Article

Post, Predict, and Rank: Exploring the Relationship Between Social Media Strategy and Higher Education Institution Rankings

Bruna Rocha [1,†] and Álvaro Figueira [1,2,*,†]

1   Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal
2   INESC TEC, Rua Roberto Frias, 4200-465 Porto, Portugal
*   Correspondence: arfiguei@fc.up.pt
†   Current address: DCC-FCUP, Rua do Campo Alegre s/n, 4169-007 Porto, Portugal.

Abstract: In today's competitive higher education sector, institutions increasingly rely on international rankings to secure financial resources, attract top-tier talent, and elevate their global reputation. Simultaneously, these universities have expanded their presence on social media, utilizing sophisticated posting strategies to disseminate information and boost recognition and engagement. This study examines the relationship between higher education institutions' (HEIs') rankings and their social media posting strategies. We gathered and analyzed publications from 18 HEIs featured in a consolidated ranking system, examining various features of their social media posts. To better understand these strategies, we categorized the posts into five predefined topics—engagement, research, image, society, and education. This categorization, combined with Long Short-Term Memory (LSTM) and a Random Forest (RF) algorithm, was utilized to predict social media output in the last five days of each month, achieving successful results. This paper further explores how variations in these social media strategies correlate with the rankings of HEIs. Our findings suggest a nuanced interaction between social media engagement and the perceived prestige of HEIs.

Keywords: higher education institutions; ranking system analysis; text mining; machine learning; topic modeling; prediction analysis

1. Introduction

As the global academic environment evolves, understanding the relative positions of higher education institutions (HEIs) has become increasingly significant, suggesting the creation of structured external assessments. In this context, global university rankings offer essential insights into the performance, reputation, and overall standing of HEIs [1–3]. These rankings have become more sophisticated, employing a range of metrics and methodologies to assess institutions with enhanced accuracy [4]. Typical evaluation criteria include academic accomplishments, research output, international presence, and industry revenue. Some prominent ranking systems such as Times Higher Education (THE) and QS World University Ranking (QS) exemplify the diversity and impact of these assessments. The varied approaches in ranking methodologies underscore their relevance as decision-making tools, aiding stakeholders in making informed decisions about collaborations and strategic partnerships, thereby promoting innovation and excellence in academia.

Alongside traditional rankings, social media has emerged as a critical platform for universities to connect with prospective students, alumni, and potential investors [5].

As audiences increasingly shift towards social media platforms, HEIs have adapted by developing new communication strategies tailored to these platforms [6]. These strategies aim to achieve specific institutional objectives, such as increasing visibility, engaging a wider community, and ultimately influencing enrollment numbers and funding prospects. Effective social media presence can greatly impact an institution's ability to attract new students and obtain financial support by enhancing its accessibility and relatability, key factors for appealing to new students and stakeholders. Therefore, examining the relationship between successful social media strategies and traditional metrics of university success is crucial.

Despite universities having common objectives, their social media strategies can differ significantly, resulting in varied results. For these strategies to be successful, social media initiatives must be aligned with and supported by the institution's overall strategic management objectives [7]. Effectively categorizing social media topics ensures that content aligns with the institution's priorities while meeting the expectations and interests of key stakeholders, including prospective students, alumni, and research collaborators. When social media strategies are aligned with broader organizational goals, HEIs can achieve more focused and impactful communication efforts. This strategic alignment enhances the effectiveness of social media activities and allows for a more comprehensive analysis of how different HEIs perform concerning traditional success metrics like rankings. By examining these categorizations and their effects, institutions and those managing other educational organizations can compare their social media strategies and outcomes, providing valuable insights into the role of content strategy in achieving academic success.

In this work, we explore the relationship between the social media strategies employed by HEIs and their positions in global rankings. The layout of this paper is presented as follows. In Section 3 (Towards a Universal Ranking Standard), we delve into the search for a universal ranking system by analyzing the similarities among five existing rankings, aiming to identify the one that most accurately represents the overall standings of institutions. After selecting the world ranking system, we chose 18 HEIs. Sections 4 and 5 (Data Collection and Preprocessing, and Exploratory Data Analysis) present the data collection process and an exploratory data analysis to uncover underlying patterns and insights within the dataset. Further, Section 6 (Post Categorization) focuses on categorizing social media posts into five distinct topics—engagement, research, image, society, and education—utilizing BERTopic for initial topic attribution followed by manual refinement to enhance accuracy. Subsequently, we employ machine learning models, including Random Forest and Long Short-Term Memory (LSTM), to predict the number of posts in the last five days of each month for each HEI. This methodological approach allows us to assess the effectiveness of various social media strategies in influencing institutional rankings.

2. Related Work

In recent years, social media has emerged as a powerful communication, marketing, and engagement platform, transforming how organizations, including educational institutions, interact with their audiences [8]. This evolution has led to different studies on how social media works and affects various sectors, particularly within higher education. These studies reveal diverse approaches to optimizing social media use in the educational sector, highlighting both the opportunities and challenges presented.

As we consider the impact of social networks on educational institutions, it becomes important to determine whether there is a significant correlation between social media publication strategies and global university rankings; for example, do top-ranked institutions employ different strategies in their communication compared to other institutions? Different rankings employ different metrics to evaluate higher education institutions, resulting

in discordant rankings for the same institutions. In this work, we use similarity metrics to assess the level of agreement among these rankings. Shehatta et al. [9] investigated the correlation between various ranking systems, using the number of overlapping elements and Pearson's/Spearman's correlation coefficients. Çakir et al. [10] conducted a systematic comparison of national and global university ranking systems in terms of their indicators, coverage, and ranking results, applying the Aguillo et al. [11] methodology for calculating the similarity among two rankings, the Inverse Rank (*M*) measure. Further extending this work, Figueira et al. [12] took a further step by employing clustering techniques to group rankings based on their similarity, providing a subtle understanding of the robustness and consistency of rankings and publishing strategies.

In addition, BERTopic has become a versatile and accurate topic-mining technique for categorizing social media content and can, therefore, be used to detect latent topics in the posts. Mendonça et al. [13] showcased the modularity of this model, demonstrating how it can be adapted and fine-tuned, while de Groot et al. [14] and Egger et al. [15] explored the integration of deep learning technologies like BERT into the topic classification process. In particular, BERTopic has been effectively applied to analyze social media content, as shown by Futterer et al. [16], Grigore et al. [17], and Schneider et al. [18], who utilized BERTopic to examine trends, public sentiment, and specific communication patterns on platforms such as Twitter. Antypas et al. [19] also used BERTopic to uncover trends and topics in real-time, demonstrating its effectiveness in extracting and interpreting themes from the text corpus in diverse applications.

Predictive modeling using Long-Short-Term Memory (LSTM) networks and Random Forest (RF) algorithms has gained significant attention for their effectiveness in complex predictive tasks across various domains. Lasri et al. [20] demonstrated how self-attention-based Bi-LSTM networks can enhance predictive accuracy in sentiment analysis of social media content, specifically analyzing tweets related to distance learning in higher education institutions (HEIs). Similarly, Pandey et al. [21] developed a model to detect sarcasm in code-mixed social media posts, enhancing the accuracy of sentiment analysis in multilingual online communications. For instance, Nti et al. [22] used RF along with other machine learning algorithms, to predict the effects of social media on students' academic performance, while Hooda et al. [23] used the algorithm to assess and provide feedback on student performance, improving student success in higher education.

This small but illustrative selection of research shows that increasing attention is being paid to the relationship between social media strategies and the performance and visibility of higher education institutions. The clustering of HEIs based on rankings and the detection of latent topics in social media posts offer valuable insights into how these institutions communicate, engage with their audiences, and position themselves globally. By decoding these strategies, researchers can uncover patterns and approaches that can in turn lead to more effective communication and engagement strategies.

3. Towards a Universal Ranking Standard

In the landscape of global university rankings, it is essential to recognize the diverse opinions and methodologies that shape these evaluations. Each ranking system utilizes distinct objectives, criteria, and data sources, resulting in differing positions for the same institutions [24]. This variability highlights the challenges of directly comparing universities and achieving a consensus on their performance. Consequently, a distinction approach to interpretation is necessary. To identify which rankings are most similar, we selected five world university rankings for a comparative analysis of the positions held by higher education institutions across these lists.

3.1. World University Rankings

Five university rankings were considered in this study:

- Times Higher Education (September 2023) [25];
- Quacquarelli Symonds (June 2023) [26];
- Center for World University Rankings (May 2023) [27];
- Academic Ranking of World Universities (August 2023) [28];
- Webometrics Ranking of World Universities (July 2023) [29].

The Times Higher Education (THE) World University Rankings is an annual publication by Times Higher Education magazine. From 2004 to 2009, THE collaborated with Quacquarelli Symonds (QS) to release the joint THE-QS World University Rankings. Following their separation, both organizations began producing their independent ranking systems. THE rankings evaluate institutions based on 17 performance indicators, which are categorized into five main areas: Teaching (29.5%), Research Environment (29%), Research Quality (30%), International Outlook (7.5%), and Industry Income (4%). This comprehensive assessment covers over 2500 HEIs across various countries and regions.

In comparison, the QS ranking features 1500 HEIs across 104 countries. In 2023, three new metrics were incorporated into the methodology: Sustainability (5%), Employment Outcomes (5%), and International Research Network (5%). The remaining criteria include Academic Reputation (30%), Employer Reputation (15%), Faculty–Student Ratio (10%), Citations per Faculty (20%), International Faculty Ratio (5%), and International Student Ratio (5%).

