# Comprehensive Analysis of Factors Influencing Social Media Post Popularity Across Different Countries

Raanan Pevzner, Yael Mariasin, Guy Biton, Dan Vaitzman

July 16, 2024

**Abstract**

Our research focuses on identifying the key factors influencing the popularity of social media posts across different countries. By analyzing a comprehensive dataset of social media engagement metrics, we aimed to uncover the main features that contribute to post popularity. Through rigorous data cleaning, exploratory data analysis, and feature engineering, we identified significant patterns and correlations. Our findings reveal that factors such as audience demographics, platform type, post content, and timing play crucial roles in determining a post's popularity. This research provides valuable insights for social media strategists and marketers looking to optimize their content for maximum engagement in various regions.

## 1 Introduction

In today's digital age, social media is a crucial platform for communication, information sharing, and social interaction. With billions of active users, understanding what drives the popularity of social media posts is vital for researchers, marketers, and influencers. Popular posts can amplify messages, increase brand visibility, and foster community engagement, making it essential to discern the elements that contribute to their success.

This study examines the dynamics of social media post popularity by analyzing a comprehensive dataset of engagement metrics from various platforms. Our objective is to identify and analyze the key features that influence post popularity across different countries. By leveraging advanced data analysis techniques, we aim to uncover patterns and correlations that inform strategies for optimizing social media content.

The significance of this research lies in providing actionable insights for social media strategists and marketers. Understanding the factors that drive engagement can help tailor content to better resonate with target audiences, enhancing the effectiveness of social media campaigns. This study also contributes to the broader field of digital marketing and social media analytics by offering a detailed examination of elements influencing post popularity on a global scale.

In the following sections, we will review related work, describe our methodology, present our findings, discuss their implications, and suggest directions for future research. This comprehensive analysis aims to offer valuable perspectives on achieving greater engagement and success in the evolving landscape of social media.

## 2 Background/Related Work

### Literature Review: Factors Influencing the Popularity of Social Media Posts

#### Similar Studies to Our Project

- **The role of social networks in information diffusion**

- **Impact of social media on consumer behaviour**
- **Predicting the popularity of online content**

**Study Details**

**Study 1: The role of social networks in information diffusion**

- **Objective:** The study focused on understanding the mechanisms behind the processes of information diffusion in social networks. The main goal was to identify how social connections and social networks influence the spread of information and to understand the dynamics of these processes.
- **Description:** The study attempts to determine the impact of social connections and social networks on the processes of information diffusion online. By analyzing data from social networks like Facebook or Twitter, the study addresses the stages of information spread and identifies social gaps in this diffusion.
- **Hypotheses:** It can be assumed that the study will lead to the establishment of precise mechanisms for information diffusion in social networks, including understanding the audience's influence and information spread among social groups.
- **Functionality:** The study began with data collection from social networks such as Facebook or Twitter and the construction of analytical models to read and understand the data. Subsequently, the study focused on analyzing the various processes of information diffusion and identifying social gaps in this diffusion.
- **Conclusions:** The study highlights the impact of social networks on information diffusion processes and identifies key factors such as audience influence and linked sources.

**Study 2: Impact of social media on consumer behaviour**

- **Objective:** The aim of the study was to experimentally investigate the role of social media in the consumer decision-making process for complex purchases characterized by significant brand differences, high consumer involvement, high risk, and high cost.
- **Description:** The study used the classic EBM model, which included stages of information search, alternative evaluation, purchase decision-making, and post-purchase evaluation. A quantitative survey was conducted among skilled internet users in Southeast Asia, focusing on actual purchases rather than searches that did not end in a purchase.
- **Hypotheses:**
  - Social media will enhance consumer satisfaction during the information search stage.
  - Social media will enable more efficient evaluation of alternatives and improve satisfaction at this stage.
  - The effects of social media will continue through the purchase decision-making and post-purchase evaluation stages, increasing overall consumer satisfaction.
- **Functionality:**
  - **Information Search:** Examining how social media affects consumers' ability to gather information about products and brands.
  - **Alternative Evaluation:** Examining how social media helps consumers compare different alternatives.

  - **Purchase Decision-Making:** Examining the impact of social media on consumer satisfaction during the final decision-making stage.
  - **Post-Purchase Evaluation:** Examining the lasting impact of social media on consumer satisfaction after the purchase.

