 2025-05-01  46953339  4.866039e+07
4 2025-05-02  51885170  4.894164e+07
```

Figure 15 - Accuracy of estimated Reach

```
============================
Checking forecast for: sentiment_score\n
Training samples: 335
Testing samples: 30\n
00:57:37 - cmdstanpy - INFO - Chain [1] start processing
00:57:37 - cmdstanpy - INFO - Chain [1] done processing
Model Accuracy Metrics:
MAE:  4.63
RMSE: 5.82
MAPE: 119.65%

First few rows of actual vs. forecast:
          ds     y      yhat
0 2025-04-28 -4.03  0.700608
1 2025-04-29 -4.39  1.879279
2 2025-04-30 -6.58  1.052691
3 2025-05-01  0.43  1.227141
4 2025-05-02  0.14 -0.309186
```

Figure 16 - Accuracy of Sentiment Score

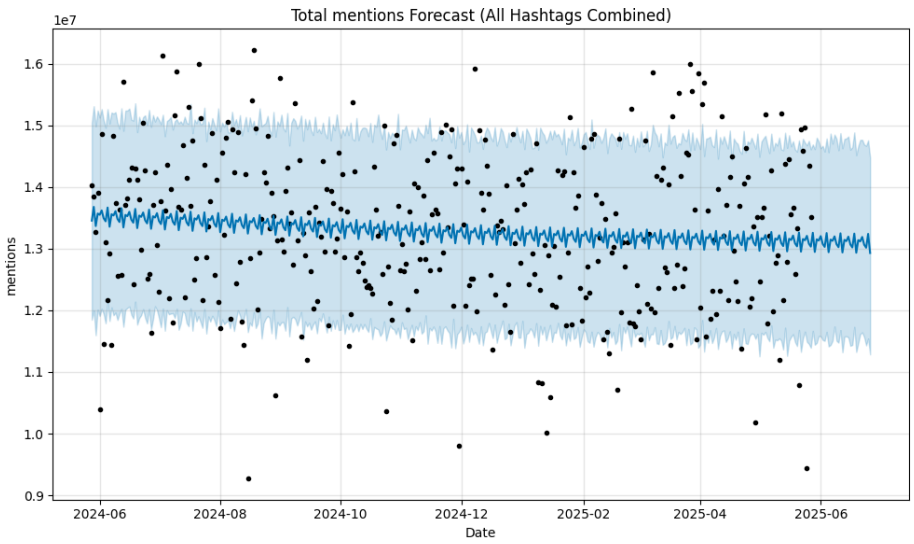


Figure 17 - Total mentions Forecast

Step 5 – Power BI Dashboard

Finally, we integrated our processed data and results into a comprehensive Power BI dashboard. This dashboard included multiple interactive visuals, such as engagement charts, forecast trendlines derived from, and a heatmap visualizing hashtag usage by country. These visuals allow stakeholders to explore engagement trends, forecasted patterns, and regional hashtag popularity in a dynamic and intuitive way.