The Academic Ranking of World Universities (ARWU), commonly known as the Shanghai Ranking, is one of the most established and respected global university rankings. Initially developed by the Center for World-Class Universities (CWCU) at Shanghai Jiao Tong University, ARWU has been published by Shanghai Ranking Consultancy since 2009. This ranking system uses six main indicators, namely, quality of education (measured by the number of alumni winning Nobel Prizes and Fields Medals, 10%), number of staff elements winning awards (20%), number of highly cited researchers selected by Clarivate (20%), highly cited researchers identified by Clarivate (20%), publications in the Science Citation Index-Expanded and Social Science Citation Index (20%), and per capita academic performance (10%). Annually, ARWU evaluates over 2500 HEIs, ranking the top 1000 institutions globally.

The Center for World University Rankings (CWUR) stands out by assessing the quality of education, alumni employment, faculty excellence, and research performance without relying on surveys and university data submissions. CWUR uses seven indicators grouped into four categories, such as Education (25%), Employability (25%), Faculty (10%), and Research (40%). This ranking provides a comprehensive assessment of 2000 universities across 108 countries and regions.

Lastly, the Webometrics Ranking of World Universities (WEB) is published by the Cybermetrics Lab, a research group within the Consejo Superior de Investigaciones Científicas (CSIC), the largest Spanish public research organization. To evaluate the different HEIs, this ranking uses three indicators, namely the normalized and average value of the number of external networks (subnets) linking to the institution's webpages (50%), the presence of top-cited researchers in Google Scholar profiles, and the citation count of top papers (40%). In 2023, WEB ranked 12,000 HEIs, offering a broad perspective on global university web presence and impact.

3.2. Similarity Metrics

Three similarity measures were selected to ensure an objective comparison among the previous rankings in Section 3.1, each with varying degrees of coverage. These measures are Overlap Size (*OC*), Spearman's footrule (*F*), and *M* measure [30].

The Overlap Size (*OC*) counts the number of shared elements between the two top-k lists. Despite its simplicity, it allows us to analyze the number of elements present in each top-k list and helps to interpret and analyze the other two metrics.

The Spearman's footrule (*F*) compares two ranked lists where the items in both lists are identical. However, in this case, the two top-k lists have different elements, meaning that some items only appear in one of them. So, to overcome this difficulty, after identifying and removing the non-overlapping items of the lists, a new relative rank is assigned to each item in both lists based on the number of elements left. For example, given a ranking $\{A, D, C\}$, where A and C are the only elements present in some other ranking, we remove element D, leading to the new ranking simply being $\{A, C\}$. Here, A is now in position 1 and C in position 2 instead of 3.

The result of the re-rankings is two permutations σ_1 and σ_2 on top-Z, where $|Z|$ is the number of overlapping items. Spearman's footrule is calculated based on these transformations [31,32] as

$$Fr^{|Z|}(\sigma_1, \sigma_2) = \sum_{i=1}^{|Z|} |(\sigma_1(i) - \sigma_2(i))|. \tag{1}$$

This sum is equal to 0 if and only if the two reshaped lists are identical. Naturally, it is only defined if the number of overlapping elements is greater or equal to two. Its maximum value is given by $1/2 \times |Z|^2$ if Z is even and $1/2 \times (|Z| + 1)(|Z| - 1)$ if Z is odd. We can then compute the normalized Spearman's footrule as

$$NFr = \frac{Fr^{(|Z|)}}{\max Fr^{(|Z|)}}. \tag{2}$$

This quantity is normalized is normalized by dividing it by its maximum value ensuring that it ranges from 0 to 1. A result of 0 indicates identical lists, while a score of 1 signifies completely distinct lists. Hence, for calculating the desired similarity value, *F* is defined as:

$$F = 1 - NFr \tag{3}$$

Another metric used to evaluate the similarity between ranked lists is the Inverse Rank (*M*) measure. The *M* measure assigns greater weight to matches and non-matches that appear higher in the rankings, ensuring that top-ranked elements have a more significant impact on the similarity score. This approach provides a subtle comparison by emphasizing the importance of the positions of elements within the lists [30]. The normalized value of the *M* measure is calculated using the following formula:

$$M^{(k)} = 1 - \frac{N^{(k)}}{\max N^{(k)}} \tag{4}$$

where

$$\max N^{(k)} = 2 \sum_{i=1}^{k} \left(\frac{1}{i} - \frac{1}{k+1} \right) \tag{5}$$

and

$$N^{(k)}(\sigma_1 \sigma_2) = \sum_{i \in Z} \left| \frac{1}{\sigma_1(i)} - \frac{1}{\sigma_2(i)} \right| + \sum_{i \in S} \left| \frac{1}{\sigma_1(i)} - \frac{1}{(k+1)} \right| + \sum_{i \in T} \left| \frac{1}{\sigma_2(i)} - \frac{1}{(k+1)} \right|. \tag{6}$$

$\sigma(i)$ represents the rank of item i in a list, k the length of the ranked lists, Z the set of common elements among the two top-k lists, S the set of elements unique to the first top-k list and T the set of elements unique to the second top-k list.

In the selected university rankings, tied positions are observed, and sometimes, rankings below a certain threshold do not provide individual ranks for all institutions. For instance, ARWU groups HEIs into sets of 50 past the top 100, without providing individual distinction. In these situations, the mid position [11] is employed, which involves

$$\text{Mid Position} = \frac{2k + (n-1)}{2} \tag{7}$$

where k corresponds to the tied position and n is the number of elements that are tied.

Results

In assessing global university performance, certain thresholds within rankings often represent critical benchmarks for institutions and prospective students. To explore these central segments, our research focused on the top 10, 100, and 200 of five world university rankings published in 2023, namely QS, THE, ARWU, CWUR, and WEB.

The similarity metrics across various rankings generally increase as longer lists are considered. Most results exceed the 50% threshold, with only a few occurrences in the top 10 rankings falling below this mark for the M measure. Despite the relatively small sample size, the overlap count consistently surpasses half of the analyzed sample. However, the M measure reveals lower similarity results, suggesting that while the institutions remain consistent across rankings, their specific positions vary. This indicates a significant consensus among the top 10 institutions regardless of the differing metrics used for evaluation.

It is important to highlight that the Spearman's footrule metric presents the worst result when comparing QS and ARWU rankings, with a value of 0.17. This result arises from the HEIs shared by both rankings, encompassing the top five positions. Consequently, after reshaping the rankings between the overlapping elements, the substantial difference between the positions among those five contributes to the significantly low metric value. Additionally, visualizing data through heatmaps (Figures 1 and 2) illustrates that for a sample of 200 HEIs, the QS presents the worst results for the M measure, as indicated by its bright colors. Among the four comparisons, QS achieves its highest result at 0.64 when compared to THE, while the other comparisons yield low positive values. Consequently, we conclude that QS is the ranking with the lowest similarity.
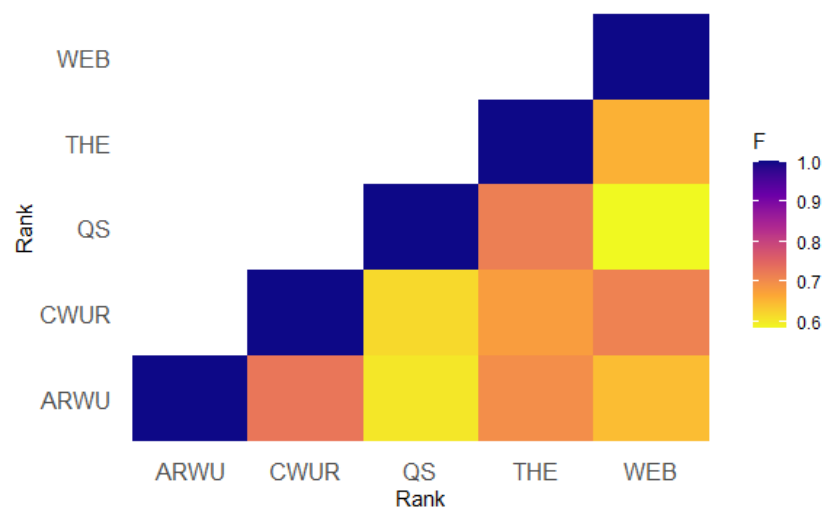

Figure 1. Spearman's footrule F for the top 200.
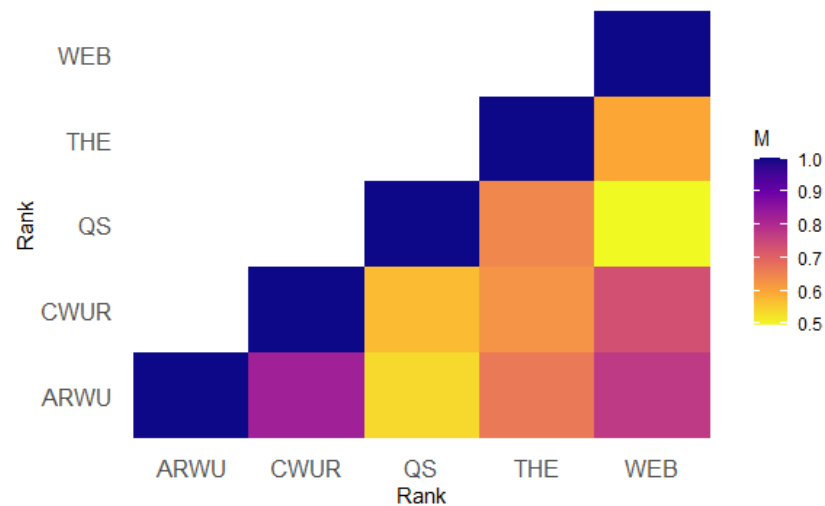
Figure 2. Inverse Rank *M* for the top 200.