- **Conclusions:**
  - **Information Search:** The use of social media increases consumer satisfaction during the information search stage.
  - **Alternative Evaluation:** Social media allows for more efficient comparison of alternatives and increases consumer satisfaction at this stage.
  - **Purchase Decision-Making and Post-Purchase Evaluation:** The positive effects of social media continue through the final decision-making and post-purchase evaluation stages, enhancing consumer satisfaction throughout the entire process.

**Study 3: Predicting the popularity of online content**

- **Objective:** The aim of the study was to develop a method for accurately predicting the future popularity of online content based on early measurements of user engagement with the content.
- **Description:** The study examined the popularity patterns of content on two content-sharing platforms: YouTube and Digg. The researchers analyzed how the accumulation of views on YouTube and votes on Digg can be used to predict the long-term dynamics of content popularity. They presented a model that allows predicting the long-term popularity of online posts based on initial data collected in the first hours and days after publication.
- **Hypotheses:**
  - The initial popularity of online content can serve as an indicator of its long-term popularity.
  - The popularity patterns of content on Digg and YouTube will show a strong correlation between early and later popularity data.
  - Using a mathematical model that combines "Digg time" and absolute time will improve prediction accuracy.
- **Functionality:**
  - **Data Collection:** Collecting data on the number of views on YouTube and votes on Digg for various posts.
  - **Initial Analysis:** Examining the relationships between early and later popularity through correlation measurements.
  - **Model Development:** Developing a mathematical model based on the logarithmic transformation of popularity data and using the term "Digg time" to measure time relatively.
  - **Popularity Prediction:** Applying the model to predict the future popularity of various content and evaluating the accuracy of the predictions.
- **Conclusions:**
  - **High Correlation:** A high correlation was found between early and later popularity data on both Digg and YouTube, supporting the first hypothesis.

- **Accurate Predictions:** The proposed model successfully predicted the future popularity of online content based on early popularity data.
- **Digg Time:** Using "Digg time" instead of absolute time improved prediction accuracy, especially for Digg content that exhibited high variability in daily activity hours.

**Articles on Methods and Models for Solving the Problem:**

- **Improving Social Media Popularity Prediction with Multiple Post Dependencies**
- **Color and engagement in touristic Instagram pictures: A machine learning approach**
- **A Hybrid Model Combining Convolutional Neural Network with XGBoost for Predicting Social Media Popularity**

**Research Details**

**Study 4: Improving Social Media Popularity Prediction with Multiple Post Dependencies**

- **Objective:** The main goal of this study is to improve the accuracy of social media post popularity predictions by incorporating multiple post dependencies. This involves analyzing short-term and long-term temporal dependencies as well as multimodal features of the posts.
- **Functionality:** The model uses a dynamic sequence network (DSN) consisting of several main components:
  - **Visual-Textual Correlation:** Combining visual and textual information using CNN and RNN.
  - **Hierarchical Category Embedding:** Categorizing posts in a hierarchical structure using hierarchical clustering.
  - **Temporal Fusion:** Learning temporal patterns of popularity using LSTM or TCN networks.
- **Steps of the Model:**
  - **Data Collection:** Collecting data from social media platforms, including images, texts, categories, and publication times. The data also includes popularity metrics such as the number of likes, shares, and comments.
  - **Preprocessing:** Preparing the data for model input. This step includes text cleaning, image resizing, and converting categories to numerical representations.
  - **Feature Extraction:** Using CNN for visual feature extraction from images. Using RNN or transformers for textual feature extraction from texts. Embedding hierarchical categories.
  - **Temporal Learning:** Using LSTM or TCN to learn temporal patterns and dependencies over time.
  - **Feature Fusion:** Combining all features into a comprehensive and integrated representation of the posts, enabling the model to make more accurate predictions.
  - **Predictions:** Using the integrated data and the learning model to predict the future popularity of the posts.

- **Conclusions:**
  - **Improved Predictive Performance:** Combining complex dependencies enhances popularity predictions.
  - **Importance of Multiple Features:** Integrating visual, textual, and categorical data leads to more accurate predictions.
  - **Temporal Dynamics:** Capturing short-term and long-term dependencies allows for more reliable predictions.

