Analysis of Posting Strategies for Higher Education Institutions

1st Adriano Chessa
Department of Computer Science
University of Porto
Porto, Portugal
up201405297@up.pt

2nd Carlos Vilela
Department of Computer Science
University of Porto
Porto, Portugal
up202203836@up.pt

3rd Gabriel Guimarães
Department of Computer Science
University of Porto
Porto, Portugal
up201903627@up.pt

4th Pedro Leite
Department of Computer Science
University of Porto
Porto, Portugal
up201906697@up.pt

***Abstract*—Through an extensive data analysis of a collection of tweets from Higher Education Institutions (HEIs), this project explores different posting strategies. Data was visualized to find the post's frequency, length, interactions and content trends. This data was grouped by HEI, in order to see which HEIs have similar posting strategies, and these groups were further explored and explained. The posts were analyzed to a deeper extent, in their text content, where their most important words were extracted and the sentiments and emotions analyzed. Using different approaches, these posts are categorized with one of the predefined categories: "Image," "Education," "Research," "Society," or "Engagement". The findings provide insights into the social media strategies of HEIs and suggest methods for optimizing digital communication.**

***Index Terms*—Twitter, X, Social Media, University, HEI, Data Analysis, Cluster Analysis, Pattern Analysis, Content Analysis, TF-IDF, Sentiment Analysis, Emotion Recognition**

I. Introduction

Nowadays a digital presence has become imperative for any entity attempting to promote themselves. Social media platforms have risen to fill that need and currently represents one of the most used tools to acquire talent, faculty, promote novel research and overall achievements. This is especially true for organizations such as Higher Education Institutions (HEIs) and the usage of this tools has been on the rise for last decade [1]. Twitter (now known as X) is one of the most used platforms for this effect and the ability to find the optimal frequency, content and timing is crucial to improve influence, outreach and institutional visibility.

With this in mind, this project aims to analyse a sample of Tweets from various HEIs in order to understand the patterns and publication strategies of each entity trough exploratory data, clustering and content analysis. By developing this project, the aim is to improve the understanding of social media dynamics in the context of higher education, while also offering actionable insights for enhancing digital communication strategies that may be implemented in future projects.

II. Materials

The project was made using a Python Notebook, with the following packages: "Pandas", "Numpy", "Seaborn", "Matplotlib", "Sklearn, "Emoji", "String", "Nltk", "Nrclex", "Collections", "Text2emotion", "Wordcloud", "Transformers", "RE", "Plotly", "Textblob", "Spacy", "Networkx".

III. Pattern Analysis and HEI Grouping

A. Pre-Processing

The pre-processing of the data involved several steps to clean and enhance the dataset. First, the CSV file was read, and the ".csv" suffix was removed from the "id" column values. The "tweet_id" column was dropped as it did not provided actionable information. The "created_at" column was parsed to extract the "day_of_week" and "hour_of_day". New binary columns were added to indicate the presence or absence of URLs, media, photos, and videos in tweets. Additionally, the length of each tweet was calculated (spaces and punctuation included), and functions were implemented to count hashtags, emojis, and mentions within the tweet text. As the HEI "complutense" had only a single observation it was removed and therefore not considered for further analysis. After an initial analysis, a tweet for "mit" relating to the subscription to "twitter Blue" was removed, although referring to a real tweet, skewed the observations of five distinct variables. For the missing values in the "view_count" variable data imputation was performed, using the median of this variable for each HEI and attributed this value to the missing entries to better reflect reality.

B. Exploratory Data Analysis (EDA)

1) Univariate Analysis: When evaluating the variables independently, it becomes apparent that "bookmark_count," "favorite_count," "retweet_count," and "reply_count" exhibit high skewness values of 34.39, 22.86, 33.99, and 31.29, respectively. This indicates right skewness caused by a small number of tweets having disproportionately high interaction counts, as evidenced by the maximum values being substantially higher than the 75th percentile. Similar positive skewness is observed in ”hashtag_count” and ”emoji_count,” although not to the same extent. Regarding the posting patterns, as shown in Figure 1, tweets are concentrated primarily during the weekdays, especially on Thursdays, where the highest concentration of created tweets is found. In terms of the hour of the day, the trend indicates that tweets are mainly posted from 11 UTC to 16 UTC. This is further supported by the mode, as all HEIs are active during the following time frame: Tuesday to Thursday from 11 UTC to 14 UTC.

