Analysis of Posting Strategies for Higher Education Institutions

Faryal Tarique
Department of Computer Science
University of Porto
Porto, Portugal

Inês Oliveira
Department of Computer Science
University of Porto
Porto, Portugal

Maria Miguel Ribeiro
Department of Computer Science
University of Porto
Porto, Portugal

***Abstract*—This study explores the dynamics of Twitter usage by Higher Education Institutions (HEIs) to improve strategic communication and outreach. Through comprehensive data cleaning and transformation, we employed a variety of exploratory data analysis techniques, including clustering and text mining, to dissect the temporal and content-based patterns of HEI Twitter posts. Our findings reveal distinct posting behaviours among HEIs, characterized by their frequency, timing, and thematic focus.**

I. Introduction

This project explores Twitter data from various HEIs over the past year to systematically identify common posting behaviours and forecast emerging trends. Our analysis aims to categorize these posts, pinpoint peak activity periods, and understand content themes and sentiments.

II. Data Preparation and Transformation

A. Initial Exploratory Data Analysis (EDA)

As the first step, we created a diagnostic report that revealed that our dataset was robust, with generally well-distributed data but highlighted areas requiring detailed investigation. We then systematically approached each concern, ensuring data integrity by confirming there were no complete duplicate rows and validating that numeric columns contained no inappropriate values, such as negative counts. This validation process confirmed the reliability of our data, which is crucial for accurate analyses.

From our drill-down investigations, we learned that while our dataset contained outliers, these were not removed immediately, preserving data points that might represent significant trends or anomalies essential for understanding broader patterns or rare occurrences. Additionally, the identification and planned imputation of missing data using the K-Nearest Neighbors method ensured that our dataset would remain comprehensive without losing valuable information.

B. Cleaning Processes

After gaining insights about the data quality from our preliminary exploratory data analysis (EDA), we proceeded to clean the dataset based on our observations. The cleaning process began by removing erroneous values; specifically, we eliminated rows where the 'id' was 'complutense.csv', which appeared to be a misclassified entry with only one occurrence. This step ensured that our dataset did not include any misleading data that could skew the analysis.

For handling missing values, we applied log transformations to numerical columns to facilitate more stable imputation, particularly avoiding issues with outliers. Missing values in the $'view_count'$ column and other count columns were addressed through K-Nearest Neighbors (KNN) imputation using log-transformed predictors. This allowed us to leverage the underlying patterns in the data, predicting missing values based on the similarities between entries. After imputation, we transformed the log values back to their original scale and removed temporary log columns to clean up our dataset.

C. Feature Engineering

In our effort to prepare the dataset for deeper analytical tasks, we significantly enhanced it by extracting new temporal features from the $'created\_at'$ column. We started by converting the timestamps into a more usable POSIXct datetime format, ensuring accuracy for time-based calculations. We then extracted the hour of the day and the day of the week, enriching the dataset with details that are vital for analyzing posting time distributions and identifying daily or weekly trends. The hour function was used to retrieve the hour part of the date-time, while the weekday was extracted using the wday function, returning full-day names for better readability and understanding.

Further, we included extracting the year and the month from the date. We combined them into a $'year_month'$ feature to facilitate trend analysis over extended periods. The month was noted in both numerical and abbreviated textual formats to enhance the utility of the temporal data in various analytical contexts.

III. Data Analysis Through Visualizations

Through the detailed visualizations provided, we analyzed posting frequencies, content types, and engagement metrics across various academic institutions.

A. Temporal Analysis

Harvard leads in posting frequency, suggesting high activity levels compared to Duke and EPFL, while Leicester and Manchester show moderate activity.

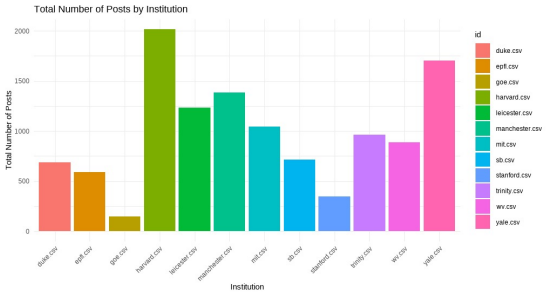

Fig. 1: Hourly Posting Frequency by Institution

The **Hourly Posting Frequency** graph illustrates peak activity between 7 AM and 5 PM for each HEI, suggesting a strategic focus on peak engagement times.

