

A Machine Learning Approach to Identify HEIs' Publication Strategies

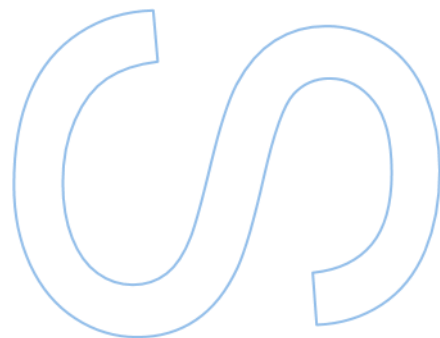
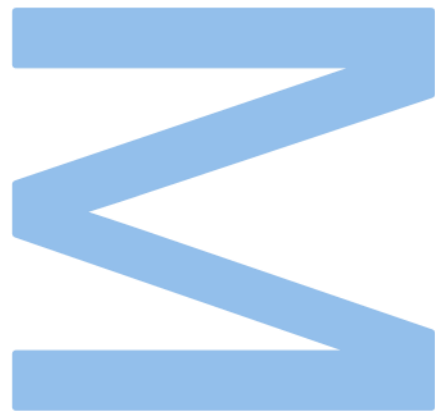
Bruna Catarina Mesquita Rocha

Master in Computer Science

Department of Computer Science

Faculty of Sciences of the University of Porto

2024



A Machine Learning Approach to Identify HEIs' Publication Strategies

Bruna Catarina Mesquita Rocha

Dissertation carried out as part of the Master in Computer Science

Department of Computer Science
2024

Supervisor

Prof. Dr. Álvaro Figueira,
Assistant Professor,
Faculty of Sciences of the University of Porto

“A vida é mais larga do que comprida.”

Avó Maria

Acknowledgements

Quero agradecer ao meu orientador Prof. Álvaro Figueira por todo o acompanhamento, dedicação e incentivo constante que me deu ao longo deste percurso. Um obrigada não chega por toda a aprendizagem de sobrevivência na vida de investigação.

Aos meus pais, por me terem dado a chance de estudar aquilo que me faz feliz, por me acompanharem incondicionalmente e por me darem forças ao longo de todos os bons e momentos menos bons. Ao meu irmão, por me ter massacrado a cabeça, e a quem espero retribuir o favor agora que começa ele o mesmo percurso.

A todos os meus amigos, fora e dentro de água, pelos convívios e motivação que me deram às suas próprias maneiras.

E, por fim, ao Miguel, obrigada por toda a paciência, amor e carinho. Por estares presente desde o início deste percurso académico e para além dele.

Resumo

No universo competitivo do ensino superior, as instituições dependem cada vez mais de rankings internacionais para assegurarem financiamento, atraírem talento e melhorarem a sua reputação global. Simultaneamente, estas universidades têm expandido a sua presença nas redes sociais, utilizando estratégias de publicação sofisticadas para disseminar informação e para aumentar o seu reconhecimento e envolvimento. Este estudo examina a relação entre os rankings de Instituições de Ensino Superior (HEI) e as suas estratégias de publicação nas redes sociais. Colecionámos e analisámos tweets de 18 HEIs presentes num sistema consolidado de ranking, focando-nos em várias características das suas publicações. Esta organização permitiu-nos ganhar um melhor esclarecimento das estratégias de publicação, levando à identificação de quatro agrupamentos distintos de HEIs, que melhoraram a interpretação das estratégias de publicação. Cada grupo foi indicado a uma de três camadas de ranking: posição alta, moderada ou baixa. Para ganhar uma visão mais completa destas estratégias, categorizámos as publicações de cada instituição em cinco tópicos predefinidos - envolvimento, investigação, imagem, comunidade e educação. Esta catalogação, juntamente com o modelo LSTM, foi usada na previsão da produção nas redes sociais nos últimos cinco dias de cada mês, que concluímos ter sucesso, indicando a presença de um claro padrão. Esta dissertação explora como é que variações nestas estratégias de redes sociais impactam os rankings das HEIs. Os resultados sugerem uma interação subtil entre envolvimento nas redes sociais e o prestígio percecionado das HEIs.

Palavras-chave: Análise de Redes Sociais, Instituições de Ensino Superior, Mineração de texto, Análise de Clusters, Análise de Ranking, Aprendizagem Automática

Abstract

In the competitive landscape of higher education, institutions increasingly rely on international rankings to secure funding, attract talent, and enhance their global reputation. Concurrently, 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 collected and analyzed tweets from 18 HEIs featured in a consolidated ranking system, focusing on various features of their social media posts. This grouping allowed for a deeper understanding of the posting strategies, leading to the identification of four distinct clusters of institutions, which allowed for an easier interpretation of the publishing strategies. Each cluster was assigned to one of three ranking layers according to the positioning of its HEI constituents: high, moderate, or low-ranking. To gain a more thorough insight into these strategies, we categorized the posts into five predefined topics — engagement, research, image, society, and education. This categorization, together with the LSTM model, was used in the prediction of the social media output in the last five days of each month, which we found to be successful, indicating the presence of a clear pattern. The dissertation then explores how variations in these social media strategies relate to the rankings of HEIs. Each institution's posting patterns were then compared to the previously mentioned layers to determine their level of correlation. Our findings suggest a subtle interaction between social media engagement and the perceived prestige of HEIs.