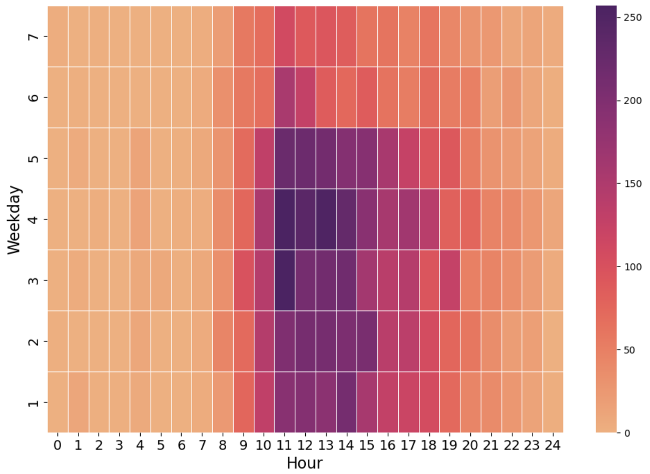

Fig. 1. Heatmap of Tweet Frequency by hour of day and Weekday.

2) Bivariate Analysis: For this analysis a correlation matrix was calculated with the average of each variable. The matrix can be seen in Figure 2 and with a superficial analysis it was possible to spot some trends that may indicate publication strategies. From this analysis the following observations can be made:

- Variables related to engagement, namely ”average_likes”, ”average_retweets”, ”average_views” and ”average_bookmarks” all show high correlation values, indicative that with more views will also lead to more likes, retweets and bookmarks.
- The length of the tweet led to negative correlations with the engagement metrics showing that bigger tweets tend to have less likes, replies and retweets. On the other hand, ”tweet_length” shows positive correlation with ”average_media” and ”average_urls” which can lead to the conclusion that bigger tweets will also contain a higher number or urls and media such as photos.
- ”tweet_count” showed a moderate positive correlation to ”average_views” and a moderate negative correlation with ”tweet_length” this interesting result may indicate that HEIs that tweet more often tend to get more views and tend to write shorter tweets.

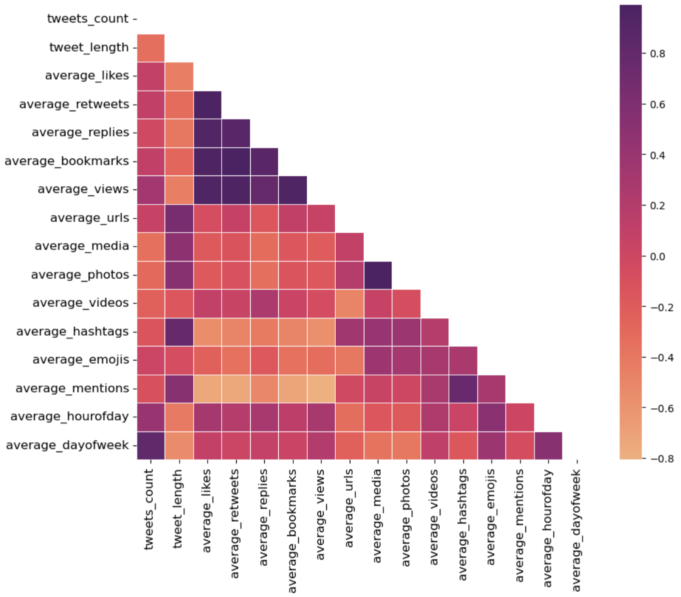

Fig. 2. Correlation matrix of average key metrics.

C. Clustering Analysis

This section focuses on grouping HEIs based on their Twitter publication strategy. By examining various attributes, the aim was to uncover patterns and trends that can inform strategic decisions. This process involves several key steps including data standardization, dimensionality reduction, and clustering, culminating in a comprehensive evaluation of the resulting groups.

