Analysis of Posting Strategies for Higher Education Institutions

Daniela Tomás up202004946@edu.fc.up.pt

Diogo Almeida up202006059@edu.fc.up.pt

Diogo Nunes up202007895@edu.fc.up.pt

João Veloso up202005801@edu.fc.up.pt

Abstract In today's digital age, social media has become an essential platform for communication and engagement, particularly for Higher Education Institutions. This report aims to provide a comprehensive understanding of different social media strategies and analyze the sentiments and emotions the posts are trying to convey. For that we developed a series of functions and tables to better understand the different strategies. We also used sentiment analysis techniques from the R library syuzhet to analyze the emotional tone and sentiment of the various posts made by the various institutions. We found that most of the posts were made during the week with more posts being in the middle of the day. In the content patterns we found that trust and anticipation were the most common emotions, while disgust was the least expressed, indicating an overall positive communication tone. Overall, this study underscores the importance of social media as a tool for communications and engagement for Higher Education Institutions. The insights gained from this analysis can help them refine their social media strategy to enhance their outreach and engagement with their target audiences.

I. Introduction

In today's digital age, social media has become an essential platform for communication and engagement, particularly for Higher Education Institutions (HEIs). These institutions utilize platforms like X/Twitter to disseminate information, promote events, share research findings, and engage with a broader audience. Understanding the patterns and content of these posts can provide valuable insights into their communication strategies and public engagement efforts.

Through this study, we aim to provide a comprehensive understanding of HEIs' social media strategies and then analyze the sentiments and emotions the posts are trying to convey.

II. Data Pre-Processing

In order to get a comprehensive idea into the patterns we first need to handle any NA values to ensure the accuracy of our statistical calculations.

To view how many values are missing from each column we implemented the function na_count that iterates through each column and counts the number of NA values, then prints them into a table. From that table we learned that only one column, view_count, contains NA values, but since there are many rows that contain the NA value, we can't just remove them, since we would lose too much information.

Our approach to handling these missing values was to look at the other metrics associated to view_count, such as favourite_count, retweet_count and reply_count, where we calculated for each row the percentile of each metric and then averaged the calculated percentile of the interactions each post had for each HEI into a new attribute. With that averaged percentile we then calculated the view_count by multiplying the maximum view count each HEI had with the percentage acquired from the averaged percentile.

When analyzing the NA problem we also realized that one of the HEI, complutense, only had 1 one post in the entire dataset. Given the small value we cannot extract any information from it so we removed it from our data.

III. Data Processing

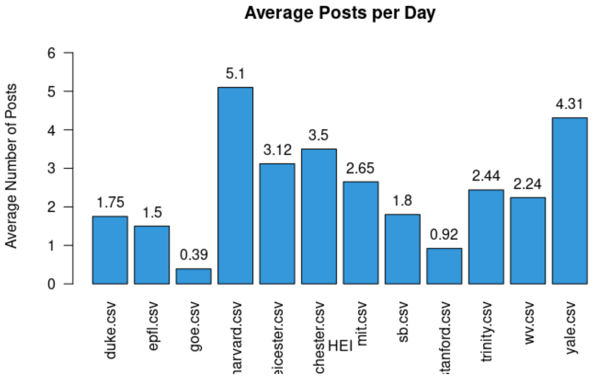
After the initial pre-processing, we utilized a number of tools and functions to visualize and get a better understanding on the different strategies between HEIs.

A. Average Number of Posts per Time frame

To analyze the posting strategies of HEIs during a specific timeframe we created a function named average_posts, whose purpose is to compute the average number of posts per given timeframe.

The first step in the function is to calculate for each HEI the date range. This can be done by summarizing the minimum and maximum date of the posts. Then the function creates a new column whose name is based on the timeframe it received, whose values are going to be the number of posts divided by the number specified in the timeframe. This timeframe can either be weeks or days.

From this function we created two new tables, posts_per_day and posts_per_week, and also a barplot for each table.



With the help of this function we can observe that almost all but one HEI, Goe, has an average post of one or near one post per day, with the maximum being Harvard with over five posts per day.