The THE ranking stands out for its impressive coherence compared to the other four rankings, as evidenced by its values consistently falling within the range of 0.60 to 0.70 for the Spearman's footrule and Inverse Rank metrics. This pattern is evident in the consistent color variations within the same range in the heatmaps visualization, distinguishing THE from the other rankings and emphasizing its constant in evaluating higher education institutions. On the other hand, the WEB ranking presents a similar behavior exhibiting more favorable correlations with the ARWU and CWUR rankings, presenting slightly better results than THE. Also, it consistently achieves average similarity scores above 60% for the *M* measure across different top rankings analyzed. Therefore, when considering the importance of the Inverse Rank metric, which accounts for the overlapping and non-overlapping institutions, it becomes evident that WEB is slightly better than THE.

For the three similarity measures, ARWU and CWUR demonstrate the highest similarity, particularly in the top 100 and 200 rankings. In the top 100, these two rankings share 78 institutions in common, while in the top 200, they share 171 institutions, resulting in a similarity match of 0.83 when considering both overlapping and non-overlapping elements. In addition, the average result of the M measure of the top 200 in CWUR is 68.8% and ARWU is 70.0%, indicating a high degree of similarity between these rankings across the entire batch of data.

The world university ranking with the best result based on the different similarity measures used is the ARWU. However, it is important to note that the ARWU does not individually classify all institutions below position 100, which means that the result obtained was greatly influenced by the method chosen to handle tied positions in this investigation (middle position). Therefore, these findings should be interpreted with this consideration in mind. We thus determine the ranking with the best similarity is the ARWU with a median overlap of 145 HEIs. This high level of agreement indicates that the typical performance level of institutions is recognized despite the varying criteria used by different ranking systems.

4. Data Collection and Preprocessing

This similarity analysis aims to determine the world university ranking that best reflects the overall positioning of HEIs. This is used as a reference to characterize different editorial standards from the HEIs by comparing their respective publishing strategies. Since the previous results demonstrated that the ARWU and CWUR rankings are quite

similar, we have chosen to use the CWUR for this purpose. This is especially beneficial because the CWUR does not allow for ties between the institutions under analysis, allowing a clearer distinction.

In this research, we focus on identifying publication patterns from different institutions and relating them to their position in global rankings. Due to the scope of the study, it was not feasible to include every university. Alternatively, a strategic subset of HEIs was chosen to reflect overall trends by comparing institutions from diverse countries with similar geopolitical contexts and comparable positions in the rankings, allowing for meaningful comparisons. The selected institutions are listed in Table 1.

Table 1. Position of the chosen HEIs in CWUR as published in 2023.

Higher Education Institution	CWUR
Harvard University	1
Massachusetts Institute of Technology	2
Stanford University	3
University of Cambridge	4
University of Oxford	5
Yale University	10
University of California, Los Angeles	18
Duke University	20
University of Manchester	50
École Polytechnique Fédérale de Lausanne	96
University of Göttingen	97
University of California, Santa Barbara	98
Trinity College Dublin	247
University of Leicester	248
Complutense University of Madrid	249
University of Porto	309
University of Coimbra	420
West Virginia University	501

In gathering data for this analysis, we considered not only the publication outputs of each HEI from the social media platform X (formerly known as Twitter) but also the broader operational context in which these outputs occur. The academic year encompasses more than the start and end of classes; it includes significant events such as commencement activities, which not only mark the conclusion of academic journeys but also serve as crucial engagement points with stakeholders such as alumni, local communities, and potential employers. The inclusion of these events adds insight to our understanding of each institution and its impact on communication strategies. Our selection criteria were tailored to accommodate the distinct academic schedules of each institution. For example, in the U.S., the academic year can be extended to include an optional summer term, leading to three or four academic terms, while EPFL in Switzerland offers a unique summer program focused on hands-on laboratory experience, fostering connections between students and the scientific community [33].

To ensure a comprehensive analysis, the academic calendars of the selected institutions were considered, as cited in [34–51]. These dates are visually represented in Figure 3, which highlights the distinct timelines across the analyzed institutions.

To capture the full academic cycle, including preparations for the start of the year and various institutional activities, the time window selected for analysis is between August 2022 and August 2023. Then, it enables a complete observation of university operations, from welcoming new students to the culmination of the academic year.

Since some selected HEIs are located in non-English-speaking countries, publications in other languages were translated into English to ensure consistency in the analysis. Since the data consisted of tweets, preserving their full context for accurate embeddings was

crucial. Therefore, specific preprocessing steps were applied, including the removal of @mentions, URLs, hashtags, and formatting elements such as italics and bold text.
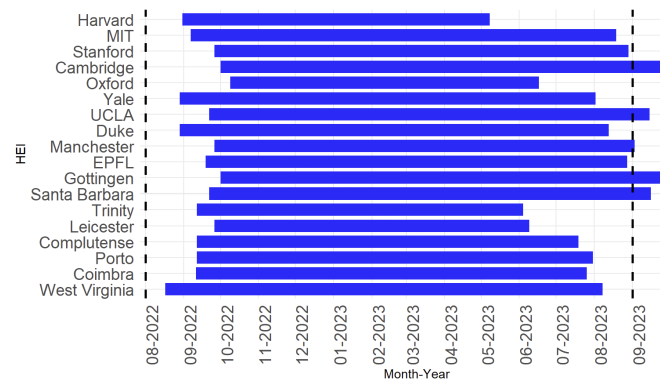

Figure 3. Academic calendars of selected higher education institutions.

5. Exploratory Data Analysis

Following the selection of the analysis period, our analysis aims to uncover patterns and key trends within the posts, providing insights into each institution's communication patterns and engagement levels. The analysis of the data in Figures 4 and 5 reveals that the University of Oxford leads in tweet frequency, significantly surpassing institutions like the Complutense University of Madrid, which shows distinctly lower activity. Stanford University, despite its high global ranking, exhibits a unique behavior with only 351 tweets during the analysis period, making it the second least active institution and the only one not surpassing 1000 tweets for the top five. Across most universities, the majority of posts are original tweets rather than replies or retweets, indicating a strategy that emphasizes broadcasting information over active engagement. MIT and the University of Porto are notable exceptions, each engaging minimally through replies and retweets, respectively, with the University of Manchester standing out by allocating 55% of its posts to replies, the highest proportion among the analyzed institutions. Additionally, only four universities, the University of Porto, the University of California Santa Barbara, the University of Göttingen, and EPFL, utilize all three types of posts, demonstrating a more balanced communication approach.
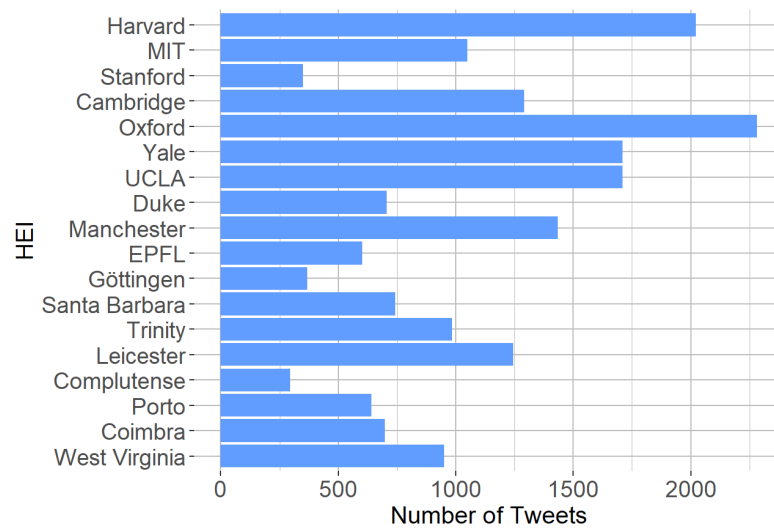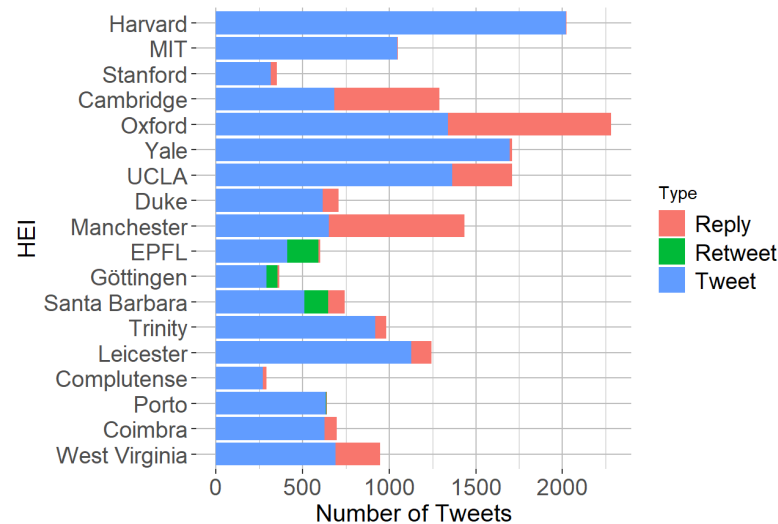

Figure 4. Data distribution per HEI.

Figure 5. Engagement types per HEI.