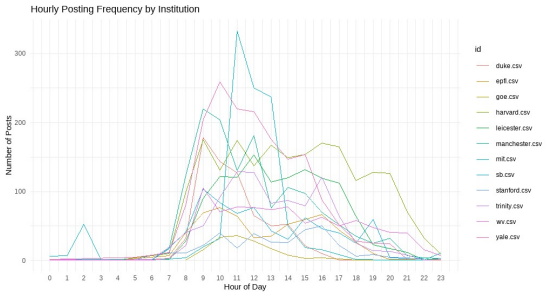

Fig. 2: Hourly Posting Frequency by Institution

Additionally, Wednesday and Thursday emerge as the most active days in the **Daily Posting Frequency** analysis, indicating a mid-week engagement strategy.

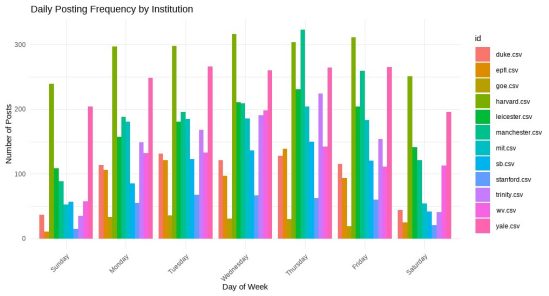

Fig. 3: Daily Posting Frequency by Institution

B. Content Analysis

Analysis of the **Average Post Length by Institution and Type** reveals that EPFL and GOE have longer posts, especially in retweets, indicating detailed interactions.

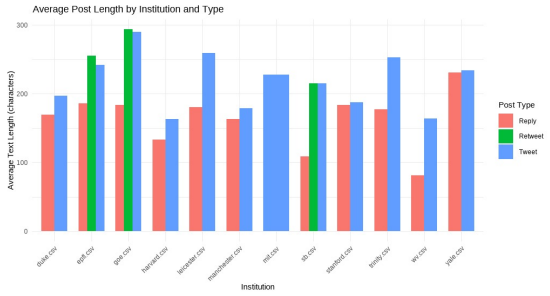

Fig. 4: Average Post Length by Institution and Type

The **Proportion of Post Types by Institution** graph shows most institutions predominantly use tweets, with Manchester exhibiting a balanced approach, suggesting a strategy geared towards interactive engagement.

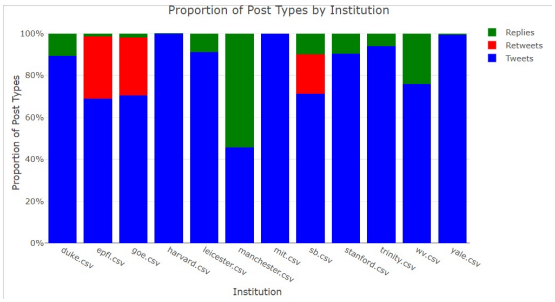

Fig. 5: Proportion of Post Types by Institution

C. Engagement Analysis

The **Bubble Chart of Engagement by Day and Hour** indicates the highest engagement on Wednesday morning and Saturday nights, optimal times for posting.

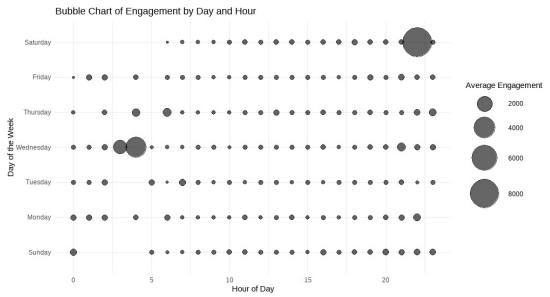

Fig. 6: Bubble Chart of Engagement by Day and Hour

The **Impact of Media Type on Engagement** pie chart shows a dominant preference for photos (82.4%), with videos (17.1%) and animated GIFs (0.652%) far less common. This suggests that photos are the most effective media type for engagement in institutional settings.