Keywords: Social Media Analytics, Higher Education Institutions, Text Mining, Cluster Analysis, Ranking System Analysis, Machine Learning

Table of Contents

List of Tables	viii
List of Figures.....	ix
List of Acronyms.....	x
1. Introduction	1
1.1. Objectives	2
1.2. Dissertation Structure	3
2. Literature Review	5
2.1. World University Rankings.....	6
2.2. Cluster Analysis	7
2.3. Natural Language Processing (NLP).....	7
2.3.1. Topic Modeling	8
2.3.2. Topic Classification	9
2.4. Long Short-Term Memory (LSTM).....	11
3. Towards a Universal Ranking Standard.....	12
3.1. World University Rankings.....	12
3.2. Similarity Metrics.....	13
3.2.1. Results	15
3.3. The sample for HEIs	18
4. Data Collection and Preliminary Insights	20
4.1. Exploratory Data Analysis	21
4.1.1. Data Distribution.....	23
4.1.2. Publication Period by Hour, Day and Month.....	24
4.1.3. Retweet and Favorite.....	25
4.1.4. Media Type, Hashtags, and URL.....	26
4.1.5. Emojis.....	27
4.1.6. Counting Characters	28
4.1.7. Sentiment and Emotion Analysis	29
5. HEIs Grouping and Analysis	32
5.1. Analysis of Bi-plot	33
5.2. Box-plot Analysis	35
5.3. Ranking and Posting Strategies	36
6. Post Categorization	38
6.1. Search Topic and Manual Topic Refinement.....	38
6.1.1. Analysis of Assigned Topics	41

6.2. Predictive Modeling	43
6.2.1. End-of-Month Prediction Insights.....	45
7. Interpretation and Evaluation of Results	50
7.1. Posting Strategies.....	50
7.2. Topic Distribution	51
7.3. Publication Prediction Analysis.....	52
8. Conclusion.....	55
8.1. Contributions.....	57
8.2. Limitations and Future Work.....	58
References.....	58
Appendix	68

List of Tables

3.1. Similarity measures for 2023 world university rankings.....	16
3.2. Position of the chosen HEIs in CWUR as published in 2023	19
4.1. Average, highest frequency and standard deviation of daily publication frequency of HEIs	22
5.1. Formed clusters for the chosen HEIs in CWUR.	33

List of Figures

3.1. Resulting values between the similarity measures for the top 200 positions in the university rankings.....	17
4.1. Academic calendars of selected higher education institutions	20
4.2. Data distribution between August 2022 and August 2023.....	22
4.3. Monthly tweet distribution over the selected period.....	23
4.4. Hour versus day of the week of HEIs.....	25
4.5. Retweet and favorite distributions by HEI	26
4.6. Average distribution of Media Type, Hashtag, and URL usage per HEI	27
4.7. Top 5 emojis through the period selected by month	28
4.8. Number of characters during the period selected by month.....	29
4.9. Positive, neutral, and negative emotion distribution of HEIs selected.....	30
4.10Sentiment distribution across HEIs of the period selected	31
5.1. Biplot PCA for HEIs selected	34
5.2. Distribution of all the features grouped by cluster.....	36
6.1. Temporal distribution of publication frequencies by topic	42
6.2. Topic distribution per day for each HEI during July 2023	43
6.3. Prediction example for the month of April (30 days).....	45
6.4. Prediction accuracy for posts: last 5 days of each month	47
6.5. Topic-wise accuracy distribution across each HEI.....	49
1. Month versus day of the week of HEIs	68

List of Acronyms

AHP Analytic Hierarchy Process.

ARWU Academic Ranking of World Universities.

CWUR Center World University Rankings.

HEIs Higher Education Institutions.

IQR Interquartile Range.

KPI Key Performance Indicators.

LDA Latent Dirichlet Allocation.

LSTM Long short-term memory.

NLP Natural Language Processing.

QS QS World University Rankings.

SSE Sum of Square Error.

THE Times Higher Education.

UMAP Uniform Manifold Approximation and Projection.

WEB Webometrics Ranking of World Universities.

1. Introduction

As the global academic environment evolves, understanding the relative standing of Higher Education Institutions (HEIs) has become increasingly important, leading to the development of organized external assessments. In this context, global university rankings provide crucial insights into the performance, reputation, and overall standing of HEIs [1–3]. These rankings have become increasingly sophisticated, employing a variety of metrics and methodologies to assess institutions with greater accuracy [4]. Common evaluation criteria in these rankings include academic achievements, research output, international outlook, and industry income. Some prominent ranking systems such as Times Higher Education (THE), QS World University Rankings (QS), and Academic Ranking of World Universities (ARWU) exemplify the variety and influence of these assessments. The diversity in ranking methodologies highlights their importance as decision-making tools, also aiding stakeholders in making informed choices about collaborations and strategic partnerships, thereby fostering innovation and excellence in academia.

In parallel to traditional rankings, social media has emerged as an arena for universities to engage with prospective students, alumni, and potential investors for research projects [5]. As audiences have gravitated towards social media platforms, HEIs have adapted by developing new communication strategies tailored to these platforms [6]. These strategies are designed to achieve specific institutional goals, such as enhancing visibility, engaging with a broader community, and ultimately influencing enrollment decisions and funding opportunities. Effective social media presence can significantly impact an institution's ability to attract new students and secure financial support, as it enhances the institution's accessibility and relatability, which are crucial for appealing to new students and stakeholders. Therefore, examining the intersection between successful social media strategies and traditional metrics of university success is essential.

Despite having common objectives, the strategies employed by the HEIs can vary significantly, leading to different results. For these strategies to be effective, social media efforts must be aligned with and framed within the overall organizational strategic management goals of the institution [7]. Effectively categorizing social media topics not only ensures that content aligns with the institution's priorities but also balances these with the expectations and interests of key stakeholders, including prospective students, alumni,

and research partners. When social media strategies are aligned with broader organizational objectives, HEIs can achieve more targeted and impactful communication efforts. This strategic alignment reinforces the effectiveness of social media efforts and enables a deeper analysis of how different HEIs perform concerning traditional success metrics like rankings. By examining these categorizations and their impact, not only the institutions but also those responsible for managing other educational institutions, can compare their social media strategies and outcomes, offering valuable insights into the role of content strategy in achieving success.