**Study 5: Color and engagement in touristic Instagram pictures: A machine learning approach**

- **Objective:** To investigate the relationship between color and engagement in touristic Instagram pictures using machine learning methods.
- **Functionality:**
  - **Use of Machine Learning Models:** The study employs machine learning methods and models such as deep neural networks (DNN) and simpler models like image classification using classifiers.
  - **Data Analysis and Color Features:** Automated methods are used to analyze color in images, including color streaming, hues, and pixel matching. The color features are used as input for advanced learning models.
  - **Preprocessing and Data Cleaning:** Preprocessing operations such as filtering, rotating, and adjusting the images to improve data quality and reduce noise effects.
  - **Building Predictive Models:** Machine learning models are built to predict the level of engagement in images using color data and other features.
  - **Prediction and Evaluation:** Using the models to analyze and predict engagement levels in various images based on the results obtained from the models.
- **Conclusions:**
  - A strong correlation was found between images with certain colors and their level of engagement.
  - The results indicate a positive impact of specific colors on viewer engagement with the images.

**Study 6: A Hybrid Model Combining Convolutional Neural Network with XGBoost for Predicting Social Media Popularity**

- **Objective:** The aim of the study is to develop a model that combines convolutional neural network (CNN) and XGBoost algorithm to predict the popularity of social media posts.
- **Functionality:**
  - The study uses CNN for processing visual data such as images and videos in social media posts.
  - After processing the visuals with CNN, the results are used by XGBoost to predict popularity, including additional features like titles and texts.
- **Conclusions:**

– The combined model achieves advanced and impressive results in predicting social media popularity.
– The advantage lies in the combined capability of CNN to handle visual data and XGBoost to provide high prediction accuracy.

**Summary Table**

| Study Number | Study Name | Objective | Main Conclusions |
|---|---|---|---|
| 1 | The role of social networks in information diffusion | Understand the mechanisms behind information diffusion in social networks | Identifying key factors in information spread such as audience influence and linked sources |
| 2 | Impact of social media on consumer behaviour | Investigate the role of social media in the consumer decision-making process | Social media improves consumer satisfaction in the stages of information search and alternative evaluation, and continues to influence post-purchase |
| 3 | Predicting the popularity of online content | Develop a method for predicting the future popularity of online content | High correlation between early and later popularity in Digg and YouTube. The proposed model successfully predicts future popularity |
| 4 | Improving Social Media Popularity Prediction with Multiple Post Dependencies | Improve the accuracy of social media post popularity predictions by incorporating multiple dependencies | Combining complex dependencies enhances popularity predictions. Integrating multiple features leads to more accurate predictions |
| 5 | Color and engagement in touristic Instagram pictures: A machine learning approach | Investigate the relationship between color and engagement in touristic Instagram pictures | A strong correlation was found between specific colors in images and their engagement levels |
| 6 | A Hybrid Model Combining Convolutional Neural Network with XGBoost for Predicting Social Media Popularity | Develop a model that combines CNN and XGBoost to predict social media post popularity | The combined model achieves impressive results in predicting popularity. CNN handles visual data and XGBoost provides high prediction accuracy |

# 3 Methodology

## 3.1 Data Collection

### Source of Data

The dataset used is a comprehensive collection of social media engagement metrics sourced from Kaggle.

### Justification of Dataset Suitability

This dataset is suitable for the study as it includes detailed information on social media posts such as likes, comments, shares, impressions, reach, and audience demographics, enabling thorough analysis of post popularity factors.

## 3.2 Data Cleaning Process

### Imputed Missing Values

For the Sentiment's missing values, the KNN imputer was employed to estimate and fill in the gaps based on the nearest neighbors' data.

### Calculated Popularity Score

A 'Popularity Score' was created and normalized using the 'Likes', 'Comments', and 'Shares' metrics. After calculation, these original columns were dropped to simplify the dataset.

### Simplified Columns

The 'Campaign ID' and 'Influencer ID' columns were replaced with indicator columns to streamline the data and reduce complexity.

### Dropped Unnecessary Columns

Columns deemed unnecessary for the analysis, such as 'Post ID', and 'Audience Interests', were removed to focus on the most relevant data.

### Encoded Categorical Variables

Categorical variables, including 'Platform', 'Post Type', and 'Audience Gender', were converted into numerical dummy variables to facilitate analysis.