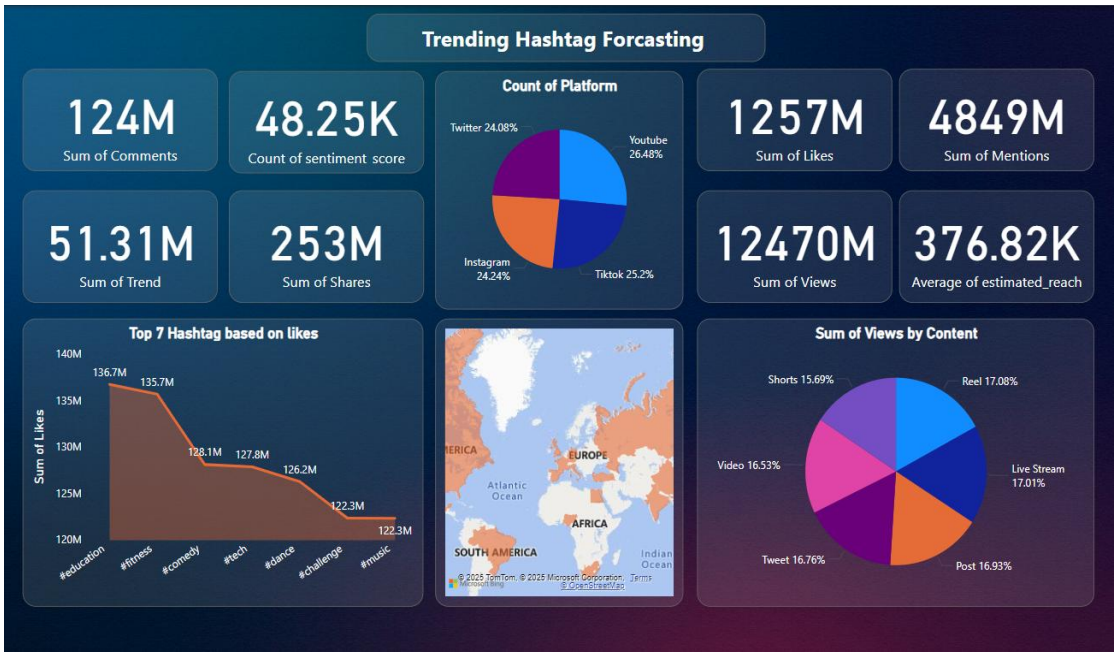


Figure 18 - Exploratory Data Analysis (EDA)

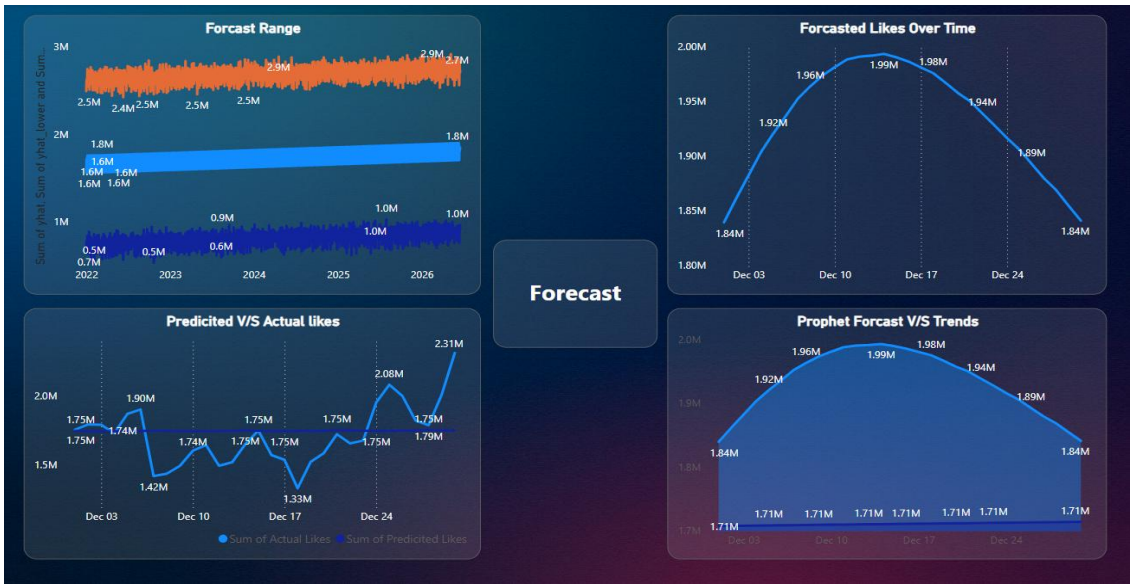
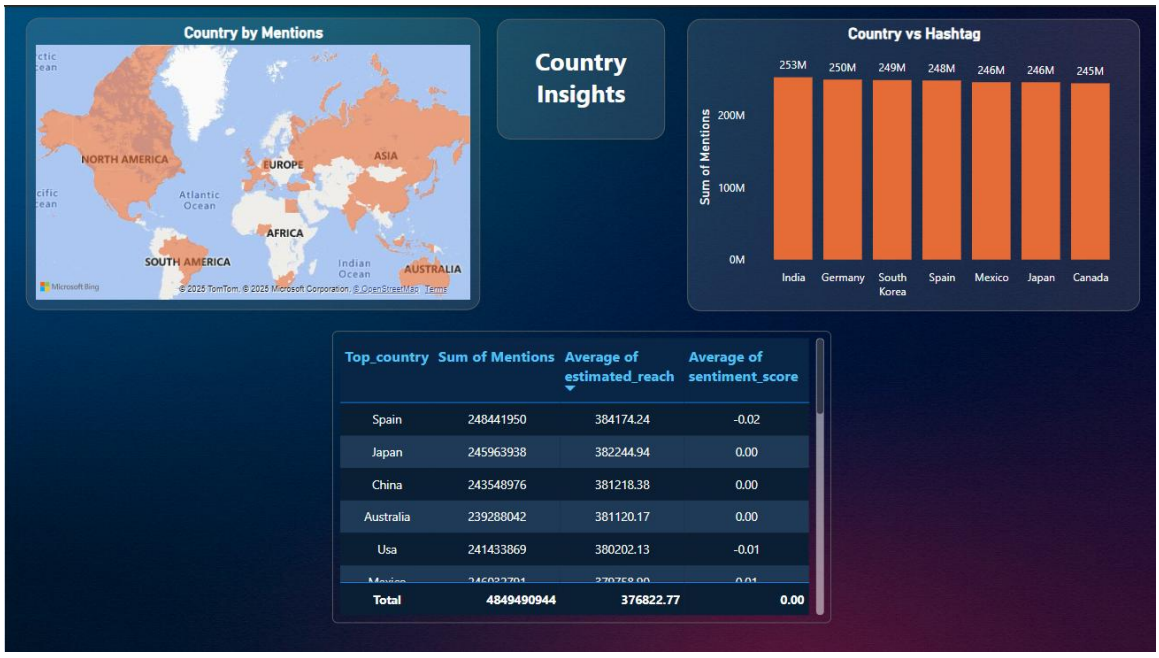


Figure 19 - Forecast vs Actual Dataset

Figure 20 - Engagement Level

Figure 21 - Trending Hashtag



Top_country	Sum of Mentions	Average of estimated_reach	Average of sentiment_score
Spain	248441950	384174.24	-0.02
Japan	245963938	382244.94	0.00
China	243548976	381218.38	0.00
Australia	239288042	381120.17	0.00
Usa	241433869	380202.13	-0.01
Total	4849490944	376822.77	0.00

Figure 22 - Country Insights

Results & Analysis

The analysis revealed the top-performing hashtags based on likes and mentions, highlighting the trends that drive the most engagement. Forecasting results using the Prophet model provided an overview of predicted popularity trends for these hashtags, offering insights into their potential future performance. The Power BI dashboard further enriched the analysis by showcasing platform-specific engagement patterns, identifying the content types that generate the highest interaction, and visualizing regional hashtag performance through heatmaps and trend charts.

Tools & Technologies Used

1. **Data Storage & Management**

   CSV/Excel Files: Used for storing and organizing the raw and cleaned datasets containing social media post-level engagement metrics and hashtag popularity statistics. Excel was also used for quick inspections, basic cleaning, and manual validation of the dataset before deeper analysis.

2. **Programming Language & Libraries**

   Python: The primary programming language for data **pre-processing**, analysis, and forecasting due to its flexibility.

- Pandas: Used for reading datasets, cleaning data, handling missing values, removing duplicates, and performing transformations like standardizing hashtag formats.
- Facebook Prophet: Time series forecasting library applied to predict future hashtag popularity.