1.1. Objectives

Our research enhances our understanding of the relationship between modern communication on social media and the dynamics of engagement, recruitment, and funding. It also underscores the extent to which HEIs are aligning their digital outreach strategies with their academic and operational objectives.

To further explore these insights, this study addresses the following key research questions:

RQ1. Which ranking provides the most consensus in institutional positioning compared to the other major ranking systems?

To delve further into the analysis, it is important to recognize that HEIs often occupy different positions across various rankings, largely due to the differing methodologies each ranking employs. Therefore, it becomes essential to identify a level of agreement or similarity among these rankings, allowing for a more accurate evaluation of the institutions' positions and their relative significance.

RQ2. How do different online strategies and behaviors of HEIs correlate with their positioning in academic rankings?

It is important to analyze the social media strategies that different institutions employ to achieve their goals over a given period. By doing so, we can identify patterns and trends that help institutions enhance their visibility, engage with their community, and influence important outcomes. Furthermore, understanding these common strategies and their effectiveness is essential for evaluating how social media efforts align with traditional academic success metrics.

RQ3. What patterns can be identified in the content of the publications across different higher education institutions?

Based on previous studies [8, 9], HEIs consistently focus on systematic topics in their social media content. By leveraging this knowledge and analyzing the content published within a period, it becomes possible to define keywords corresponding to these topics. Therefore, allowing the effective categorization of social media publications into distinct categories provides deeper insights into the communication strategies employed by these institutions.

RQ4. How accurately can their activity be predicted using a sequential training approach?

The research also aims to explore whether it is possible to predict social media posts for specific days based on prior data. By categorizing posts, the research seeks to identify patterns in the content and trends over time that could enable accurate predictions.

RQ5. Is there a significant correlation between social media publishing strategies and university rankings?

This is the main research question. As indicated by the title and the previous questions, it represents the culmination of our analysis and aligns with the study's main objective.

1.2. Dissertation Structure

The present dissertation is organized into 8 chapters, including this one:

- Chapter 2, **Literature Review**: which provides a theoretical background and the introduction to previous research relevant to the present investigation.
- Chapter 3, **Towards a Universal Ranking Standard**: conducts a world university ranking analysis to identify the ranking that most closely aligns with our chosen similarity metrics, based on the positions of higher education institutions, and selects a targeted group of HEIs for further evaluation.
- Chapter 4, **Data Collection and Preliminary Insights**: delves into the analysis of social media posts from the selected HEIs over an academic year. This chapter begins by detailing the criteria of the chosen period, followed by an in-depth exploratory data analysis. Thus, uncovering patterns and key trends within the posts,

and providing insights into each institution's communication patterns and engagement levels.

- Chapter 5, **HEIs Grouping and Analysis**: classifies the selected higher education institutions into distinct groups based on their characteristics. It then aims to identify patterns and variations in the approaches employed by each HEI, relating these to their positions within the selected global ranking.
- Chapter 6, **Post Categorization**: involves defining a set of categories to analyze the thematic content of the posts published during the period chosen. This categorization aims to classify publishing strategies but also identify and predict emerging trends and patterns.
- Chapter 7, **Interpretation and Evaluation of Results**: provides a concise summary of the key findings presented in the previous chapters, offering a holistic analysis and its significance.
- Lastly, Chapter 8, **Conclusion**: will address the research questions elaborated in the Introduction describing the research process and present the final conclusions inferred from this study.

2. Literature Review

In recent years, social media has emerged as a powerful platform for communication, marketing, and engagement, impacting how organizations, including educational institutions, interact with their audiences. [10]. This change has led to different studies on how social media works and affects various sectors. Thus, it reveals varied strategies and impacts but also demonstrates an in-depth awareness of how these platforms can be optimized within the educational sector. This dissertation employs "backward citation searching" to track down relevant articles as well as a straightforward search.

A study by Oliveira et al. [11] explored ways to improve social media strategies for businesses, providing insights that are equally applicable to education contexts. The study analyzed user interactions and feedback from Higher Polytechnic Portuguese Institutions, grouping them into clusters by types of social media content strategies and the input of the Key Performance Indicators (KPI). The findings suggest that tailored content boosts engagement levels.

Similarly, Peruta et al. [12] examined the social media strategies of top U.S. colleges and universities based on Facebook posts. It explored how universities can conduct social media to strengthen community ties and foster communication between students and administration. Further, it emphasizes the value of strategic content creation in building a cohesive brand identity but also offers insight for university social media managers on how to optimize their content and posting strategies to enhance engagement and expand their online reach.

Sengupta et al [13] research delves into the significant increase in social media usage in higher education during the COVID-19 pandemic. Their research delves into the analytics of social media trends, utilizing different methodologies including bibliometric analysis, thematic maps, and the Analytic Hierarchy Process (AHP). By analyzing both primary and secondary data, and consulting various stakeholders through virtual brainstorming sessions and the Delphi method, the study offers a comprehensive framework that helps educational institutions predict and adapt to the evolving digital environment. This framework provides tools for these institutions to anticipate student needs and preferences better, thereby improving interaction and support during such disruptive periods.

Additionally, Santos et al. [14] further explored how text mining and sentiment analysis can reinforce the international appeal of HEIs. Their study provides targeted communication strategies that enhance institutional attractiveness by highlighting the positive attributes most valued by students. This approach aids universities in improving their global competitiveness and appeal, ensuring they effectively meet the expectations of prospective international students.

More recently, Capriotti et al. [15] developed a model to assess the impact of social media engagements on organizational performance. This model enables institutions to measure the effectiveness of their social media strategies in real-time, facilitating a dynamic and responsive approach to digital marketing and communication. The findings reveal variations in publication frequency and content types across different regions and platforms, noting that Facebook exhibits higher than average daily activity, particularly in Latin America.