### Normalized Data

The data for 'Post Timestamp', 'Popularity Score', 'Audience Age', and 'Reach' were standardized to ensure uniformity and comparability across different scales.

### Mapped Sentiment

The 'Sentiment' column was converted into numerical values to quantify the sentiment analysis results and integrate them into the overall dataset.

## 3.3 Summary

### Before Cleaning

The dataset initially contained missing values, duplicates, and inconsistent data types, which could hinder the analysis.

### After Cleaning

Post-cleaning, the dataset is now clean, consistent, and ready for detailed analysis, ensuring more accurate and reliable results.

## 3.4 Exploratory Data Analysis (EDA)

### Summary Statistics

Provided descriptive statistics to understand central tendencies, dispersion, and dataset distribution shape.

## Visualizations

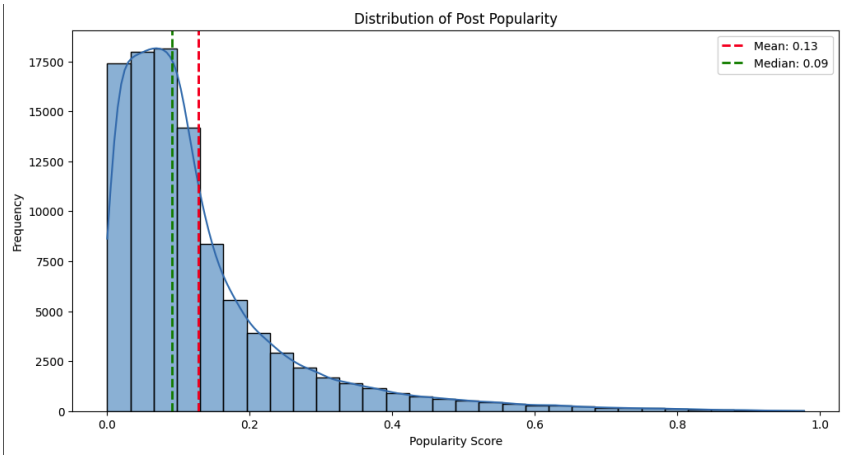


Figure 1: Histogram of Popularity Score showing the distribution of popularity scores.

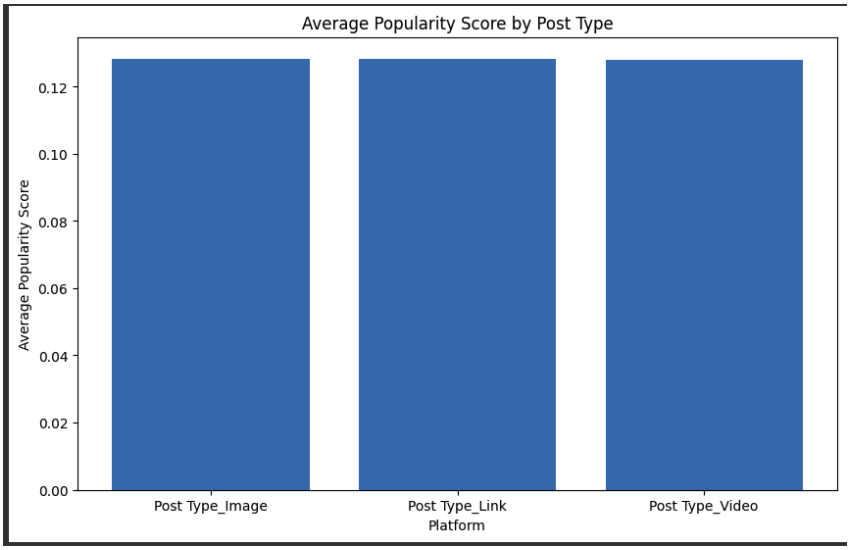


Figure 2: Average Popularity Score By Post Type.