Distinct temporal patterns emerge across institutions. American HEIs generally reduce their publication frequency between October and December, likely correlating with midterm elections. They exhibit increased activity during key periods such as open house events for admitted students in March and the commencement of academic activities from May to June. Similarly, the University of Göttingen, EPFL, and the Complutense University of Madrid showed increased posting activity in March, May, and June, aligning with significant academic milestones. On the other hand, the University of Cambridge concentrated its highest posting frequency in August of both 2022 and 2023, coinciding with important events like A-Level Results Day to engage prospective students effectively. UCLA also experiences a notable peak in March, driven by interactive tweets that enhance follower engagement. Harvard maintains a consistently high average posting frequency throughout the year, with peaks in October and March highlighting strategic communication efforts during key academic events.

Analysis of posting times, adjusted for each institution's time zone, indicates that most universities prefer traditional working hours, though there are variations. Yale and Manchester favor morning posts, while Trinity and Leicester opt for the afternoon. MIT distinguishes itself with a narrower posting window between 11 a.m. and 3 p.m. on weekdays and occasional early morning posts between 2 a.m. and 6 a.m., diverging from typical American institutions' schedules. Additionally, institutions like Duke avoid late-night postings, whereas Complutense University maintains a steady post rate during these hours.

Engagement metrics, measured through retweets and favorites, indicate that higher-ranked institutions like Harvard, MIT, and Stanford achieve the highest median counts in both categories. These universities also display wide variations in engagement levels, indicating diverse interactions with their audiences. While the top five institutions enjoy significantly higher favorite counts, other universities exhibit a broader range of engagement. Notably, outliers such as the Complutense University of Madrid and West Virginia University demonstrate unique engagement patterns despite their lower rankings. The Complutense University of Madrid, ranked 249th globally, shows retweet distributions comparable to top-tier universities, suggesting that institutional rank does not always directly correlate with social media engagement metrics.

Among the institutions analyzed, the University of Leicester emerged as a leading performer, consistently excelling in engagement categories like media types, hashtags, and URLs, as visible in Figure 6. While Leicester led overall performance, MIT stood

out with high usage of photos, emphasizing visually engaging content. URLs were the most commonly used content type across all institutions, highlighting their role in link-sharing. Manchester diverged from this trend by using notably fewer URLs, media types, and hashtags, suggesting a different engagement strategy. Coimbra was an outlier with its remarkably high usage of hashtags, contrasting with the practices of similarly ranked institutions. In contrast, Trinity and Leicester displayed similar patterns of high engagement across all categories, reinforcing Leicester's position as a leader in digital engagement.
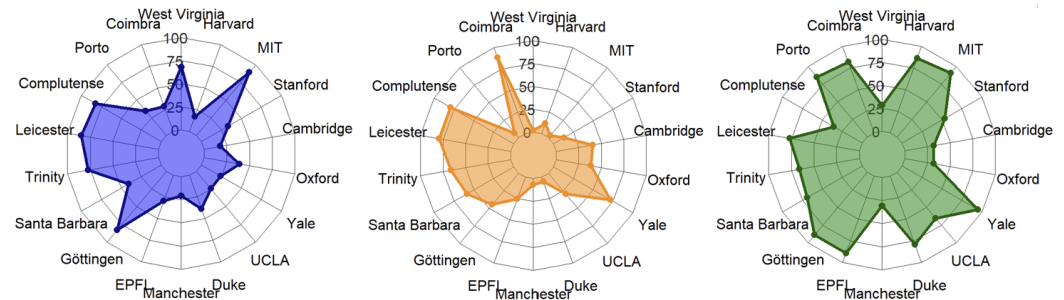

Figure 6. Average distribution of media type, hashtag, and URL usage per HEI.

6. Post Categorization

The exponential growth of social media platforms has created a wide range of content, offering both opportunities and challenges in extracting meaningful insights. Effective categorization of social media posts remains crucial for optimizing information dissemination, enhancing user experience, and delivering more targeted content. To achieve this, a set of categories has been established to analyze the thematic content of general tweets from various HEIs to classify their publishing strategies and identify and predict emerging trends and patterns. This categorization is grounded in the methodologies and findings detailed in research conducted by Oliveria et al. [6,52] and Coelho et al. [53], which have significantly contributed to our understanding.

The categories used for analysis are:

- **Image**—which encompasses the external image of the HEI, including its reputation, branding, and public perception.
- **Education**—relates to the institution's role in providing educational services, including its operations involving students, faculty, and academic programs.
- **Research**—covers activities that top-ranked HEIs are expected to pursue to maintain and foster their international status, such as scholarly research and innovation.
- **Society**—includes posts that communicate with the broader community, or provide information on topics of general interest related to the HEI and its broader context and/or to society.
- **Engagement**—focuses on the effort of the institution to interact with its audience, fostering connections and promoting engagement through comments, shares, favorites, and other forms of active participation on social media platforms.

6.1. Search Topic and Manual Topic Refinement

An automatic approach was implemented through topic classification to categorize the mentioned topics efficiently. The method selected for this task was BERTopic, a topic-mining technique that utilizes transformers and c-TF-IDF to form dense clusters. This approach enables the creation of easily interpretable topics while preserving keywords in the topic descriptions.

Initially, each document is transformed into an embedding representation using a pre-trained language model, converting the posts into numerical vectors. While other models were considered, such as SpaCy [54] and Word2Vec [55], the default choice of sentence-transformers was selected as it outperformed other alternatives. Based on the original models' list provided by the Sentence Transformers Hugging Face organization [56], the pre-trained model chosen was `all-mpnet-base-v2`, selected based on its strong general-purpose performance and its training on over 1 billion pairs, which gave it the highest average performance among the available options. Additionally, the removal of stop words as a preprocessing step is not advised as the transformer-based embedding models need the full context to create accurate embeddings. However, stop words might end up in our topic representation without adding meaningful interpretation, as they are not seed words that describe the five predefined topics. Thus, Countvectorizer is used to preprocess our posts after having generated embeddings and grouped them into clusters.

BERTopic generates topics based on a dataset provided, allowing for the unsupervised discovery of themes within the text. Nonetheless, further steps were taken since our objective was to categorize the publications into five topics (Image, Education, Research, Society, and Engagement). After the initial topic modeling, we calculated the embeddings for these predefined topics and then measured their cosine similarity with the document embeddings generated by BERTopic. By using cosine similarity, we were able to assess how each post aligned with our predefined topics, allowing us to fine-tune the topic assignments. Therefore, we were able to identify the five most similar topics that matched our categories. From these, we extracted the first five words associated with each one, since they represent the core concepts that define each topic, offering insight into the underlying themes. In this process, Topic-1, which contains outliers in BERTopic, was excluded to prevent distortion in the analysis and to maintain the accuracy and relevance of our categorization. To ensure the best result, it was processed ten times for each institution, and the coherence score metric was used to evaluate the quality and the degree of semantic similarity between high-scoring words within a topic.

After obtaining the seed words, we intended to use them to define the desired topics. However, as noted by the author in [57] and reflected in the results, HEIs with fewer publications generated a limited number of distinct topics. For instance, in the case of Complutense, this limitation resulted in the same topic being assigned to all five intended categories. To overcome this issue and achieve more accurate topic differentiation, an additional 15 manually selected words were incorporated into the seed words:

- **Education**: faculty, students, professors, courses, curriculum, teaching, classes, lecture, learning, degrees, enrollment, education, academics, exams, internships;
- **Society**: community, event, announcement, public, outreach, ceremony, celebration, congratulations, initiative, volunteer, charity, society, networking, support, collaboration;
- **Engagement**: welcome, join, participate, share, connect, engage, follow, celebrate, communicate, discuss, contribute, engage, network, respond, invite;
- **Image**: reputation, history, recognition, prestige, excellence, leadership, innovation, influence, legacy, status, accreditation, visibility, ranking, image, distinction;
- **Research**: study, findings, analysis, discovery, experiment, investigation, innovation, publication, data, breakthrough, research, development, researcher, insights, results.

To attribute topics to the documents, a pre-trained model was used to generate embeddings. First, the seed words for each desired topic were embedded using the `all-mpnet-base-v2` model, establishing a reference for comparison. The documents were then embedded using the same model to maintain consistency. Topics were assigned by calculating the average cosine similarity between each document's embedding and the seed

word embeddings. The document was assigned to the topic with the highest similarity score, indicating the closest thematic match.

Analysis of Assigned Topics

Before proceeding with the analysis, it is important to recognize that certain HEIs (Stanford, Duke, and Complutense) produced a limited number of topics using BERTopic. As previously noted, in the case of Complutense, the model generated a small number of distinct topics, leading to the same topic being assigned across all five intended categories. This limitation led to fewer descriptive words for each topic, making the topic assignment largely dependent on the manual words introduced. Likewise, Stanford and Duke show similar behavior, producing two and three topics, respectively. In addition, the coherence score of the topic modeling generally falls within the range of 40 to 60. Except for the aforementioned institutions, it appears that a lower coherence score often corresponds to a greater number of topics generated. This suggests that as the model attempts to identify more topics, the coherence of those topics tends to decrease, implying that generating more topics can sometimes dilute the coherence, making the topics less semantically related and potentially more difficult to interpret. Furthermore, during the topic attribution process, some entries resulted in a negative average cosine similarity. Although these values represent the best match among the five predefined topics, the negative result shows a level of dissimilarity between the document embeddings and the seed topic embeddings. Therefore, these entries were removed from further analysis, retaining 99% of the original data across the 18 HEIs.