Together, specifically the last three articles, exemplify the varied aspects of how digital communication affects the operational and strategic objectives of HEIs, thereby serving as a motivation for the research we have conducted.

2.1. World University Rankings

As we consider the impact of social media on educational institutions, it becomes important to determine whether there is a significant correlation between social media publishing strategies and global university rankings. Different rankings use different metrics to evaluate higher education institutions which results in different attributions of different positions. To assess the extent of agreement among various global university rankings, similarity metrics are employed to understand how closely these rankings align in their evaluations. For instance, Shehatta et al. [16] investigate the correlation between various ranking systems, using the number of overlapping elements and Pearson's/Spearman's correlation coefficients. Additionally, Çakir et al. [17] conducts a systematic comparison of national and global university ranking systems in terms of their indicators, coverage, and ranking results, applying Aguillo et al. [18] methodology for calculating the similarity among two rankings, Inverse Rank (M) measure. Furthermore, Figueira et al. [19] take a further step by not only assessing the correlation but also employing clustering techniques. This approach categorizes rankings into groups based on their similarity, providing a subtle understanding of the robustness and consistency of rankings and publishing strategies. Consequently, we aim to determine which ranking system provides the most consensus, thus enhancing our understanding of university rankings and social media strategies.

2.2. Cluster Analysis

Clustering is an unsupervised machine-learning technique used to group unlabeled objects based on their inherent similarities. Unlike supervised learning, which relies on known labels to train models, clustering identifies patterns and natural groupings in data without prior labeling. The goal is to ensure that objects within each cluster are more similar to each other than to those in other clusters, thereby uncovering hidden patterns or groupings that might not be immediately obvious [20]. It can be accomplished through several algorithms, including k-means, a straightforward yet powerful algorithm [21]. K-means is a type of partitioning clustering that divides a dataset into k distinct, non-overlapping clusters. It starts by randomly selecting k points as cluster centers and assigns each object in the dataset to the nearest cluster center based on similarity measures such as Euclidean distance. Then, the centroid of each cluster becomes the new cluster center. This process repeats iteratively until the cluster assignments no longer change, indicating that the clusters are as homogenous as possible.

Expanding on this, recent research such as that of the KS et al.[22] explore the impacts of social media on mental health, demonstrating how clustering can reveal patterns between social media habits and levels of anxiety and depression among the users. Crockett et al. [23] review various algorithms for clustering Twitter publications, demonstrating how these techniques can be applied to analyze social media for different purposes, including marketing and sentiment analysis. Furthermore, Guzman et al. [24] delve into the use of cluster analysis in higher education institution settings, in knowledge identification and production processes. Thus aiding institutions in improving their research strategies and education offerings.

2.3. Natural Language Processing (NLP)

With the growing volume of text-based content on these platforms, Natural Language Processing (NLP) has become an essential tool for analyzing and interpreting the vast amounts of unstructured data generated. NLP techniques enable the extraction of meaningful insights from social media posts, such as sentiment analysis, topic detection, and user behavior modeling, thereby enhancing our understanding of digital communication trends and their impact on various sectors. In NLP, the text data organized into datasets is known as a corpus, with documents serving as the sources of this data, such as books, news articles, and emails.

To optimize the performance of NLP models, the raw data within a corpus must first go through pre-processing steps. A few of the pre-processing techniques which will be mentioned in the following chapters include:

- Tokenization - involves breaking down the text into individual elements, called tokens, which can be words, phrases, and sentences. This process helps in structuring the text for further processing and analysis
- Lemmatization - consists of an alternative for reduction dimensionality. It takes into account the context of the word in order to apply a morphological analysis, making sure the root word (lemma) belongs to the language.
- Stop words removal - removes the stop words. Stop words are a set commonly used in any language, such as "the", "for", "with" and others for the English Language that do not contribute to the overall meaning. This elimination process reduces the noise and dimension in textual data, making it easier to identify the most important keywords and phrases.

Once the text has been pre-processed, we can then apply advanced techniques such as topic mining. This method allows us to identify and categorize recurring patterns or topics across vast datasets, thereby providing deeper insights into the content's underlying themes. Through algorithms and models, topic mining helps reveal the underlying structures in the data, offering a deeper understanding of the content. This makes this technique particularly valuable in social media analytics, where identifying trends and topics can provide insights into public opinion and cultural patterns.

2.3.1 Topic Modeling

Delving further into topic mining, two important methods are encountered: Topic Modeling and Topic Classification. Topic Modeling is a technique used to extract and identify hidden themes that are not immediately observable in the document but instead suggested by a collection of words (latent topics) within a corpus of unlabeled documents. An algorithm used in this process is Latent Dirichlet Allocation (LDA) as described by Blei et al. [25]. LDA is a generative probabilistic model that assumes each document is a mixture of a finite number of topics each characterized by a distribution of words. This algorithm works by assigning every word in the documents to a temporary topic, which is iteratively adjusted based on the words in the document and the overall topic distribution across the corpus

until the topic distribution stabilizes [26, 27]. The number of latent topics is user-defined, allowing for a tailored analysis.

This method of topic identification has broad applications, as evidenced by Strydom et al. [28], who explore the effectiveness of topic modeling within the social media market to measure consumer sentiments and trends. Similarly, Maphosa et al. [29] illustrates the application of LDA in analyzing textual data across higher education platforms to enhance curriculum development and academic research. Furthermore, the systematic review by Laureate et al. [30] underscores the versatility of topic modeling, showcasing its utility across various fields by systematically collating studies that have successfully employed this technique to extract meaningful insights from large datasets. Another contribution is by [31] which further validates the efficacy of the LDA model in extracting topics from complex textual information, thereby facilitating more accurate and insightful data interpretation across diverse research domains.

2.3.2 Topic Classification

Topic classification is distinguished by the capability of categorizing documents into predefined topics, a process distinct from topic modeling, which uncovers latent themes without prior labels. This method relies on supervised learning techniques, in which the model is trained on a corpus of documents that have already been categorized, using these labels to learn how to classify new, unseen documents accurately.