B. Average of Posts During Academic and Vacation Periods

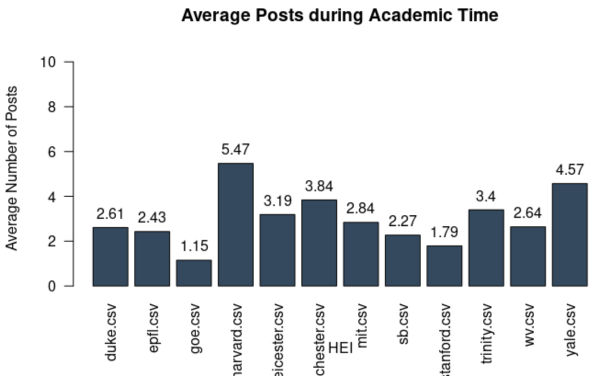
To understand the posting behavior of HEIs throughout the academic year and during vacation periods, we developed a series of functions to define specific time intervals and analyze the number of posts within these intervals.

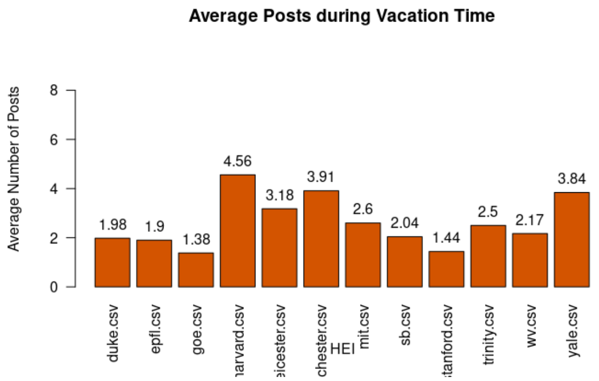
Firstly we defined some time interval corresponding to a generic academic year, with those being from 31/08/2022 to 15/12/2022, then 4/01/2023 to 1/04/2023 and finally 14/04/2023 to 15/06/2023.

With the intervals defined, we designed a function named check_interval that iterates through each interval checking if it falls into any of them. If it does then it returns TRUE, otherwise its FALSE. We applied this function to each post's creation date and created a new column academic_year that contains the boolean.

This new information allows, with the use of a new function, to filter all posts according to the vacation time or academic time. The function is analyze_posts and not only does it filter and count the posts, but it also counts the number of unique days.

These two counts are then used to calculate the average of posts per day for each period.





From these plots we can see that only two HEIs, Goe and Manchester, post more while its students are in vacation than during the academic year.

C. Highs and Lows of Posting

To gain insights into the posting strategies of each HEI we also analyzed the day and hour that each HEI prefers to posts.

Firstly we created a new table, data_posts_days that includes the day of the week and hour of the post. To do that we used the function weekdays that can extract the day of the week from the created_at attribute, and to retrieve the hour, we used the format function.

Then we grouped in another table number_posts_days, the posts by the id and day_of_week to count the number of posts of each day for all HEIs. With this table we use the functions slice_min and slice_max to retrieve for each day the HEI that has the highest and lowest count.

From these tables we can see that Harvard dominates the highest count, only losing one day to Manchester. In terms of lowest count all the weekdays are for Goe, but in the weekend its Stanford and Epfl.

To better understand the daily posting patterns, in another table we grouped the data by id, day_of_week and created_date to then count for each day the number of posts. We then found, for each day, the average post count per HEI. From this table, the only difference from the lowest count is that now Mit has the lowest average for the weekend.

D. Favorite Day and Hour

From the table created to know the count of posts of each day, we can also retrieve the day that each HEI decides to post more, for almost all HEIs that day is thursday. Despite some HEIs preferring another day, we can also observe that all HEIs fall in the interval between tuesday and thursday, in other words their favorite time falls during the middle of the week.

Then we decided to also view the favorite hour to post and observed that despite the values being spread more evenly throughout all HEIs, they all fall between 10 and 17, meaning that they are the most active during the middle of the day. To facilitate the clustering analysis we categorized the favorite posting hour into “Morning”, “Afternoon” and “Night” and assigned numerical values to these categories.