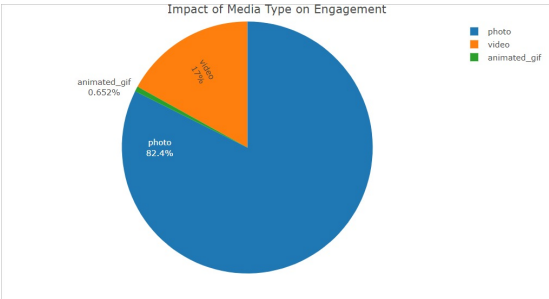

Fig. 7: Impact of Media Type on Engagement

The **Engagement by Institution** charts highlight significant disparities in engagement between regular and high-impact posts. Harvard notably achieves higher engagement on high-impact posts, with substantial engagement peaks indicating effective resonance with the audience.

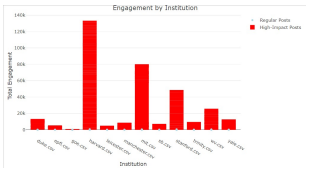

Fig. 8: High Impact Posts

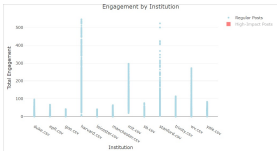

Fig. 9: Regular Posts

IV. Clustering Analysis

We analyzed the different clusters of HEIs present based on communication patterns and engagement metrics.

Data preprocessing involved extracting significant temporal and content-related features from the Twitter posts of Higher Education Institutions (HEIs). The key features calculated for clustering included: average post length, peak post hour, average posts per day, weekday vs. weekend ratio, average engagement per post, hashtag use frequency, media use rate, and evening post proportion. These metrics formed the basis for clustering the institutions, allowing for a detailed comparison of their social media strategies and audience engagement.

To determine the optimal number of clusters, we employed the Elbow Method and identified that five clusters best minimized within-cluster variance.

A. K-means Clustering

K-means clustering was executed to segment the institutions into five distinct groups based on their engagement and posting characteristics.

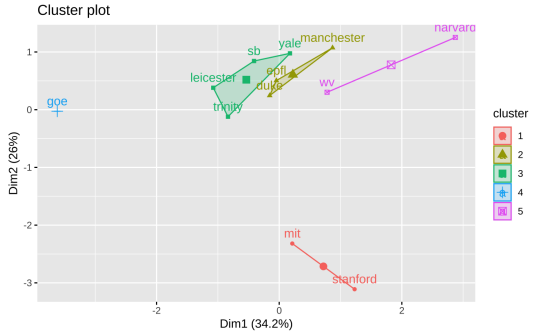

Fig. 10: K-means clusters

B. Hierarchical Clustering

Hierarchical clustering was also applied as a complementary approach to identify hierarchical relationships between institutions. The dendrogram displayed provides a visual interpretation of these relationships, offering insights into how institutions gradually group together.

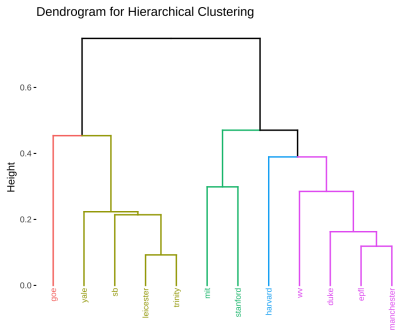

Fig. 11: Dendogram for Hierarchial Clustering

V. Topic Categorization

The categorization of each word to its correspondent category, being one of 5 (education, research, image, society and engagement) is done with 5 steps:

A. Data Preparation

The first step involves preparing the data for the categorization. We start by filtering posts from the ”data” set to exclude:

- Replies (posts with type ”Reply”).

B. Creation of the corpus cleanup function

Before categorization, the *corpus* undergoes a preprocessing step to clean it. This involves: conversion to lowercase for consistency; punctuation, number, stop word and emoji removal, etc.; text normalization; specific word removal and lemmatization.

C. Creation of the categorize_documents function

This function classifies documents (posts), based on predefined categories and their associated words.

The function takes three arguments: a set of documents to be categorized, a list of top words and a dictionary that associates each category and its associated words.

If a document is empty (no text content), it's categorized as "education", by default.

For each document with text content, a counter is initialized to track the occurrences of each category. The function iterates through each top word. If the top word exists in both the category dictionary and the current document text, the count for the corresponding category in the current document's counter is incremented.

After iterating through all top words, the category with the highest count for the current document is assigned as the document's category.

Finally, the function returns the list containing the assigned categories for all the documents.