One of the most important aspects of the process of grouping HEIs is the attributes selection, which is limited by the available data. This was achieved through a combination of engagement metrics (likes, retweets, replies, bookmarks, and views), the post’s characteristics (count and length), types of content (URLs, media, photos, videos, hashtags, mentions, and emojis), and time of posting (weekday and hour of the day).

The different scales of the variables can skew results and mislead interpretations. To mitigate that, a standardization procedure was done, using the Min-Max Scaling of the ”Sklearn” pre-processing package.

With all variables on the same scale, the dataset still has a high number of dimensions, due to the number of attributes. To lessen this high dimensionality, a new dataframe was built, only using the Principal Components (PCs) of the original variables. This was done using the ”Sklearn” decomposition package. At the end of the evaluation, only the first two PCs were kept.

Since this is an unsupervised learning task, a method was devised in order to find the perfect number of groups. The Elbow Method (a heuristic based technique) finds the perfect trade-off on the plot of the within-cluster sum of squares versus different cluster numbers, commonly called the ”elbow point”. The number of clusters chosen was four and the resulting clusters can be observed in Figure 3.

Finally, the grouping was made by applying the K-Means Cluster Algorithm, of the cluster package from ”Sklearn”. K-Means works by iteratively assigning data points to clusters based on proximity to centroids.

- Cluster 0 is composed by ”MIT”.
- Cluster 1 is composed by ”GOE”, ”Trinity” and ”Leicester”.
- Cluster 2 is composed by ”EPFL, ”SB”, ”Yale”, ”Duke”, ”WV” and ”Manchester”.
- Cluster 3 is composed by ”Stanford” and ”Harvard”.

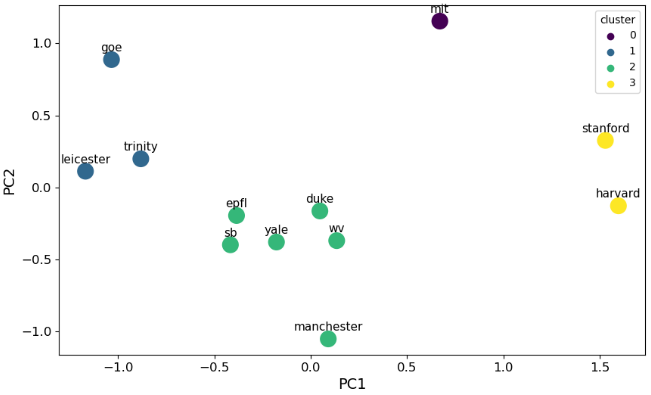

Fig. 3. K-means Clustering of Tweets based on the first and second Principal components. Each point represents an HEI and the clustering was obtained with optimal k of 4.

To evaluate the clustering results, the method Silhouette Coefficient was used, from the metrics package from ”Sklearn”. This method measures the compactness intra-cluster and separation inter-cluster, where the higher the score, the better the results. The obtained score was 0.542 which taking in account the low number of observations can be considered a good result.

In order to analyze what made each cluster, the clusters were grouped and the ranking of each attribute was extracted and evaluated.

- Cluster 0 HEIs: Uses a lot of media, mainly photos, but very little videos, uses very little hashtags and emojis, mentions the least other accounts, uses a lot of URLs, and tends to tweet later in the day.
- Cluster 1 HEIs: Tweets less frequently and the tweets are the biggest, have the less interactions and views, uses a lot of hashtags and emojis, but does not mention a lot other accounts, and tweets later in the week.
- Cluster 2 HEIs: They are average in all metrics, but the URLs usage, since they do not tend to use them.
- Cluster 3 HEIs: Tweets the most and the tweets are the smallest, have the most views and interactions, use very little media and photos, although they are the ones that use the most videos, mentions the most other accounts and uses the most URLs, and tweets sooner in the day and week.

IV. Content Analysis

A. Pre-Processing

In order to perform the top term frequency–inverse document frequency (TF-IDF) words extraction, sentiment analysis and emotion recognition, the text needs to be processed first. To do so, the following steps were performed:

- Every word was lowercased.
- Hashtags, URLs, emojis, mentions, punctuation, quotes and hyphens were removed.
- Stop words (like ”the”, ”is”, ”in”, etc) were also removed.
- Lemmatization, in order to transform the words to their root form.
- Trimmed extra spaces.