In terms of temporal distribution, Education stands out as the dominant topic closely followed by Research, while Image has the lowest frequency of publications, as shown in Figure 7. UCLA presented a sharp peak in March driven by engagement posts through interactive tweets aimed at connecting with followers. For instance, Harvard, Oxford, and Cambridge maintain a steady flow of publications across all topics year-round, showcasing their stable academic activities. In contrast, Santa Barbara and Göttingen show more variability, with noticeable peaks and troughs in their activity. This pattern may suggest a more dynamic or project-focused approach, with certain times dedicated to concentrated efforts on specific initiatives or topics.

To delve deeper into the analysis, the PrefixSpan algorithm [58] was implemented to identify frequent sequential patterns within the selected institutions, yielding over 50% accuracy to ensure that the sequences identified were both frequent and significant. To visualize the occurrence of the frequent sequential pattern results, a timespan of two weeks was selected, as illustrated in Figure 8. Harvard emphasizes both Education and Research in its activities, with recurring patterns like Education→Education on July 27th and 28th and Education→Research, reflecting the strong commitment to these areas. Thus, it expresses how the institution integrates its role in providing educational services with its commitment to maintaining international status through academic research and innovation, where one type of activity may lead to another. Similarly, Yale demonstrates a recurrent Research→Research pattern, further emphasizing its commitment to research excellence, as visible on the 22nd and 23rd of July. In contrast, while the patterns generated consisted of the predominance of Education and Research as being the most frequently occurring patterns across various institutions, MIT presents a noticeable pattern related to the topic Image, indicating discussions or content related to visual data, institutional branding, or public perception.
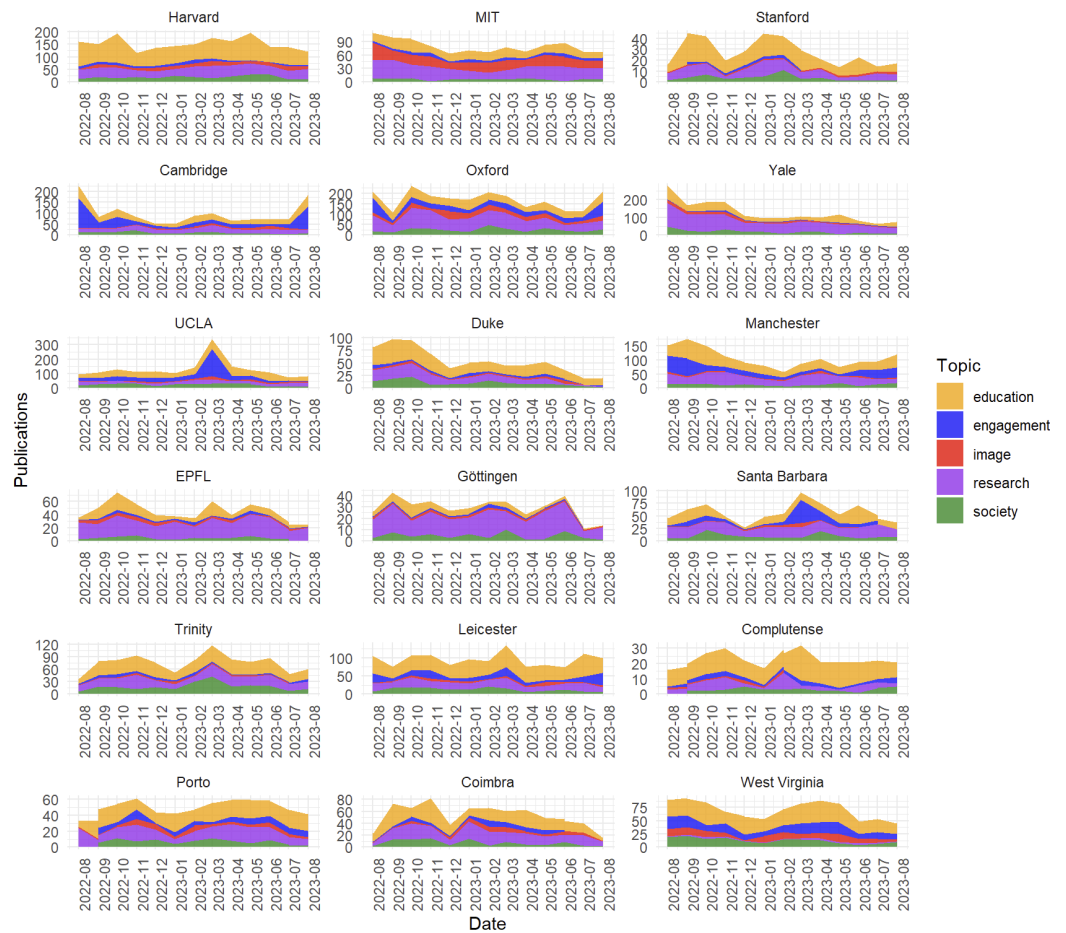
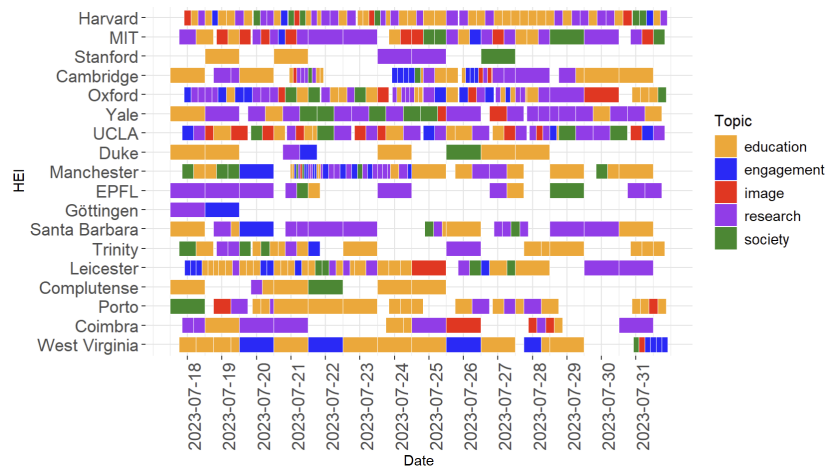**Figure 7.** Temporal distribution of publication frequencies by topic.

Figure 8. Topic distribution per day for each HEI during July 2023.

7. Predictive Modeling

To anticipate shifts in the communication strategies of higher education institutions (HEIs) on social media, we employed predictive modeling techniques on the categorized themes of their posts. Focusing on the last five days of each month for each institution, we aimed to predict future social media output based on historical data patterns. For this purpose, we utilized a Long Short-Term Memory (LSTM) neural network [59] and a Random Forest (RF) algorithm [60] for predictive analysis. These models were chosen for their effectiveness in handling time-series data and classification tasks, respectively.

The text of the posts published underwent a second round of pre-processing, incorporating additional steps to ensure cleaner and more uniform data. Additionally, all text was converted to lowercase, and stop words, digits, punctuation, and any non-standard characters (for example, special symbols or typographic dashes) were removed. The text was then tokenized and lemmatized to standardize the words and prepare them for further analysis.

The topic prediction process began by splitting the data of each HEI into training and testing sets, with the testing set consisting of posts from the last five days of each month. Initially, the training data comprised posts from the remaining days of each month. However, as noted in the topic distribution, the posts were not evenly distributed among the five topics, with Education emerging as the dominant one, leading to an imbalanced training dataset. To address it, the most frequent topic of that month was identified and additional examples were selected from adjacent months based on their temporal proximity to the target period (month and year) to compensate for the deficit to balance the remaining categories. Temporal proximity was calculated by determining the number of days between each example's date and the first or last day of the target month, ensuring the most relevant examples were selected. Those closer to the target month were prioritized to maintain relevance. Once enough examples were chosen to balance each category, the training dataset was completed by merging the original data with the newly added examples.

After the text content is tokenized and converted into a sequence of integers, padding is applied to ensure all sequences have the same length, making them more suitable for model input. Additionally, the model incorporates temporal features, particularly the month and day of the week included for seasonal publication patterns, along with the average cosine similarity score which determined the topic assignment for each month, thus ensuring that topic assignment is handled more accurately. In the training process, the model is trained for 10 epochs with a batch size of 32, allowing it to learn and adjust its parameters with each pass progressively.

The architecture of the LSTM model begins with an embedding layer that transforms tokenized text sequences into dense vector representations, effectively capturing the semantic relationships between words. The core component of the model is the LSTM layer, which processes sequences over time using forget, input, and output gates. This structure allows it to maintain and update internal states, capturing long-term dependencies in the data. Next, an intermediate dense layer with a ReLU activation further refines the extracted features. Finally, the model's output layer, utilizing a *softmax* activation function, generates a probability distribution across the possible topics, allowing for the classification of each post into one of the predefined categories. The model is trained iteratively; it first learns from the initial training data and then updates itself based on predictions made during the test phase by the order in which they were published. After each prediction, the test set is updated with the true result of the post, regardless of the prediction's accuracy, continuing this process until the last post of the final five days of the respective month. This ensures that the model adapts its ability to predict posts published during the last five days of each month.

In parallel, a Random Forest algorithm was implemented to perform the topic prediction task, initialized with one hundred trees, each trained on random subsets of the balanced training data containing the same features as the LSTM as previously mentioned. After each prediction on the test set, the model undergoes complete retraining with the newly incorporated test example, allowing it to incrementally refine its predictive capabilities.

To evaluate the performance either for LSTM or Random Forest, the accuracy metric is used but with some modifications. Since the target consists of assessing the model's ability to predict the last 5 days' posts and capture trends within a more balanced dataset, the accuracy is calculated by

$$\text{Accuracy}_t = \frac{1}{t} \sum_{i=1}^{t} \mathbb{I}(\hat{y}_i = y_i) \tag{8}$$

where y_i represent the true label for the i-th post, and $\hat{y}_i$ represent the predicted label for the i-th post. Therefore, it helps to measure the effectiveness of the machine learning model in classifying data correctly.