D. Creation of the dtm_remove_lowfreq function

This step refines the data further by removing terms (words) that appear infrequently and removes documents that are empty:

E. Institution-Specific Categorization

This step iterates through each institution's data, and is where the functions mentioned above are used:

First, the documents belonging to the current institution are filtered based on their ID. The documents are converted into a suitable format for analysis (e.g., VCorpus). Then, the cleaning function is applied to the documents.

A document-term matrix (DTM) is created to represent word frequencies within the institution's data. The DTM is converted into a term frequency-inverse document frequency (TF-IDF) matrix for weighting terms.

Low-frequency terms are removed (optional: empty documents resulting from this removal might also be removed).

Now the top 20 most frequent words can, finally, be identified, by extracting the highest frequencies of the words obtained in dtm.tfidf.

With those 20 words, we created a dictionary that associates each word with one of 5 categories, being: education, society, image, engagement and research. A wordcloud is also plotted with those 20 top words.

Fig. 12: Example: wordcloud of Yale institution

The categorization function is applied to classify the documents, and the assigned categories are saved.

Finally, a $'category'$ column is added to the $'documents'$ dataset. This column is populated with the category assigned to each post based on the classification results. After that, a data frame $'doc\_text\_cateogry'$ is created and contains each posts' institution id, text and category.

To analyze the frequency of each category for each institution, we created a category frequency barplot:

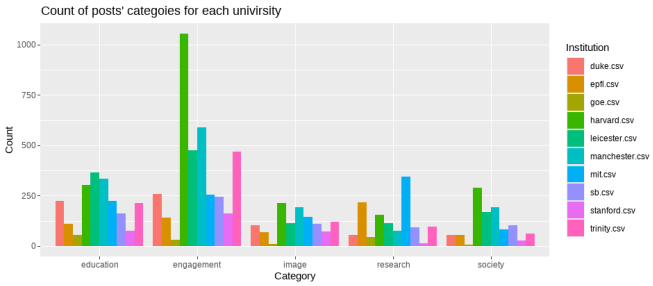

Fig. 13: Categories' frequencies for each university posts $'doc\_text\_cateogry'$

As we can see, there is an exceptionally high count of engagement related posts related to the Harvard university. There are also counts close to zero for image, research and society for "goe.csv", "stanford.csv" and "goe.csv", respectively. The "image" category is the one that has the smallest range in terms of quantity of posts.

VI. Sentiment Analysis

We analysed the post sentiments to gain comprehensive insights into the public perception and internal communication tone of each HEI. Here are the analyses and visuals we achieved.

A. Overall Sentiment Analysis Over Time

The **Monthly Sentiment Trend** chart illustrates the evolution of average sentiment in posts over time for the examined HEIs.The trend shows a general increase in positive sentiment, particularly noticeable in the first and last quarters of the period analyzed. There are positive peaks in May and June 2023, which is roughly the same time when Covid-related bans were uplifted in the US. This suggests that certain events or time periods, potentially aligning with public health updates, have a pronounced impact on the sentiment conveyed in institutional communications.

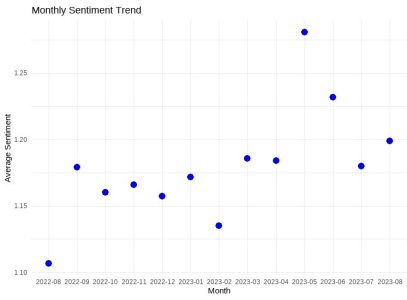

Fig. 14: Monthly Sentiment Trend

Further analysis of the **Daily Sentiment Trend** graph in our code supports the findings from the monthly trends, showing that on average, the sentiments expressed by the HEIs were generally positive, with noticeable fluctuations and peaks throughout the two-year period analyzed.

B. Emotion Analysis in HEIs

This part aimed to identify and quantify the prevalence of various emotions such as joy, trust, anticipation, and others within the social media communications of Higher Education Institutions (HEIs).

The **Distribution of Emotions** bar chart provides a clear view of the overall emotional makeup of the posts. It shows that positive emotions, especially trust and anticipation, dominate the emotional spectrum. This suggests a generally optimistic tone in the communications across HEIs.