A fundamental aspect of topic classification is word embedding, which transforms individual words into a numerical representation, known as vectors. A technique named Word2Vec, by Mikolov et al. [32] generates word embeddings by combining a group of two-layer networks. Each vector captures the meaning of the word based on the surrounding words. By doing so, it ensures that words with similar meanings are positioned closely in vector space, facilitating more complex text analysis.

Building on the concept of word embeddings, advancements in topic classification have been achieved through the use of BERTopic, a technique influenced by the capabilities of BERT (Bidirectional Encoder Representations from Transformers) [33]. BERT outdoes at understanding the context of a word concerning all the other words in a sentence, allowing the model to weigh the importance of each word in the sentence differently.

The process of implementing BERTopic is highly modular and can be divided into five steps, being the last optional. Initially, the model converts all documents in a corpus into

dense vectors using BERT. Thus capturing semantic meanings and contextual relationships between words within the documents. Following this, the high-dimensional embeddings are reduced using techniques such as Uniform Manifold Approximation and Projection (UMAP). This step aids the clustering process by simplifying the complex structures in the data while retaining essential characteristics. Once the dimensionality is reduced, for example, HDSCAN is employed to cluster these embeddings, which groups similar document vectors based on their density in the vector space. Each cluster formed in this step represents a potential topic. The subsequent step involves applying class-based TF-IDF (c-TF-IDF) which refines the traditional TF-IDF approach by calculating terms important specifically within the context of each topic cluster rather than across the entire document corpus. The final, optional step allows for fine-tuning the topics. This can be done by merging similar topics, splitting broad topics, or discarding irrelevant ones, thus enhancing the clarity and relevance of the topics identified.

The flexibility and accuracy of BERTopic have been demonstrated in numerous studies. For instance, Mendonça et al. [34] showcase the modularity of this model, demonstrating how it can be adapted and fine-tuned. De et al. [35] and Egger et al. [36] further explore the integration of deep learning technologies like BERT into the topic classification process, revealing significant enhancements in analytical capabilities. BERTopic's applicability extends into various fields, particularly in analyzing social media content. Futterer et al. [37] demonstrates how BERTopic can be used to analyze conversations and trends in social media platforms, which is critical for understanding public sentiment and behavioral patterns. Similarly, Grigore et al. [38] employed transformer-based topic modeling to measure the severity of eating disorder symptoms, while Schneider et al. [39] to detect hate speech and its targets. Additionally, Antypas et al. [40] focus on the same social media platform to uncover trends and topics in real-time. Collectively, these studies demonstrate the potential of extracting and interpreting themes from the corpus of the text across diverse applications.

These capabilities make BERTopic an excellent tool for our project, where we aim to systematically categorize posts from different institutions into distinct topics. Therefore, providing a clear understanding of the predominant discussions, concerns, and communication strategies within each institution.

2.4. Long Short-Term Memory (LSTM)

Long short-term memory (LSTM) networks are a supervised machine learning algorithm that specializes in sequence prediction tasks, capturing long-term dependencies where order is crucial [41]. This is achieved through the network's unique architecture design, which involves a memory cell that is controlled by three gates to regulate the flow of information:

- Input gate - controls what information from the input should modify the memory state, selecting which information is significant to retain for long-term memory.
- Forget gate - it helps optimize the network's performance since it determines what portions of information of the memory cell to erase, allowing it to discard irrelevant information.
- Output gate - filters the output information that goes through the memory cell to the output of the network unit.

Recent studies emphasize the relevance and utility of LSTM in complex predictive tasks across various domains. For instance, Lasri et al. [42] demonstrate how self-optimizing LSTM networks can enhance predictive accuracy in dynamic environments, similar to forecasting academic schedules and resource requirements in HEIs. Pandey et al. [43] examine the latest developments in natural language processing, focusing on how the BERT model can be enhanced by integrating LSTM to better understand and predict textual data within academic settings. Additionally, Zhou [44] explores innovative applications of LSTM in forecasting technological trends, suggesting that similar methodologies could be adapted to forecast educational trends and needs.

This architecture is useful for applications where understanding the order and context of data over time is essential. In the context of HEIs, such capabilities can be particularly valuable, especially when predicting categories for the last five days of each month.

3. Towards a Universal Ranking Standard

In the dynamics of world university rankings, it is crucial to recognize the inherent diversity of opinions and methodologies that underpin these assessments. Each ranking system is designed with different objectives, criteria, and data sources, consequently causing the resulting rankings to diverge [45] and, leading to variations in the positions of each university. This divergence leads to variations in the positions of each university, underscoring the difficulty of comparing institutions directly. As such, achieving a consensus on university performance is intricate, necessitating a more subtlety approach to interpretation. As a result, to identify the ranking with the greatest similarity, five world university rankings were selected to conduct a comparative analysis of the varying positions occupied by higher education institutions across these ranked lists.

3.1. World University Rankings

Five university rankings were considered in this study:

- Times Higher Education (September 2023) [46]
- Quacquarelli Symonds (June 2023) [47]
- Center for World University Rankings (May 2023) [48]
- Academic Ranking of World Universities (August 2023) [49]
- Webometrics Ranking of World Universities (July 2023) [50]

The Times Higher Education World University Rankings, often referred to as THE is the annual publication by Times Higher Education magazine. Previously, the magazine jointly with Quacquarelli Symonds (QS) published the THE-QS World University Rankings from 2004 to 2009 when the two organizations parted ways producing independent rankings. THE Ranking considers 17 performance indicators that can be grouped into 5 groups, teaching (29.5%), research environment (29%), research quality (30%), international outlook (7.5%), and industry (4%). THE provides a comprehensive assessment of 2671 HEI across different countries and regions.