And the overall accuracy of the T posts (for the last 5 days of the month) is then:

$$\text{Final Accuracy} = \frac{1}{T} \sum_{t=1}^{T} \text{Accuracy}_t. \tag{9}$$

End-of-Month Prediction Insights

To evaluate the performance of the machine learning models throughout the academic year, we present the results for LSTM and Random Forest in Figures 9 and 10. Figure 9 illustrates the monthly accuracy for each of the 18 higher education institutions (HEIs), with point sizes indicating the number of publications shared during the last five days of each month.

For the LSTM model, Duke consistently recorded accuracies below 50%, whereas with the Random Forest model, it occasionally surpassed the 50%. Stanford showcased the highest variability with the LSTM model, with accuracy ranging from 0% in August 2023 to 100% in July 2023, indicating a sensitivity to changes in publication volume or topic shifts.

When considering the volume of data published in the last five days of each month, Yale stands out, particularly during the initial four months (August to November 2023). However, despite this high volume, Yale's accuracy fluctuates considerably around the baseline, with noticeable peaks and troughs. This inconsistency may be linked to data variations, external academic events, or a gradual reduction in publication numbers over time, which might impact the model's performance. Similar patterns of fluctuating accuracy with a general trend toward improvement were observed for Stanford, EPFL, and the University of Coimbra, possibly indicating adaptations or refinements in their social media strategies.

In contrast, MIT, Santa Barbara, Trinity, and Leicester demonstrate relative stability in their accuracy, with only minor fluctuations around the baseline. This consistency suggests robust model performance less influenced by monthly publication variations. Göttingen, although exhibiting volatility similar to Manchester, manages to achieve the best overall results with an accuracy of 1 in two separate months and only one instance (April 2023) where accuracy decreased below 50%, indicating strong and reliable performance despite some fluctuations.

Both MIT and the University of Oxford showed consistent publication volumes and stable accuracies, implying that their social media posting strategies are likely more uniform and effectively managed. This uniformity minimizes the impact of end-of-month publication variations in the models' accuracy.

When comparing models, the Random Forest algorithm exhibited more significant fluctuations across all HEIs than the LSTM model, which generally showed smoother performance curves. Santa Barbara had the lowest performance with the Random Forest model, while Duke University had the lowest accuracy with the LSTM. For Duke, the Random Forest model showed variability with two peaks surpassing the 50% threshold, including in the last month where there were three data points but only two peaks.

The University of Coimbra's performance with the Random Forest model was relatively better, featuring multiple instances of positive accuracy percentages. Notably, Coimbra displayed a unique pattern where accuracy decreased almost uniformly by 10%

each month between August 2022 and March 2023, except for February, where the accuracy percentage was higher than 10%.

Comparing the LSTM and Random Forest models, our analysis reveals distinct performance dynamics in predicting topic values over time for the HEIs studied. Specifically, the LSTM model achieved a mean accuracy of 51% with a standard deviation of 18%, whereas the Random Forest model attained a mean accuracy of 43% with a standard deviation of 19%. The better performance of LSTM can be attributed to its ability to capture temporal dependencies in sequential data, making it more suitable for time-series predictions in our dataset. Although the LSTM model requires more computational time to execute due to its complex architecture and training process, the higher accuracy and consistency it provides justify its use in contexts where prediction precision is critical. These results demonstrate that the LSTM model is the most appropriate choice for our predictive task.

Given the LSTM model's higher accuracy and better handling of temporal data compared to the Random Forest model, we decided to pursue the study using only the LSTM model. Consequently, Figure 10 displays the distribution of accuracy across the five desired topics, showcasing the data for each month obtained using the LSTM model. The accuracy for each topic was calculated using the following expression:

$$\text{Accuracy Topic}(k) = \frac{C_k}{N_k} \tag{10}$$

where the C_k denotes the number of instances where the topic K was correctly predicted, and N_k is the total count of instances of topic K in the test set for that month.

Across the majority of HEIs, the topic of Engagement consistently achieves higher accuracy, with greater consistency and higher median results. In contrast, the topic of Society displays lower accuracy and greater variability, with some institutions like Complutense even recording an accuracy of zero, which includes an outlier. The topic of Education, while generally having lower accuracy across institutions, is notable for its smaller IQR.

Complutense exhibits the worst performance among the institutions showing a median accuracy of zero in three out of five topics during the selected period. This indicates consistent difficulties in accurately predicting topics like Research despite the outliers indicated by the dots on the plots. This performance may be associated with the high number of posts categorized as Education, which has been noted as the dominant topic for this HEI, potentially leading to insufficiently diverse data for effectively training the model on other topics. In contrast, Yale exhibits the best performance, with three out of five topics achieving median accuracies above 50%. Yale's box plots are well distributed and balanced, indicating robust model performance across various topics.

Furthermore, the topic of Image reveals substantial variability across institutions. Oxford and Cambridge, for instance, have wider IQRs and lower median accuracies, suggesting difficulties in modeling this topic accurately. Moreover, MIT, Oxford, and Yale display a broader distribution of accuracy across all topics, pointing to more varied performance. On the other hand, Coimbra and Leicester display unique accuracy distributions that diverge from general trends, which may suggest differences in data handling.

In conclusion, while some institutions like MIT and Cambridge demonstrate more consistent performance, others, such as Göttingen and Stanford, show significant variability across different topics. This highlights the importance of ongoing model refinement and adaptation to handle the complexity and diversity of content more effectively. In this study, all HEIs were evaluated uniformly without adjustments for their unique characteristics to maintain a standardized and scalable approach, facilitating broad applicability of the model. However, to enhance model accuracy, improvements should be pursued independently,

using tailored strategies that incorporate domain-specific insights to address the specific needs and features of each institution.
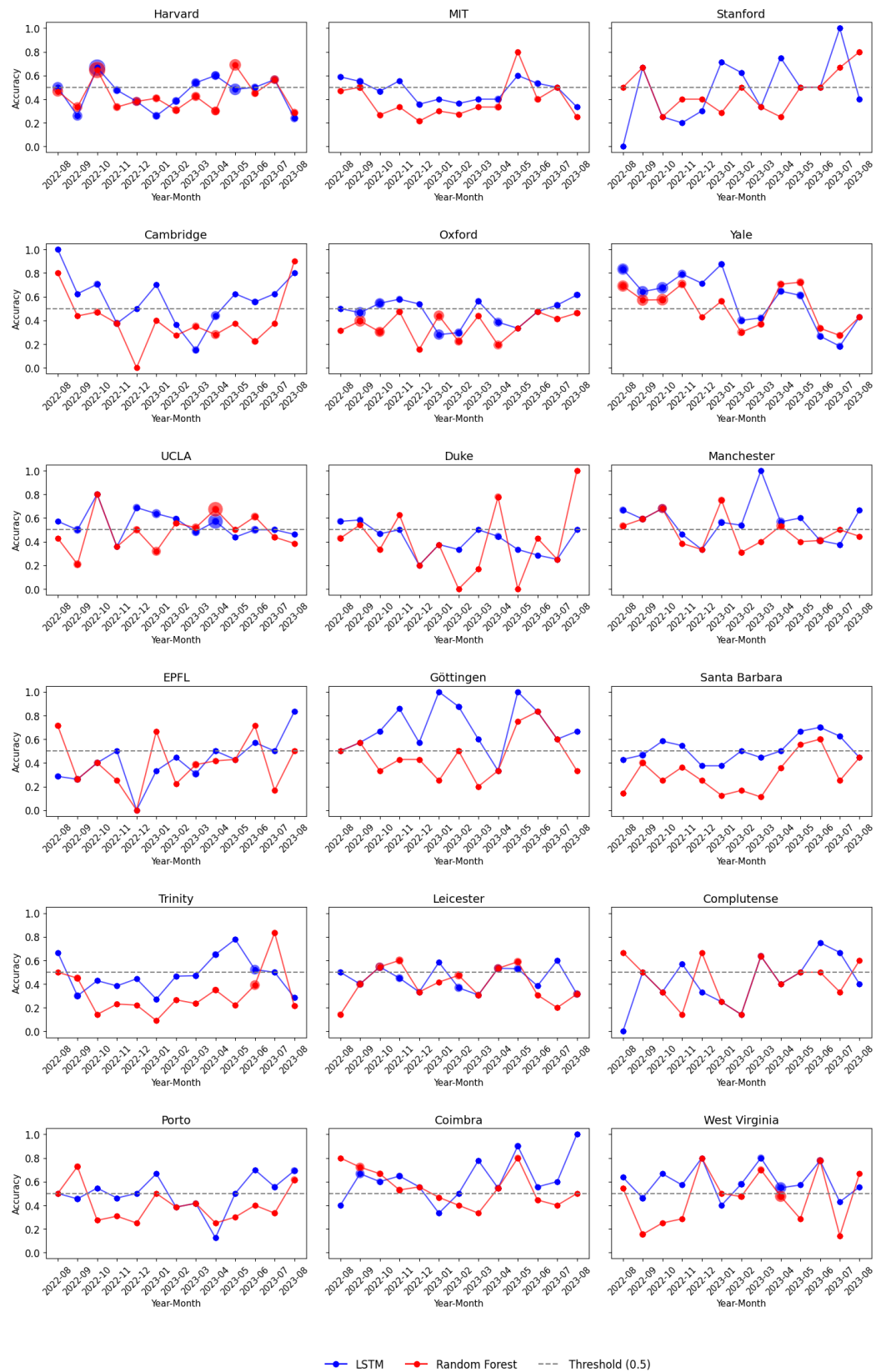