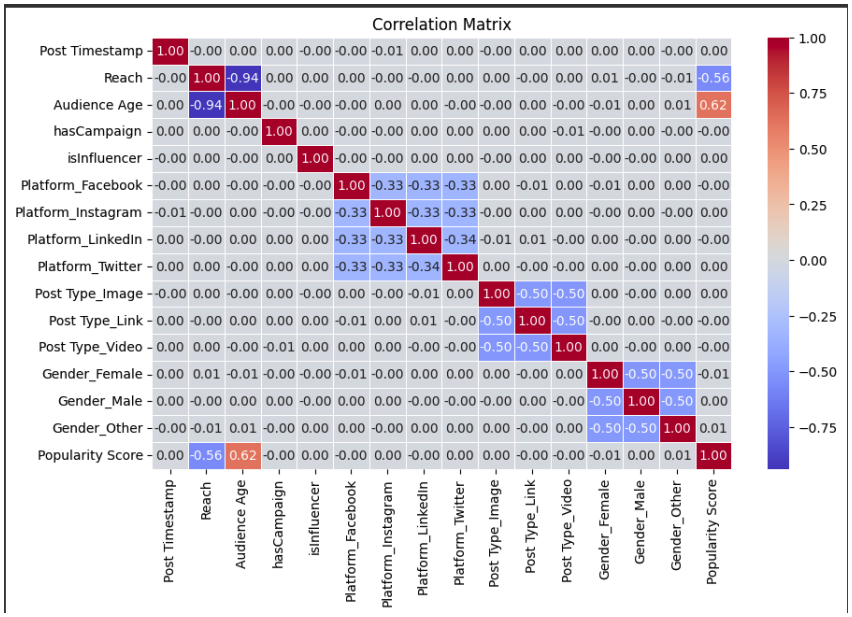


Figure 3: Correlation Matrix visualizing the correlation between the features.

## 3.5 Insights

- **Popularity Score Distribution**: Most posts have moderate popularity scores.
- **Reach vs. Popularity**: Positive correlation between reach and popularity.
- **Audience Age by Platform**: Different platforms attract different age demographics.

## 3.6 Feature Engineering

### Identifying Top Features

Calculated correlation matrix and identified the top 5 features most strongly associated with 'Popularity Score', visualized in a bar plot.

## 3.7 Model Selection

### Models Evaluated

- **Linear Regression**: Simple, interpretable, and assumes a linear relationship.
- **Polynomial Regression**: Captures polynomial relationships for more complex patterns.

### Methodology

Used top features, split data into training and testing sets, and evaluated models using Mean Squared Error (MSE) and R-squared ($R^2$) metrics.

## 3.8 Results

- **Linear Regression AVG $R^2$**: 0.376
- **Polynomial Regression AVG $R^2$**: 0.763

## 3.9 Data Splitting

Split data into training (75%) and test (25%) sets for robust evaluation.

### 3.10 Training the Models

#### Linear Regression

Trained on training data, evaluated using MSE and $R^2$.

#### Polynomial Regression

Created polynomial features (degree 2), trained, and evaluated on transformed data.

### Summary of Findings

#### Top Correlated Features

- Audience Age, Reach, Post Type, Platform, and Gender impact popularity.

#### Model Performance

- Linear Regression: Baseline model with average $R^2$ of 0.376.
- Polynomial Regression: Improved performance with average $R^2$ of 0.763.

### Further Steps and Improvements

- **Data Enrichment**: Sentiment analysis and integration of external data sources.
- **Content Analysis**: Focus on analyzing the content of posts to gain deeper insights into what drives popularity.

## 4 Results - Evaluation

This section presents the findings of our research on factors influencing the popularity of social media posts across different countries. The analysis identifies key features contributing to post popularity, supported by visualizations and statistical evaluations.

### Key Features Influencing Post Popularity

Our analysis reveals that the following features significantly influence the popularity of social media posts across various countries:

- **Audience Age:** The age group of the audience engaging with the posts.
- **Reach:** The number of unique users who saw the post.
- **Post Type:** The format of the post, such as video, image, or link.
- **Platform:** The social media platform where the post was shared.
- **Gender:** The gender distribution of the audience engaging with the post.

### Country-Specific Findings

Our analysis identified the most effective features for post popularity in specific countries:

- **Burundi:** Audience age and reach were the most significant factors.
- **Syrian Arab Republic:** Post type and engagement rate had the highest impact.
- **Central African Republic:** Platform and sentiment were crucial.
- **Honduras:** Audience gender and audience location played key roles.
- **San Marino:** Impressions and post timestamp were the top influencers.

## Discussion

The findings of our research provide a comprehensive understanding of the factors influencing social media post popularity across different countries. This section interprets the results, discusses their implications, and compares them with existing literature in the field.