The QS Ranking features 1498 HEI across 104 locations. In 2023, in comparison with previous years, three new metrics were added to the methodology: sustainability (5%),

employment outcomes (5%), and international research network (5%). The remaining indicators consist of 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), also known as the Shanghai Ranking, is one of the most renowned and oldest global university rankings. Originally developed by the Center for World-Class Universities (CWCU) at Shanghai Jiao Tong University, since 2009 it has been published and copyrighted by Shanghai Ranking Consultancy. ARWU uses 6 main indicators, namely, quality of education (number of the alumni of an institution winning Nobel Prizes and Fields Medals - 10%), number of staff elements winning awards (20%), number of highly cited researchers selected by Clarivate (20%), number of papers published in Science Citation Index-Expanded and Social Science Citation Index (20%), number of articles indexed in SCI-Expanded and SSCI (20%) and per capita academic performance of an institution (10%). The Shanghai Ranking assesses more than 2500 HEI annually, ranking the top 1000.

Center World University Rankings (CWUR) is distinguished for ascertaining the quality of education, alumni employment, quality of faculty, and research performance without relying on surveys and university data submissions. CWUR uses 7 indicators grouped in 4 areas, such as education (25%), employability (25%), faculty (10%), and research (40%). CWUR provides a comprehensive assessment of 2000 universities across 108 countries and regions.

And, finally, Webometrics Ranking of World Universities (WEB) is published by the Cybermetrics Lab i.e. a research group of the Consejo Superior de Investigaciones Científicas (CSIC), recognized as the largest Spanish public research body. To evaluate the different HEI, this ranking uses 3 indicators as the normalized and average value of the number of external networks (subnets) linking to the institution's webpages (50%), top-cited researchers from Google Scholar profiles and top cited papers, excellence (40%). In 2023 it ranked 12001 HEI.

3.2. Similarity Metrics

To ensure an objective comparison among the previous rankings in Section 3.1, each with varying degrees of coverage, three similarity measures were selected. These measures are Overlap Size (OC), Spearman's Footrule (F), and M Measure [51].

The Overlap Size (OC) counts the number of elements that the two top- k lists have in common. Despite its simplicity, it allows us to analyze the number of elements present in each of the top- k lists 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 [52, 53] as

$$Fr^{|Z|}(\sigma_1, \sigma_2) = \sum_{i=1}^{|Z|} |(\sigma_1(i) - \sigma_2(i))|. \quad (3.1)$$

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 2. 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

$$NFr = \frac{Fr^{|Z|}}{\max Fr^{|Z|}}. \quad (3.2)$$

This quantity is normalized in the sense that dividing by its maximum value guarantees that it ranges from 0 to 1, corresponding to exactly equal and maximally different lists, respectively. When both ranked lists are identical, the resulting value is 0. Hence, for calculating the desired similarity value, F is defined as:

$$F = 1 - NFr \quad (3.3)$$

Another metric was used to consider all elements of the ranked list, the Inverse Rank (M) measure. Originally, a similarity measure was proposed [54] which involved extending

the footrule by assigning a rank to non-overlapping elements. Consequently, when comparing the similarity of two top-k lists to the non-overlapping elements are attributed k+1 position. However, this approach assigns excessive weight to non-overlapping elements, regardless of their position within the lists. Therefore, an improved version has been developed to address this limitation [51]. The formula for its normalized value is computed as

$$M^{(k)} = 1 - \frac{N^{(k)}}{\max N^{(k)}} \quad (3.4)$$

where

$$\max N^{(k)} = 2 \sum_{i=1}^k \left(\frac{1}{i} - \frac{1}{k+1} \right) \quad (3.5)$$

and

$$N^{(k)}(\sigma_1 \sigma_2) = \sum_{i \in Z} \left| \frac{1}{\sigma_1(i)} - \frac{1}{\sigma_2(i)} \right| + \sum_{i \in S} \left| \frac{1}{\sigma_1(i)} - \frac{1}{(k+1)} \right| + \sum_{i \in T} \left| \frac{1}{\sigma_2(i)} - \frac{1}{(k+1)} \right|. \quad (3.6)$$

$\sigma(i)$ 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 is 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 [18] is employed, which involve

$$\text{Mid Position} = \frac{2 \times k + (n - 1)}{2} \quad (3.7)$$

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

3.2.1 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 as QS, THE, ARWU, CWUR, and WEB.

Table 3.1: Similarity measures for 2023 world university rankings

2023	THE			QS			CWUR			ARWU			WEB		
	OC	<i>F</i>	<i>M</i>	OC	<i>F</i>	<i>M</i>	OC	<i>F</i>	<i>M</i>	OC	<i>F</i>	<i>M</i>	OC	<i>F</i>	<i>M</i>
Top 10															
THE				7	0.58	0.45	7	0.58	0.47	8	0.56	0.54	5	0.50	0.49
QS	7	0.58	0.45				5	0.17	0.47	6	0.22	0.36	5	0.33	0.34
CWUR	7	0.58	0.47	5	0.17	0.47				8	0.81	0.85	6	0.89	0.76
ARWU	8	0.56	0.54	6	0.22	0.36	8	0.81	0.85				6	0.89	0.86
WEB	5	0.50	0.49	5	0.33	0.34	6	0.89	0.76	6	0.89	0.86			
Top 100															
THE				78	0.71	0.62	73	0.65	0.60	75	0.71	0.65	71	0.59	0.58
QS	78	0.71	0.62				63	0.60	0.55	61	0.60	0.52	64	0.57	0.47
CWUR	73	0.65	0.60	63	0.60	0.55				78	0.74	0.83	72	0.68	0.73
ARWU	75	0.71	0.65	61	0.60	0.52	78	0.74	0.83				70	0.70	0.79
WEB	71	0.59	0.58	64	0.57	0.47	72	0.68	0.73	70	0.70	0.79			
Top 200															
THE				158	0.72	0.64	158	0.68	0.62	151	0.70	0.67	145	0.66	0.60
QS	158	0.72	0.64				139	0.62	0.57	134	0.60	0.54	132	0.58	0.50
CWUR	158	0.68	0.62	139	0.62	0.57				171	0.73	0.83	151	0.71	0.73
ARWU	151	0.70	0.67	134	0.60	0.54	171	0.73	0.83				139	0.65	0.77
WEB	145	0.66	0.60	132	0.58	0.50	151	0.71	0.73	139	0.65	0.77			