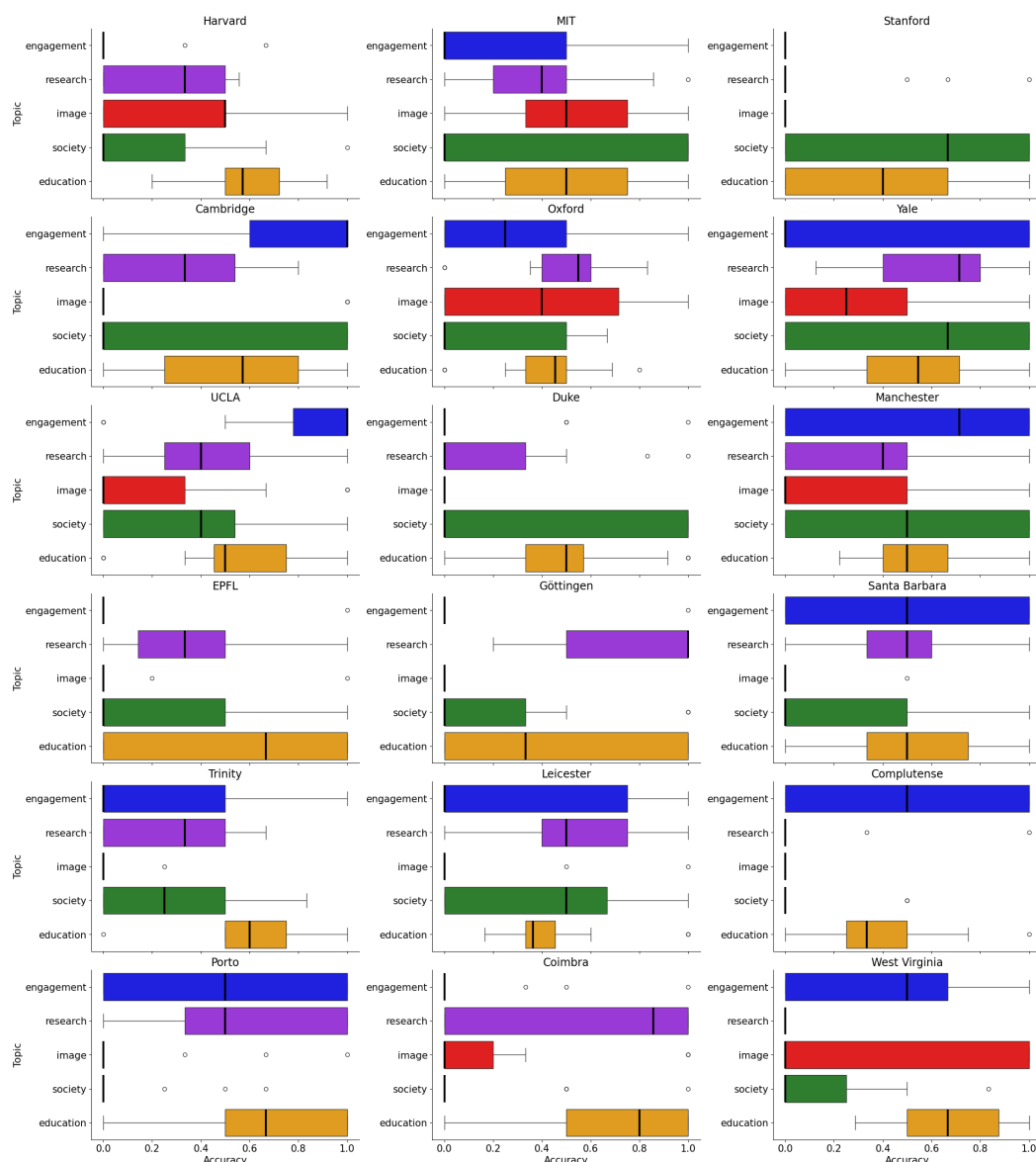**Figure 9.** Prediction accuracy for posts: last 5 days of each month.

Figure 10. Topic-wise accuracy distribution across each HEI.

8. Conclusions

This study examined the relationship between social media strategies employed by higher education institutions (HEIs) and their positions in global rankings. Our findings indicate that low-ranking institutions often adopt innovative and experimental approaches in their social media presence, focusing on engaging local communities and promoting educational content. These universities prioritize posts related to internal events and cultural activities, demonstrating a strong commitment to education-centric communication. The variability and adaptability of their strategies frequently result in high engagement levels, highlighting the effectiveness of customized, community-focused content in enhancing their online impact.

Moderately ranked institutions tend to incorporate elements from both high- and low-ranking strategies. They maintain a consistent posting schedule around midday and make pronounced use of URLs, typically directing followers to their institutional platforms. This approach allows them to disseminate larger amounts of information within the character limits of social media posts. Their focus tends to be on research, especially for those institutions closer to the top of this middle tier, who often shift their attention away from

educational content, as exemplified by MIT and Duke. Despite varied emphases across this group, the predictive accuracy of their social media impact generally remains under 50%.

High-ranking institutions exhibit more sophisticated and consistent social media strategies. They leverage online platforms to reinforce their globally recognized brands and public images, engage with widespread audiences, and highlight their world-leading academic achievements and research breakthroughs. Evidenced by their significantly high volumes of retweets and favorites, these universities effectively boost their reputations and continue to attract top students and researchers, which is vital to maintaining their high positions in the rankings.

Overall, our analysis suggests a pattern where social media strategies are aligned with institutional rankings, highlighting tailored approaches that resonate with their respective audiences. However, it is important to note that these conclusions are based on an examination of 18 selected HEIs and may not fully represent the broader landscape of institutional social media accounts. Future research could expand upon this work to include a more comprehensive range of institutions, thereby validating and extending these findings.

Our approach not only provides deeper insights into how social media is leveraged to enhance institutional visibility and engagement but also offers a means for understanding the broader implications of these strategies on global rankings. The methodologies employed in predictive modeling, topic mining, and clustering methodologies indicates a promising direction for future research, enabling a more comprehensive analysis of the relationship between social media strategies and the competitive positioning of HEIs.

Limitations and Future Work

A limitation of this study lies in the limited engagement metrics available in our dataset, which constrained our ability to thoroughly analyze audience responses to publications. Future studies could address this by assigning varying levels of importance to different engagement types, enabling a more detailed evaluation of their impact on overall outcomes. Moreover, incorporating image and video content from posts into the analysis could provide a more comprehensive perspective on the data. Finally, exploring the expectations and preferences of higher education stakeholders, along with extending the analysis to cover multiple academic years, would enhance the depth and robustness of such research.

Author Contributions: Conceptualization, A.F.; methodology, A.F. and B.R.; software, B.R.; validation, B.R. and A.F.; formal analysis, B.R.; investigation, B.R.; resources, A.F.; data curation, B.R.; writing—original draft preparation, B.R.; writing—review and editing, A.F.; visualization, B.R. and A.F.; supervision, A.F.; project administration, A.F.; funding acquisition, A.F. All authors have read and agreed to the published version of the manuscript.

References

1. Rafique, T.; Awan, M.U.; Shafiq, M.; Mahmood, K. Exploring the role of ranking systems towards university performance improvement: A focus group-based study. *Heliyon* **2023**, *9*, e20904. [CrossRef] [PubMed]
2. Pavel, A.P. Global university rankings—A comparative analysis. *Procedia Econ. Financ.* **2015**, *26*, 54–63. [CrossRef]
3. Hazelkorn, E. How rankings are reshaping higher education. In *Los Rankings Univeritarios: Mitos y Realidades*; Tecnos: Madrid, Spain, 2013.