3. **Visualization**
   Power BI:

   - Designed and developed an interactive dashboard to display top-performing hashtags, engagement patterns, and forecasted popularity trends.
   - Integrated charts such as forecast lines, bar graphs, and heatmaps for a multi-dimensional view of hashtag trends.

Limitations

While the project achieved its core objectives, several limitations influenced the scope, accuracy, and scalability of the results:

1. **Dataset Size & Coverage**
   - The dataset contained only a few thousand rows, which restricted the forecasting model's ability to learn robust seasonal and cyclic patterns.
   - Data was sourced from static Kaggle datasets rather than real-time social media feeds, limiting the ability to respond to sudden viral trends.
   - The dataset did not include multiple years of historical records, reducing the capacity to identify long-term seasonal effects.
2. **Data Quality & Granularity**
   - Missing and inconsistent entries (e.g., variations in hashtag formatting, missing engagement metrics) required manual cleaning, which may have resulted in minor data loss.
   - Engagement metrics were aggregated at the hashtag level without separating audience segments (e.g., age groups, platform demographics), limiting the depth of analysis.
3. **Modelling Constraints**
   - The forecasting relied solely on Facebook Prophet, which is effective for long-term trend and seasonality detection but less suited to capturing sudden, short-lived spikes in popularity.

- Advanced deep learning models such as LSTM, transformers, or hybrid ensemble approaches were not used due to hardware constraints, team expertise, and project timelines.

4. **Visualization & Interactivity**
   - While the Power BI dashboard provided interactive exploration, it was static in terms of data refresh, as no live data connection was established.
   - Regional engagement was visualized based on available fields but lacked integration with external geographic or demographic datasets.

Conclusion & Future Scope

This project successfully identified and forecasted hashtag trends using a combination of data analysis, visualization, and predictive modelling techniques. By integrating Python-based pre-processing and forecasting with interactive Power BI dashboards, the study provided stakeholders with actionable insights into current and predicted hashtag performance. These insights can help brands, marketers, and content creators align their strategies with emerging trends, ultimately enhancing audience engagement and reach.

The approach adopted in this project emphasizes accessibility, reproducibility, and adaptability, making it feasible for smaller teams or organizations without access to large-scale computing infrastructure. Although the methodology did not incorporate resource-heavy machine learning architectures, it demonstrated that lightweight solutions like Facebook Prophet can still yield meaningful results when applied to well-prepared datasets.

However, certain limitations must be acknowledged. The absence of real-time data streams meant that forecasts were based on static historical datasets, limiting responsiveness to sudden viral spikes. Additionally, the lack of raw textual data prevented the incorporation of sentiment analysis, which could provide a deeper understanding of audience perception. Prophet's forecasting strength in identifying long-term seasonal patterns also means it is less suited to capturing abrupt,

short-term fluctuations caused by breaking events or sudden influencer endorsements.

Future enhancements could include:

- Real-time API integration from platforms such as Twitter/X, Instagram, or TikTok to enable continuous trend monitoring and up-to-date forecasting.
- Hybrid modelling approaches that combine Prophet with advanced deep learning methods like LSTM or transformer-based architectures for improved short-term accuracy.
- NLP-powered sentiment analysis to complement quantitative engagement metrics with qualitative insights on audience emotions and attitudes toward trending topics.
- Multimodal trend analysis, incorporating both text and visual content to capture richer trend signals.
- Automated alert systems that notify stakeholders when specific hashtags show early signs of going viral, enabling proactive marketing responses.

By addressing these areas, the framework developed in this project could evolve into a robust, real-time trend intelligence platform, supporting data-driven decision-making in fast-moving digital environments.

References

1. https://facebook.github.io/prophet/
2. https://www.kaggle.com/datasets/tfisthis/global-trending-hashtags-dataset-300k-rows
3. https://www.kaggle.com/datasets/atharvasoundankar/viral-social-media-trends-and-engagement-analysis?resource=download
4. https://www.youtube.com/watch?v=6cV3OwFrOkk&t=32s