THE - Times Higher Education

QS - Quacquarelli Symonds

CWUR - Center for World University Rankings

ARWU - Academic Ranking of World Universities

WEB - Webometrics Ranking of World Universities

OC - The size of the overlap

F - Spearman's Footrule

M - M measure

As can be seen in Table 3.1, The values of similarity metrics for the different rankings tend to increase as we consider longer lists. Nearly all results surpass the 50% threshold except for a few occurrences in the top 10 for the *M* measure. Despite the relatively

small sample size, it is clear that the number of overlaps exceeds half of the sample under analysis. However, the findings at the M measure reveal low results, indicating that although the institutions being analyzed remain consistent, their rankings differ. Thus, presenting a notable consensus among the top 10 institutions, regardless of the diverse metrics used to assess them.

It is important to highlight that in Spearman's footrule metric, the worst result obtained, corresponds to the comparison between QS and ARWU (0.17). This result stems from the HEIs shared by both rankings, encompassing the top 5 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. Also, visualizing data through heatmaps (Figure 3.1), it is possible to note that for a sample of 200 HEIs, the QS presents the worst results for the M measure, visible to its bright colors. Out of 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.

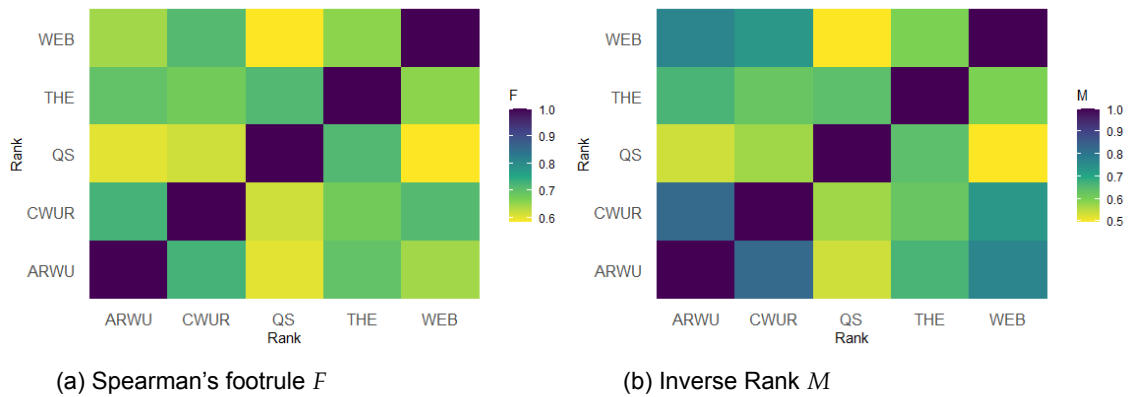


Figure 3.1: Resulting values between the similarity measures for the top 200 positions in the university rankings.

Remarkably, 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. It also consistently achieves average results above 60% for the M measure metric across different top rankings analyzed. Consequently, when

considering the importance of the Inverse Rank metric, which takes into deliberation the overlapping and non-overlapping institutions, it becomes evident that WEB is slightly better than THE.

For the three similarity measures, the best results for the top 100 and 200 are found between ARWU and CWUR. In the top 100, these two rankings share 78 institutions in common, while in the top 200, they share 171 institutions in common, 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%. We thus infer that these rankings present a high degree of similarity with 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's noteworthy to point out that 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 (mid position). Therefore, these results should only be taken into consideration with this assumption in mind. We thus determine the ranking with the best similarity as ARWU with the median result of 145 number of overlapped HEIs, with QS presenting the worst with 136.5. This high level of agreement indicates that the typical performance level of institutions is recognized despite the varying criteria used by different ranking systems.

3.3. The sample for HEIs

This similarity analysis serves to identify the world university ranking that best encompasses the general positioning. This is used as a reference to characterize different editorial standards from the HEIs by comparing their respective publishing strategies. Since the previous results between ARWU and CWUR are quite similar, CWUR is ideal for this purpose. This is especially true as CWUR is a ranking that does not allow for ties between the institutions under analysis, allowing for a more thorough distinction.

In this work, we aim to extract publication patterns from different institutions and relate these to their position in global rankings. Due to space restrictions, we are not able to present an analysis for every single university, forcing us to tactically choose the best set of these, which would ideally mimic the trends of the more general list. This is best captured by comparing institutions in approximately comparable geo-political scenarios.

We thus take examples in batches from different countries while also comparing nearby positions in the ranking. The chosen institutions are

Table 3.2: 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

4. Data Collection and Preliminary Insights

To analyze publication trends across diverse Higher Education Institutions (HEIs) and discern their correlation with respective positions in global rankings, we extracted the posts from each HEI from the social network X (Twitter). To ensure the most adequate period, we carefully considered the academic calendars of the HEIs in different countries, recognizing that these vary significantly.