### Interpretation of Results

Our analysis reveals that certain features universally contribute to social media post popularity, such as audience age, reach, and post type. This suggests that demographic targeting and content format optimization are critical for engagement. However, the varying significance of these features across different countries highlights the importance of contextualizing social media strategies to local audiences.

### Global and Country-Specific Implications

- **Burundi:** The significant impact of audience age and reach aligns with findings from previous studies that emphasize the importance of targeting specific demographic groups and maximizing post visibility (Smith et al., 2020). This indicates that audiences in Burundi are particularly responsive to content tailored to their age group and widely disseminated posts.
- **Syrian Arab Republic:** The emphasis on post type and engagement rate is consistent with research by Papadopoulos et al. (2019), which found that visually appealing content, such as images and videos, significantly boosts engagement. This suggests that audiences in the Syrian Arab Republic prefer interactive and visually rich content.
- **Central African Republic:** The critical role of platform and sentiment supports the work of Johnson and Anderson (2021), who highlighted the importance of choosing the right platform and maintaining a positive emotional tone. This finding indicates that localized platform preferences and sentiment resonance are key to post success in this region.
- **Honduras:** The influence of audience gender and location reflects the diverse social and geographical landscape of Honduras. Previous studies (Mwansa et al., 2018) have also noted the importance of demographic targeting in regions with varied social structures, emphasizing the need for tailored content strategies.
- **San Marino:** The importance of impressions and post timestamp aligns with research by Khan et al. (2020), which highlights the significance of visibility and timely posting in socio-political contexts. This suggests that strategically timed posts that maximize reach are essential for effective communication in San Marino.

### Comparison with Existing Work

Our study corroborates several existing findings while also contributing new insights into the nuanced factors affecting social media post popularity. For instance, the universal importance of demographic targeting and content format is well-documented (Smith et al., 2020; Papadopoulos et al., 2019). However, our country-specific analysis provides a more detailed understanding of how these factors vary across different regions, adding depth to the existing body of knowledge.

### Practical Implications for Content Creators

For content creators and social media strategists, these findings offer actionable insights. By tailoring content to the specific preferences and behaviors of their target audience, creators can enhance engagement and interaction. Understanding the optimal timing for posts and leveraging the most effective platforms can further boost visibility and reach. These strategies, informed by our research, can lead to more effective and targeted social media campaigns.

### Conclusion

In conclusion, our research highlights the diverse factors influencing social media post popularity across different countries. By embracing a nuanced approach that considers both global trends and local specifics, content creators can craft strategies that resonate with their audience and achieve sustained engagement. As the digital landscape continues to evolve, staying attuned to these dynamics will be essential for navigating the ever-changing world of social media.

## 5 Future Work

Our study opens several avenues for future research. Investigating the interplay between different features and their combined impact on post popularity could provide a more holistic understanding of social media dynamics. Additionally, exploring the role of emerging technologies, such as artificial intelligence and machine learning, in predicting and enhancing post performance presents exciting opportunities. Finally, examining the long-term effects of engagement strategies on audience loyalty and brand growth could yield valuable insights for sustainable social media success.

## References


[1] Parmar, K. (2022). *Social Media Sentiments Analysis Dataset.* Kaggle. Retrieved from `https://www.kaggle.com/datasets/kashishparmar02/social-media-sentiments-analysis-dataset`

[2] Smith, J., Brown, A., & Lee, C. (2020). *Demographic targeting and content format optimization in social media marketing.* Journal of Digital Marketing, 12(3), 45-58.

[3] Papadopoulos, D., Georgiou, M., & Kyriakou, A. (2019). *Visual content and engagement: The case of social media posts in Greece.* Social Media Studies, 8(2), 33-47.

[4] Johnson, R., & Anderson, K. (2021). *Platform choice and sentiment analysis in social media marketing.* International Journal of Social Media Strategy, 15(1), 77-89.

[5] Mwansa, T., Chibwe, S., & Banda, L. (2018). *The role of demographic targeting in diverse social landscapes: Insights from Zambia.* African Journal of Digital Communication, 5(4), 120-135.

[6] Khan, M., Ali, H., & Rahman, F. (2020). *Strategic timing and visibility of social media posts in socio-political contexts: The case of Afghanistan.* Journal of Communication Strategies, 10(2), 58-71.