Social Media Usage & Development Indicators Analysis

Final Project Report (DSA210)

Abstract

This study examines the relationship between social media usage patterns and country-level development indicators. By integrating platform-specific usage data with global development metrics, the project investigates whether differences in economic, institutional, and social development are associated with how societies engage with major social media platforms. Using correlation analysis, normalization techniques, and a regression-based machine learning model, the results indicate that while certain platforms exhibit statistically significant associations with development indicators, social media usage alone has limited explanatory and predictive power. These findings suggest that social media behavior reflects broader development contexts but should be interpreted as a complementary signal rather than a primary determinant of development outcomes.

1. Introduction

Social media platforms play a central role in communication, information dissemination, and social interaction worldwide. However, platform adoption and intensity of use differ substantially across countries. These differences may be influenced by underlying development factors such as economic prosperity, institutional quality, education, and overall quality of life.

The motivation for this project stems from a long-term interest in building a social media platform and understanding how users from different development contexts interact with digital environments. By empirically examining cross-country data, this study aims to contribute to a more nuanced understanding of how development levels shape digital behavior.

The central research question is:

Is there a statistically significant relationship between social media usage patterns and country-level development indicators?

To address this question, the following hypotheses were tested:

- **H0 (Null Hypothesis):** There is no correlation between social media usage and development indicators.
- **H1 (Alternative Hypothesis):** There exists a statistically significant correlation between social media usage and development indicators.

2. Data Sources

Three primary datasets were used in this study:

2.1 Social Media Usage Data

- **File:** `monthly_time_spend_by_country.csv`
- **Source:** Datareportal (2025)
- **Description:** Average monthly time spent on major social media platforms for the top 50 countries.

Since the raw dataset was not publicly available, platform-specific graphs published by Datareportal were extracted and converted into tabular form using Generative AI tools. This step was limited strictly to data extraction, not analysis.

2.2 Android Phone Usage Data

- **File:** `android_phone_use_by_country.csv`
- **Source:** StatCounter Global Stats
- **Description:** Country-level Android operating system market share, used as a proxy for smartphone penetration.

2.3 Development Indicators

- **File:** `development_country.csv`
- **Source:** Kaggle – 2023 Global Country Development and Prosperity Index
- **Description:** Composite indicators measuring economic performance, governance, institutional quality, and quality of life.

---

3. Data Preparation

A structured data preparation pipeline was implemented to ensure consistency and reliability:

- Standardization of `Country_Code` values across datasets
- Replacement of missing or invalid entries ("-") with NaN
- Conversion of all numeric columns to appropriate numeric types
- Merging datasets using `Country_Code` as the key
- Removal or handling of incomplete observations where necessary

To account for cross-country differences in device accessibility, platform usage variables were normalized using Android phone penetration rates. This normalization step aimed to control for structural access constraints rather than behavioral differences.

The final cleaned and merged dataset was generated within the project notebooks and used for all subsequent analyses.

---

4. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to gain an initial understanding of the data structure and relationships. The EDA focused on:

- Distributional properties of social media usage variables
- Identification of outliers and skewness
- Missing data patterns across countries
- Pairwise relationships between platforms and development indicators

Visual tools such as correlation heatmaps and scatter plots were used extensively. These visualizations provided early indications that platform-specific usage patterns differ in how they relate to development dimensions.

5. Normalization Strategy

Raw social media usage values can be misleading when comparing countries with unequal access to smartphones. To mitigate this issue, usage metrics were normalized by Android penetration rates.

This approach assumes that Android market share serves as a reasonable proxy for smartphone availability, particularly in developing regions. Normalization helped isolate behavioral usage differences from access-driven effects.

6. Correlation and Hypothesis Testing

Two complementary statistical methods were employed:

- **Pearson correlation** to capture linear relationships
- **Spearman rank correlation** to capture monotonic, non-linear relationships

Each platform’s normalized usage was tested against each development indicator. Statistical significance was evaluated using a threshold of **$p < 0.05$**.

Comparing Pearson and Spearman results allowed for robustness checks and reduced sensitivity to outliers and non-normal distributions.

7. Machine Learning Analysis

To further evaluate the relationship between social media usage and development indicators, a regression-based machine learning model was implemented.

- **Objective:** Assess the predictive power of social media usage variables
- **Target:** Selected development indicators
- **Result:** $R^2 \approx 0.097$

The low R^2 score indicates that social media usage alone explains only a small fraction of the variance in development indicators. This confirms that development outcomes are driven by complex, multi-dimensional factors beyond digital behavior.

8. Results and Interpretation

The analysis revealed several notable patterns:

- Platform-specific associations vary across development dimensions
- **LinkedIn** and **Pinterest** tend to show positive correlations with institutional quality and quality-of-life indicators
- **TikTok** and **YouTube** exhibit more negative associations with governance and economic indicators
- Correlation strengths are generally modest, even when statistically significant

Machine learning results align with the statistical findings by reinforcing the limited standalone explanatory power of social media usage.

Overall, the null hypothesis (H0) is partially rejected: statistically significant correlations exist, but they are weak to moderate and highly platform-dependent.

9. Limitations

Several limitations should be acknowledged:

- Limited country coverage (top 50 countries)
- Potential measurement noise from graph-to-table data extraction
- Android penetration as an imperfect proxy for total smartphone access
- Cross-sectional design limits causal interpretation

These constraints suggest caution in generalizing results beyond exploratory insights.

10. Conclusion

This project demonstrates that social media usage patterns are related to country-level development indicators, but the relationship is neither strong nor sufficient for prediction on its own. Social media behavior reflects broader socioeconomic and institutional contexts, yet development outcomes depend on a wide range of structural factors.

From a platform design perspective, the findings highlight the importance of tailoring digital products to the developmental realities of different societies. Future work could extend this analysis by incorporating longitudinal data, additional access metrics, or platform-specific engagement features.

11. AI Tool Usage Disclosure

Generative AI tools were used strictly for supportive tasks, including:

- Converting publicly available graphical data into tabular format
- Assisting with code debugging and refactoring
- Improving documentation clarity

All analytical decisions, hypothesis formulation, statistical testing, interpretation of results, and conclusions were independently conducted by the author. AI tools were not used to generate results or automate analysis.