E. Heatmaps

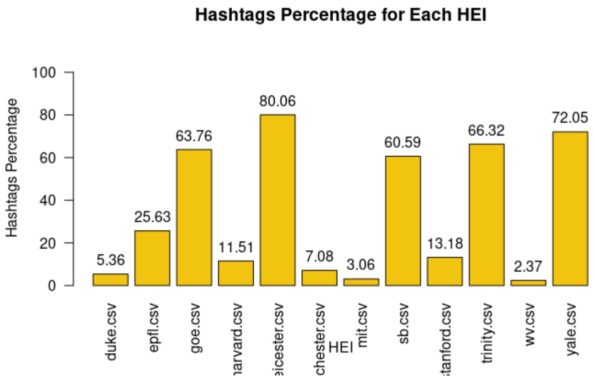
After analyzing the posts of each HEI, in relation to day and hour, we wondered why Goe didn't also have the lowest average count for the weekend, so we decided to view the data through a heatmap, using the table data_posts_days. With this we realized that Goe didn't have any post during the weekend, all of their posts were made during the week. By analyzing the heatmaps we also noticed, that almost no activity occurs during the weekend by any HEI with the outlier being Harvard.

F. Hashtags

In this section, we will be viewing the use of hashtags from each HEI. For that we looked at the number of unique hastags and the usage rate that each HEI had.

Before creating any table we need to transform every empty string that exists into a NA value. This step ensures that all posts without any hashtags can be properly identified. Then we created a table named hashtags that for each HEI includes the total number of posts, the number of posts without hashtags, the number of unique hashtags and with the first two counts we also calculated the percentage of posts with hashtags.

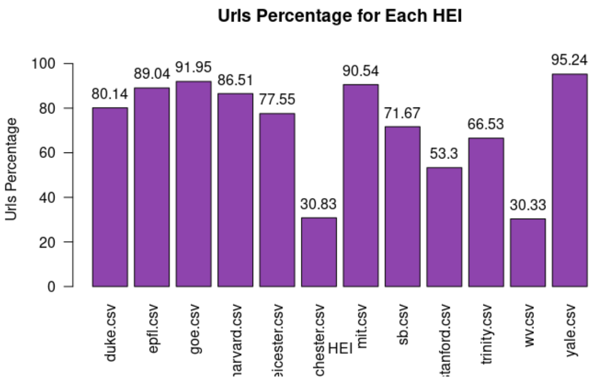
While plotting the barplot for the unique hashtags we came to the conclusion that the usage of unique hashtags is very varied between HEIs, with some passing the hundred mark, such as Epfl, Leicester, Sb, Yale and Trinity were the most creative.



In the percentage plot, we were surprised since we thought how Trinity had the most unique hashtags out of all that they could have the biggest usage percentage but its Leicester. On the other hand Wv, Mit and Duke don't employ many hashtags on their posts, which also explains why they didn't have many unique hashtags in the previous table.

G. Urls

For the usage of urls, our approach was similar to the creation of the hashtags table. Firstly we passed all empty values to NA, then for each HEI we counted the number of posts with and without a url. Afterwards we made the calculations and created a plot based on the values from the table url_usage that was were we stored all the previous values.

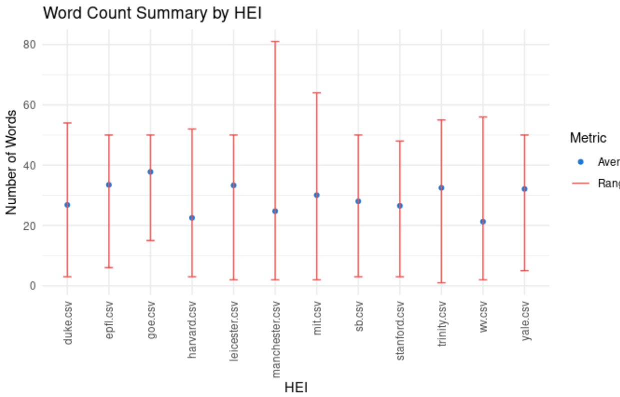


From this table we noticed that the HEIs that don't use many hashtags, use many urls, with an exception being Wv. Another HEI that doesn't use many urls either is Manchester, having very similar values to Wv. On the other hand, Yale, Mit, and Harvard all have very high percentages, surpassing the ninety mark.

H. Text Metrics

Lastly we decided to analyze the text content, specially focusing on the word count metric, including the average, minimum and maximum number of words.