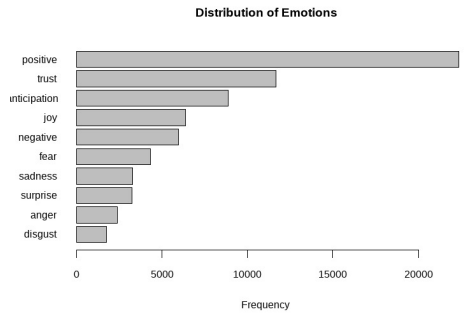

Fig. 15: Overall Distribution of Emotions in HEI Communications

The **Distribution of Emotions in HEIs** charts break down these emotions by individual institutions. These detailed charts indicate variability in emotional expression among different HEIs. Some institutions exhibit a higher prevalence of joy and trust, while others have noticeable levels of negative emotions like sadness and anger, reflecting diverse communication styles and possibly different institutional cultures or responses to specific events.

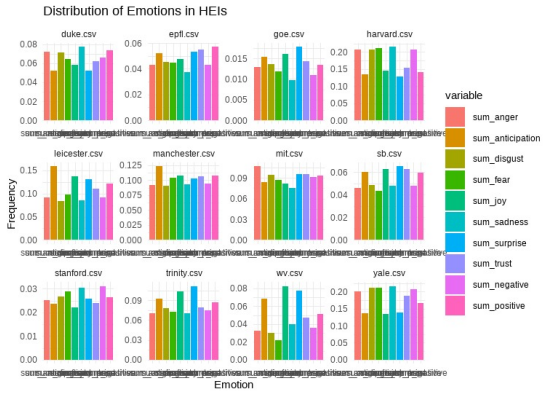

Fig. 16: Detailed Distribution of Emotions Across HEIs

These visual analyses illuminate the predominant emotional tones conveyed by each institution.

C. Time Series Analysis of Sentiment by HEI

The time series analysis utilizes the "Syuzhet Plot" to track the progression of emotional valence over the content posted by each HEI.

The x-axis, labeled "Order," represents the chronological sequence of content posted, sorted by the date and time each post was made, giving each post a sequential identifier. The y-axis displays sentiment scores, ranging from negative to positive values, derived from sentiment analysis performed on the text of each post.

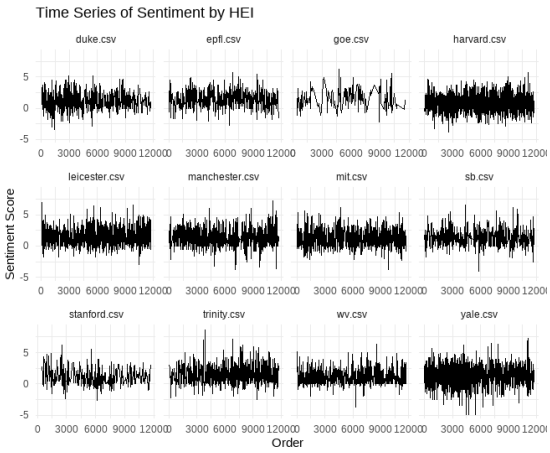

Fig. 17: Time Series of Sentiment by HEI

The plots exhibit varying levels of volatility. For example, GOE shows wider fluctuations in sentiment, indicating a broad range of content types and diverse audience reactions. In contrast, institutions like Harvard display more stability, suggesting a consistently moderated tone in communications.

Most HEIs display sentiment scores fluctuating around the zero line, suggesting a balanced mix of positive and negative sentiments over time. There are no consistent positive or negative trends, indicating dynamic interactions and diverse topics covered by the institutions. Notable spikes in the plots, both positive and negative, may indicate posts or events that elicited strong reactions.

VII. Conclusion

The comprehensive analysis of HEIs on Twitter provided significant insights into their strategic communication practices. Our study revealed diverse posting behaviors, with notable differences in frequency, timing, and content themes across various institutions. Advanced data analysis techniques including clustering and sentiment analysis allowed us to categorize HEIs based on their communication patterns, identify peak activity times, and assess the emotional tone of their posts.

The temporal patterns indicate optimal times for posting to maximize engagement, while the sentiment and emotion analysis provides a deeper understanding of the audience's perception and reactions to the content posted by HEIs. The clustering analysis highlights the potential for segmenting HEIs into distinct groups based on their social media behavior.

Overall, this project contributes valuable perspectives to the field of digital communication in education, offering actionable insights for HEIs to refine their social media strategies and foster more meaningful interactions with their communities.