B. Top TF-IDF Words

To identify the most significant words in the tweets, the text was vectorized and the mean TF-IDF score was calculated, using ”Sklearn” text feature extraction package. These words were ranked, and only the top 200 were extracted for the future tasks which resulted in the world cloud visible in Figure 4.

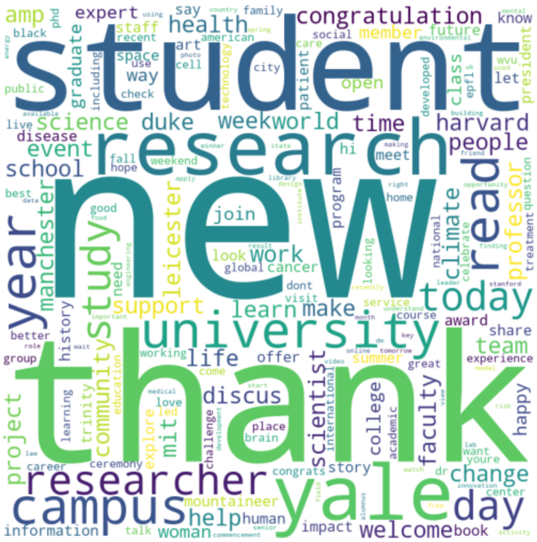

Fig. 4. Word Cloud of Top 200 Terms in analysed Tweets.

Using the package ”Wordcloud” and ”Matplotlib”, a plot was made to visualize the most significant words. ”Yale”, ”thank”, ”new, ”research” and ”student” seem to be the most important words in the dataset.

C. Sentiment Analysis

To analyze the sentiment of tweets, the ”Sentiment Intensity Analyzer” from the ”VADER” package was used. A function was defined to compute the sentiment value (a compound score ranging from -1, for extremely negative, to 1 for extremely positive). Another function categorized these sentiment values into ”negative” labels (sentiment value smaller then 0), ”neutral” labels (sentiment value equal to 0) or ”positive” labels (sentiment value bigger then 0). By applying these functions to the dataset, the sentimental tone of each tweet was quantified and labeled, providing a comprehensive view of the sentiments expressed in the textual content.

To visualize the most significant words associated with different sentiments in tweets, a word cloud was generated for each sentiment. For the negative tweets, ”new”, ”dr” and ”epfl” were the most relevant words. For the neutral tweets, ”new”, ”learn, ”explore” and ”epfl”. And for the positive tweets, ”new”, ”year”, ”learn” and ”award”.

To visualize the likes and views associated with each sentiment, two bar plots were made. With the plots visible in Figure 5 it is possible to observe that ”neutral” tweets have on average more views and likes, ”negative” tweets have more views than ”positive” ones, but less likes.

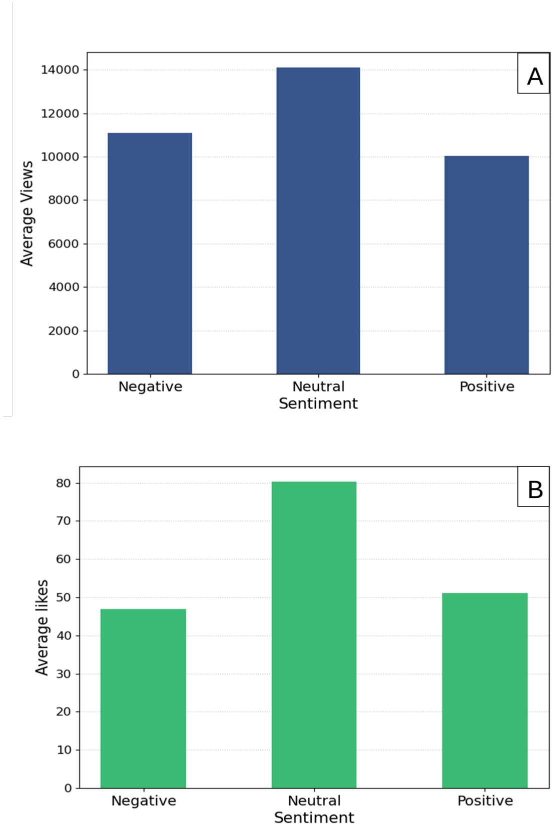

Fig. 5. Bar plots representing distribution of Tweets per sentiment. A: Average views of tweets per sentiment. B: Average likes of tweets per sentiment.