4. Peters, M.A. Global university rankings: Metrics, performance, governance. *Educ. Philos. Theory* **2019**, *51*, 5–13. [CrossRef]
5. Constantinides, E.; Zinck Stagno, M.C. Potential of the social media as instruments of higher education marketing: A segmentation study. *J. Mark. High. Educ.* **2011**, *21*, 7–24. [CrossRef]
6. Oliveira, L.; Figueira, Á. Social media content analysis in the higher education sector: From content to strategy. *Int. J. Web Portals (IJWP)* **2015**, *7*, 16–32. [CrossRef]
7. Figueira, Á. A Three-Step Data-Mining Analysis of Top-Ranked Higher Education Institutions' Communication on Facebook. In Proceedings of the Sixth International Conference on Technological Ecosystems for Enhancing Multiculturality, Salamanca, Spain, 24–26 October 2018; pp. 923–929.
8. Evans, D. *Social Media Marketing: The Next Generation of Business Engagement*; John Wiley & Sons: Hoboken, NJ, USA, 2010.
9. Shehatta, I.; Mahmood, K. Correlation among top 100 universities in the major six global rankings: Policy implications. *Scientometrics* **2016**, *109*, 1231–1254. [CrossRef]
10. Çakır, M.P.; Acartürk, C.; Alaşehir, O.; Çilingir, C. A comparative analysis of global and national university ranking systems. *Scientometrics* **2015**, *103*, 813–848. [CrossRef]
11. Aguillo, I.; Bar-Ilan, J.; Levene, M.; Ortega, J. Comparing university rankings. *Scientometrics* **2010**, *85*, 243–256. [CrossRef]
12. Figueira, A.; Nascimento, L.V. Do Top Higher Education Institutions' Social Media Communication Differ Depending on Their Rank? *Technology* **2022**, *2*, 2.
13. Mendonça, M.; Figueira, Á. Topic Extraction: BERTopic's Insight into the 117th Congress's Twitterverse. *Informatics* **2024**, *11*, 8. [CrossRef]
14. de Groot, M.; Aliannejadi, M.; Haas, M.R. Experiments on generalizability of BERTopic on multi-domain short text. *arXiv* **2022**, arXiv:2212.08459.
15. Egger, R.; Yu, J. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Front. Sociol.* **2022**, *7*, 886498. [CrossRef] [PubMed]
16. Fütterer, T.; Fischer, C.; Alekseeva, A.; Chen, X.; Tate, T.; Warschauer, M.; Gerjets, P. ChatGPT in education: Global reactions to AI innovations. *Sci. Rep.* **2023**, *13*, 15310. [CrossRef] [PubMed]
17. Grigore, D.N.; Pintilie, I. Transformer-based Topic Modeling to Measure the Severity of Eating Disorder Symptoms. In Proceedings of the CLEF (Working Notes), Thessaloniki, Greece, 18–21 September 2023; pp. 684–692.
18. Schneider, N.; Shouei, S.; Ghantous, S.; Feldman, E. Hate Speech Targets Detection in Parler using BERT. *arXiv* **2023**, arXiv:2304.01179.
19. Antypas, D.; Ushio, A.; Camacho-Collados, J.; Neves, L.; Silva, V.; Barbieri, F. Twitter topic classification. *arXiv* **2022**, arXiv:2209.09824.
20. Lasri, I.; Riadsolh, A.; Elbelkacemi, M. Self-Attention-Based Bi-LSTM Model for Sentiment Analysis on Tweets about Distance Learning in Higher Education. *Int. J. Emerg. Technol. Learn.* **2023**, *18*, 119–141. [CrossRef]
21. Pandey, R.; Singh, J.P. BERT-LSTM model for sarcasm detection in code-mixed social media post. *J. Intell. Inf. Syst.* **2023**, *60*, 235–254. [CrossRef]
22. Nti, I.K.; Akyeramfo-Sam, S.; Bediako-Kyeremeh, B.; Agyemang, S. Prediction of social media effects on students' academic performance using Machine Learning Algorithms (MLAs). *J. Comput. Educ.* **2022**, *9*, 195–223. [CrossRef]
23. Hooda, M.; Rana, C.; Dahiya, O.; Rizwan, A.; Hossain, M.S. Artificial intelligence for assessment and feedback to enhance student success in higher education. *Math. Probl. Eng.* **2022**, *2022*, 5215722. [CrossRef]
24. Zhang, Y.; Xiao, Y.; Wu, J.; Lu, X. Comprehensive world university ranking based on ranking aggregation. *Comput. Stat.* **2021**, *36*, 1139–1152. [CrossRef]
25. World University Rankings. 2023. Available online: https://www.timeshighereducation.com/world-university-rankings/2024/world-ranking (accessed on 19 November 2023).
26. QS World University Rankings 2024. 2024. Available online: https://www.topuniversities.com/world-university-rankings (accessed on 19 November 2023).
27. World University Rankings 2023 | Global 2000 List | CWUR. 2023. Available online: https://cwur.org/2023.php (accessed on 19 November 2023).
28. ShanghaiRanking's Academic Ranking of World Universities. 2023. Available online: https://www.shanghairanking.com/rankings/arwu/2023 (accessed on 19 November 2023).
29. World | Ranking Web of Universities | Webometrics Ranks. 2023. Available online: https://www.webometrics.info/en/world (accessed on 19 November 2023).
30. Bar-Ilan, J.; Levene, M.; Lin, A. Some measures for comparing citation databases. *J. Inf.* **2007**, *1*, 26–34. [CrossRef]
31. Diaconis, P.; Graham, R.L. Spearman's footrule as a measure of disarray. *J. R. Stat. Soc. Ser. Stat. Methodol.* **1977**, *39*, 262–268. [CrossRef]
32. Dwork, C.; Kumar, R.; Naor, M.; Sivakumar, D. Rank aggregation methods for the web. In Proceedings of the 10th International Conference on World Wide Web, Hong Kong, 1–5 May 2001; pp. 613–622.

33. École Polytechnique Fédérale de Lausanne. EPFL Summer Research Program, 2023. Available online: https://www.epfl.ch/schools/sv/education/summer-research-program/ (accessed on 7 September 2024).

34. Harvard University. Academic Text Calendar 2022–23. 2022. Available online: https://www.gsd.harvard.edu/wp-content/uploads/2022/07/Academic-Text-Calendar-2022-23-072522.pdf (accessed on 30 January 2024).

35. Massachusetts Institute of Technology. Key Catalog and Schedule Dates 2022–2023. 2022. Available online: https://registrar.mit.edu/sites/default/files/2022-09/Key_catalog_and_schedule_dates_2022-2023.pdf (accessed on 30 January 2024).

36. Stanford University. Stanford Academic Calendar 2022–23. 2022. Available online: https://studentservices.stanford.edu/stanford-academic-calendar-2022-23#autumn-quarter-2022-23 (accessed on 30 January 2024).

37. Cambridge University. Term Dates and Calendars. 2022. Available online: https://www.cambridge-news.co.uk/news/cambridge-news/cambridge-university-202223-term-dates-24773402 (accessed on 30 January 2024).

38. Daily Information. Oxford University Term Dates. 2022. Available online: https://www.dailyinfo.co.uk/oxford/guide/university-term-dates (accessed on 30 January 2024).

39. Yale University. Yale College Calendar 2022–2023. 2022. Available online: https://registrar.yale.edu/sites/default/files/files/2022-2023_Yale%20College%20Calendar%20with%20Pertinent%20Deadlines%20_%20Yale%20University.pdf (accessed on 30 January 2024).

40. University of California, Los Angeles. UCLA Academic Calendar 2022–23. 2022. Available online: https://registrar.ucla.edu/archives/academic-calendar-archive/academic-calendar-2022-23 (accessed on 30 January 2024).

41. Duke University. Duke Academic Calendar 2022–2023. 2022. Available online: https://registrar.duke.edu/2022-2023-academic-calendar/ (accessed on 30 January 2024).

42. The University of Manchester. Academic Calendar 2022–2023. 2022. Available online: https://studentnet.cs.manchester.ac.uk/ugt/2023/timetable/cts_calendar.pdf (accessed on 30 January 2024).

43. EPFL. EPFL Information About Dates and Deadlines 2022–2023. 2022. Available online: https://www.epfl.ch/campus/services/internal-trainings/language-centre/practical-information/information-about-dates-and-deadlines-registration-and-courses/ (accessed on 30 January 2024).

44. Georg-August-Universität Göttingen. Academic Calendar 2022–2023. 2022. Available online: https://www.uni-goettingen.de/en/academic%2Bcalendar%2B%28including%2Bfuture%2Band%2Bpast%2Bsemesters%29/24440.html (accessed on 30 January 2024).

45. University of California, Santa Barbara. UCSB Academic Calendar 2022–2023. 2022. Available online: https://registrar.sa.ucsb.edu/calendars/calendars-deadlines/academic-calendars/academic-calendar-for-2022-2023 (accessed on 30 January 2024).

46. The University of Dublin. Academic Year Structure 2022–23. 2022. Available online: https://www.tcd.ie/calendar/academic-year-structure/2022-23/academic-year-structure.pdf (accessed on 30 January 2024).

47. The University of Leicester. Term and Semester Dates 2022–2023. 2022. Available online: https://le.ac.uk/about/info/term-semester-dates/closure-days (accessed on 30 January 2024).

48. Universidad Complutense de Madrid. Complutense University of Madrid Academic Calendar 2022–2023. 2022. Available online: https://trabajosocial.ucm.es/file/calendario-ingles-2022-23?ver (accessed on 30 January 2024).

49. University of Porto. Academic Calendar 2022–2023. 2022. Available online: https://www.up.pt/portal/en/study/academic-information/academic-calendar/ (accessed on 30 January 2024).

50. University of Coimbra. University of Coimbra Academic Calendar 2022–2023. 2022. Available online: https://www.uc.pt/en/academicos/regulamentos/calendario (accessed on 30 January 2024).

51. West Virginia University. WVU Academic Calendar 2022–2023. 2022. Available online: https://registrar.wvu.edu/calendars (accessed on 30 January 2024).

52. Oliveira, L.; Figueira, Á. Benchmarking analysis of social media strategies in the Higher Education Sector. *Procedia Comput. Sci.* **2015**, *64*, 779–786. [CrossRef]

53. Coelho, T.; Figueira, Á. Analysis of Top-Ranked HEI Publications' Strategy on Twitter. In Proceedings of the 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 15–18 December 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 5875–5877.

54. Honnibal, M.; Montani, I. spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. 2017. Available online: https://spacy.io (accessed on 26 December 2024).

55. Mikolov, T. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.

56. Reimers, N.; Gurevych, I. Sentence-Transformers: Pretrained Models. 2023. Available online: https://www.sbert.net/ (accessed on 2 June 2024).

57. Grootendorst, M. BERTopic FAQ. 2023. Available online: https://maartengr.github.io/BERTopic/faq.html (accessed on 2 June 2024).

58. Han, J.; Pei, J.; Mortazavi-Asl, B.; Pinto, H.; Chen, Q.; Dayal, U.; Hsu, M. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In Proceedings of the 17th International Conference on Data Engineering, Heidelberg, Germany, 2–6 April 2001; IEEE: Piscataway, NJ, USA, 2001; pp. 215–224.

59. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]

60. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]