We started by extracting from data_posts the columns we needed, which included the HEI identifier and the text of the posts, then for each text we counted the number of words in each text to a new table named data_posts_contents. From that table we grouped the data by HEI and calculated the average, minimum and maximum number of words per post, in data_posts_content_metrics. Then we plotted the information on that table to better visualize the results.



With the plot we could visualize that the average number of words in a posts for all HEIs falls between 20 and 40. Also almost all their ranges are very similar, with the outlier being Manchester, that has a maximum much higher than the second highest word count.

IV. POSTS CLASSIFICATION

Before classifying each post for a HEI we first passed every text related to each HEI through a function named corpus_maker that transforms the plain text into a VectorSource and then we apply the Corpus function to our text vector. Afterwards that corpus is given to a cleanup function provided by professor Álvaro Figueira, that for every text received passes all words to lowercase then removed all numbers and all english common stopWords with assistance from the tm_map function. It also removes all punctuation and extra white spaces, then we lemmatize the text in order to reduce dimensionality and improve consistency by ensuring that similar concepts are treated uniformly, this will also help when we use the technique of Term Frequency-Inverse Document Frequency (TF-IDF). This function can also receive a list of terms to remove if needed.

After this step we pass the "clean" text to the function freq_terms that created a DocumentTermMatrix and uses the TF-IDF to determine the term frequency for all terms. We then selected the most frequent terms and their values and visualized all of them, with this, we realized that some of them were too generic, and as such, they were removed via the cleanup function. With the terms that remained, we started to create a dictionary where we associated every word to one of five categories ("Image", "Education", "Research", "Society",
"Engagement").

With our dictionary done we created a function named classify_text that for each text received it searches for the frequent words in the text then sorts through all words and their categories and selects the most frequent one to categorize the text.

In order to categorize all posts we repeated this process for every HEI and gave the classify_text function all their posts.

V. Sentiment Analysis

In this section we will analyze the emotional tone and sentiment of the various posts made by the various HEIs, to do that we employed sentiment analysis techniques from the R library syuzhet.

Firstly, for each HEI, we calculated the sentiment score for each post and stored it into a new column named sentiment. Then combined the sentiment scores of all HEIs into one table named all_sentiments and calculated the average sentiment of each HEI.

To analyze the emotions in the texts we created a function named emotions_maker that uses the NRC sentiment lexicon, it calculates the proportion of each emotion and generates a horizontal bar plot.

From these plots we noticed that for all HEIs the highest emotions were either trust or anticipation, followed by joy. Also the emotion with less percentage in all HEIs is disgust.

VI. Cluster Analysis of HEIs Posting Strategies

In this section, we will be performing a clustering analysis on the various attributes we already saw in order to see how HEIs are grouped based on their similarities of posting strategies.

To prepare the data for clustering we combined various metrics from several tables into a single table named cluster_table. These metrics include the percentage of tweets, replies, hashtags, and URLs, as well as the average number of posts during different time periods also the average number of words per post and the average sentiment of posts.

Afterwards we created a function named cosine_matrix_maker that calculates the cosine similarities between metrics. Based on the results from the matrix we decided to remove the following metrics: average_num_words, unique_hashtags, avg_posts_per_weeks, avg_posts_in_academic_time and avg_posts_in_vacation_time. Since all had high similarity values to other attributes.

To then determine the optimal number of clusters for our k-means clustering, we used the elbow method. This method involves plotting the within-cluster sum of squares (WCSS) for different numbers of clusters and then we selected the number where the rate of WCSS sharply slows down (the "elbow point"). With the optimal number of clusters identified, four, we then applied again the k-means clustering method, then assigned each HEI to each cluster for better understanding.

VII. Conclusion

In this report we analyzed the social media strategies of the several HEIs on X/Twitter, focusing on posting behaviors, content patterns and engagement metrics. From the posting behaviors we learned that most HEIs post consistently, with Harvard being the most active, also that posts were typically made during the week with more posts being in the middle of the day. For the content patterns the usage of hashtags and urls were significantly varied, with some HEIs showing more creativity in their creation of hashtags. With the sentiment analysis we saw how trust and anticipation were the most common emotions, while disgust was the least expressed, indicating an overall positive communication tone.

Overall, this study underscores the importance of social media as a tool for communication and engagement for HEIs. The insights gained from this analysis can help them refine their social media strategies to enhance their outreach and engagement with their target audiences.