D. Emotion Recognition

For emotion recognition, a pre-trained transformer-based model was used. It utilizes the RoBERTa (A Robustly Optimized BERT Pretraining Approach) architecture, which is a variant of the BERT (Bidirectional Encoder Representations from Transformers) model. The model utilized is a powerful tool for emotion detection in textual data, designed to identify a wide range of emotions. The pipeline in the Hugging Face library provides a high-level interface for various natural language processing (NLP) tasks. When a piece of text is input into the pipeline, the following steps are performed:

- Tokenization: The input text is tokenized into subword units compatible with the RoBERTa model.
- Encoding: The tokens are converted into numerical representations (embeddings) that capture the contextual meaning of the words.
- Inference: The encoded representations are passed through the RoBERTa model, which processes the input and generates predictions for each emotion category.
- Output: The pipeline outputs the predicted emotions along with their confidence scores.

Emotions for the cleaned text were estimated, resulting in the identification of 28 distinct emotions. The most frequent emotions found were neutral, admiration, gratitude, approval, excitement, joy, caring, optimism, fear, realization, and love.

This comprehensive emotion detection offers a nuanced understanding of the emotional undertones present in the tweets, providing a deeper layer of analysis beyond basic sentiment. By combining sentiment analysis and emotion recognition, a richer, more detailed picture of feelings expressed in tweets is achieved. This dual approach enhances the ability to derive meaningful insights from social media data.

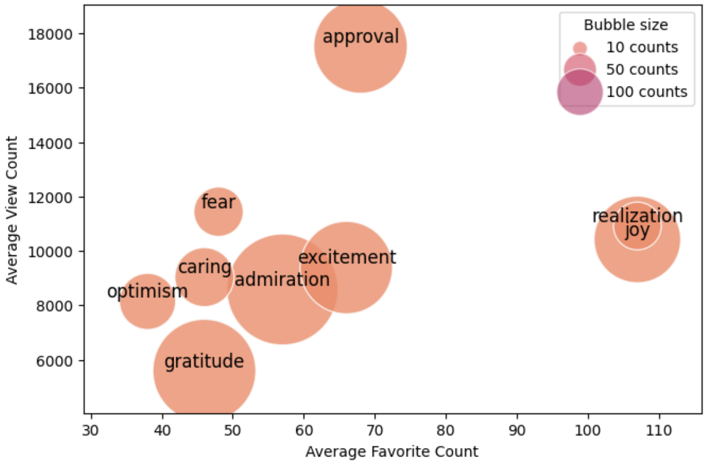

Fig. 6. Scatter Plot of the Average Favorite Count vs. Average View Count by Emotion, excluding Neutral. The Size of the Circles represents the frequency.

The scatter plot shown in Figure 6 highlights the performance of different types of emotions present in the tweets in terms of average views and average likes. Each bubble represents a different emotion, with the size of the bubble indicating the volume of tweets. As observed, certain emotions such as ’gratitude’ and ’admiration’, despite frequent have the lowest engagement metrics. Although ’approval’ has the most average views, certain emotions, particularly ’realization’, and ’joy’, are more effective at converting views into favorites. HEIs can leverage these insights to craft emotionally resonant content that maximizes engagement on social media.

E. Category Identification

The goal was to assign one of the following categories to each tweet: Image (self promotion), Research, Education, Engagement and Society. To do so, two different approaches were used.

1) First Approach: This approach uses K-Means Clustering, to group the tweets according to the top TF-IDF words usage. A manual assignment is then done, to identify each cluster with the appropriate category.

A new dataframe was built, where every column represents one of the top 200 TF-IDF words, calculated previously. The values of these columns are the count of those words per tweet. Because of the high number of attributes, a PC Analysis was made, creating a new dataframe with only two PCs. K-Means Clustering was then applied (Figure 7) in the PCA-transformed dataframe, with five clusters, each cluster representing a category. The clustering got a silhouette coefficient score of 0.819, a very positive result.

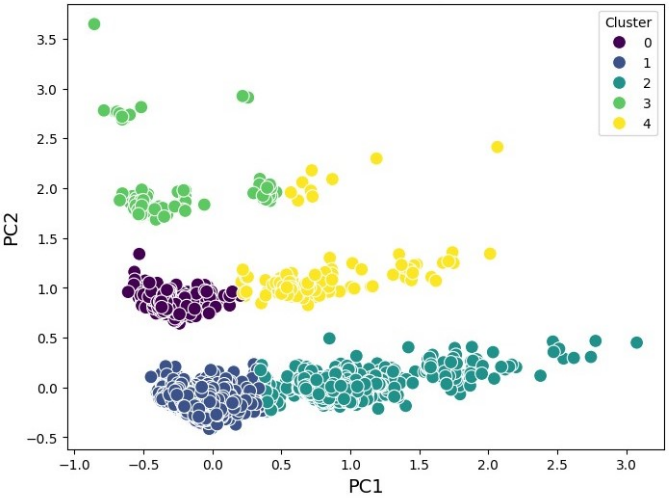

Fig. 7. K-Means Clustering of Principal Components. Color of each point corresponds to the cluster it belongs and and the clustering was obtained with optimal k of 5.

To categorize each cluster, the top emotions, sentiments and TF-IDF words were extracted, for each one. Additionally, random tweets were selected, from each cluster for further analysis. And the manual assignment was made.

2) Second Approach: The second approach employs the TF-IDF method to identify the top words for each HEI instead of directly categorizing the tweets, the focus is on categorizing the words. A function was then developed to use this word categorization to classify the tweets.

Given that each HEI has unique posting strategies, the top 20 TF-IDF words were calculated separately for each HEI and compiled into a comprehensive list. This list was then manually analyzed, with each word assigned to one of the predefined categories, resulting in a new dictionary. This dictionary was utilized to determine the presence of the most common words from each category within the tweets. The categorization of tweets was accomplished by counting the occurrence of these words in each tweet and assigning the tweet to the category with the highest word count.

To extend the approach, some additional steps were undertaken to enhance the categorization process. Firstly, the top 100 words for each HEI were identified using the TF-IDF method. This expands the vocabulary considered for categorization, making it more comprehensive.

Next, a pre-trained word vector from spaCy was loaded to provide a richer context for the words. These word vectors enable the calculation of similarities between words, which can be leveraged to assign words to the appropriate categories more accurately.

Additionally, a network was created (Example of which can be seen in Figure 8) by analyzing the co-occurrence of words within tweets, which reveals how often words appear together. The employment of network metrics allowed finding communities by the modularity classes.

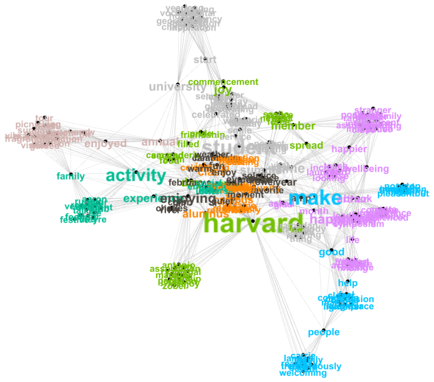

Fig. 8. Network based on co-occurrence of word within each tweet. Example of Harvard tweets.

Each community is then assigned to a category based on the cosine similarity of the words, ensuring that words within a community are semantically related. Finally, tweets are classified based on the word communities they contain.

V. Conclusion

By developing this project we managed to create a successful pipeline that can be used to analyse a given dataset and extract relevant information and reveal trends not visible at first glance. We were able to infer posting trends and social media strategies utilised by various HEIs to maximise their influence and reach. This project also allowed for sentiment and emotion analysis that revealed which features have a bigger impact on the general population. In future work this project would benefit of a machine learning model that could accurately predict the category of features posts and the developed pipeline could be tested with alternative datasets so its use may be further widened.

References

[1] Malik, A., Heyman-Schrum, C., & Johri, A. (2019). Use of Twitter across educational settings: A review of the literature. International Journal of Educational Technology in Higher Education, 16(1), 36. https://doi.org/10.1186/s41239-019-0166-x