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 depth 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 a summer term, which is typically optional and designed for students who need or want to take additional courses. This often results in American HEIs having a higher number of academic terms, typically ranging from three to four academic terms. In contrast, the EPFL in Switzerland offers a distinctive summer term, where students can immerse themselves in the vibrant scientific community through eight weeks of laboratory experience [55]. This not only enhances their educational experience but also connects them with leading researchers and industry professionals, fostering a practical understanding of scientific careers. The visual below displays the varied academic calendars of the institutions analyzed, illustrating the distinct timelines:

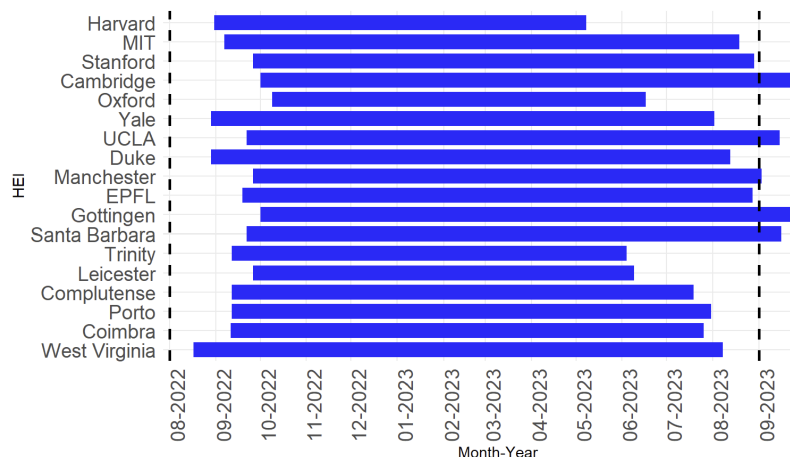


Figure 4.1: Academic calendars of selected higher education institutions

To determine the ideal timeframe for our analysis, the academic calendar for each HEI was considered, as cited in [56–73]. Consequently, to capture the entirety of an academic year, 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, enabling observation of the full cycle of university operations, from welcoming new students to campus to the culmination of the academic year.

4.1. 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 the communication patterns and engagement levels of each institution. The analysis of the data in Figure 4.2, reveals that the University of Oxford is the institution with the highest frequency of publication, significantly more than Complutense University of Madrid, which has a substantially lower rate of tweet activity. Stanford, despite its third position in the global CWUR ranking, seems to deviate from the behavior of nearby institutions in the top 5 positions. It only published 351 posts during this period, making it the sole institution not to exceed 1000 tweets. Additionally, it ranks as the second least active institution in terms of publications.

Moreover, nearly all the universities have published more tweets than replies and retweets, deliberating on choosing a strategy that prioritizes broadcasting information rather than actively engaging with its followers or other users on the platform. With this behavior, MIT stands out with one reply while the University of Porto has one retweet, indicating that the majority of its activity consists of publishing original content. In contrast, in the University of Manchester, out of 1435 posts, 784 consist of replies, and the other 651 tweets, which makes it an exception among the other institutions analyzed since the majority consists of replies, representing 55% of the data. Thus, making the institution the one with the higher percentage of replies. In addition, among Higher Education Institutions, only four have all three types of possible posts: the University of Porto, University of California Santa Barbara, University of Gottingen, and EPFL, which exhibit the most conscious distribution among them.

Furthermore, Complutense University of Madrid suggests a more consistent posting strategy, with the lowest standard deviation and an average of approximately one post per day (Table 4.1). Conversely, the University of Cambridge demonstrates the highest standard deviation (8.29), indicating a greater degree of variability or dispersion in publication frequency compared to other HEIs.

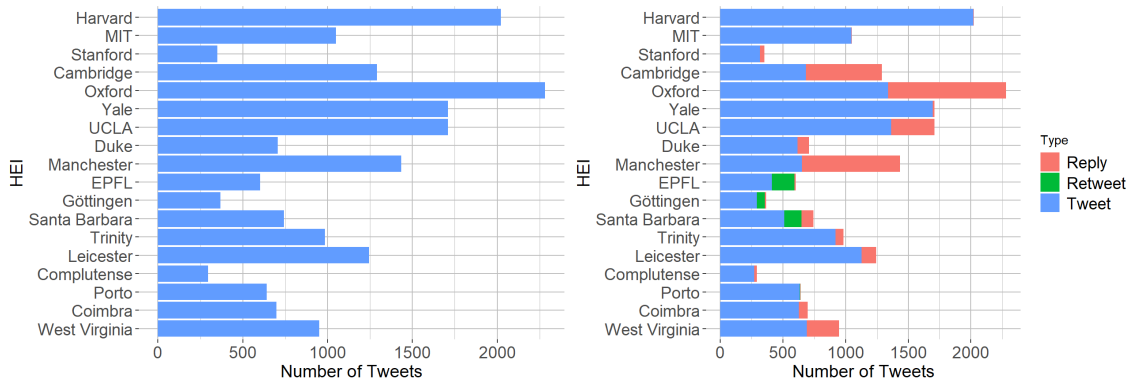


Figure 4.2: Data distribution between August 2022 and August 2023

Table 4.1: Average, highest frequency and standard deviation of daily publication frequency of HEIs

Rank	Higher Education Institution	Average	Highest Frequency	Std (post/day)
1	Harvard University	5.16	19	1.45
2	Massachusetts Institute of Technology	2.75	15	1.52
3	Stanford University	1.70	9	1.26
4	University of Cambridge	3.86	106	8.29
5	University of Oxford	5.81	64	6.67
10	Yale University	4.32	13	2.35
18	University of California, Los Angeles	4.50	94	6.31
20	Duke University	2.47	9	1.51
50	University of Manchester	4.00	54	4.17
96	École Polytechnique Fédérale de Lausanne	2.30	11	1.51
97	University of Göttingen	1.64	6	0.76
98	University of California, Santa Barbara	2.29	26	1.92
247	Trinity College Dublin	3.14	12	2.10
248	University of Leicester	3.21	19	2.37
249	Complutense University of Madrid	1.25	4	0.54
309	University of Porto	1.95	5	0.95
420	University of Coimbra	2.53	8	1.61
501	West Virginia University	2.62	24	2.97

4.1.1 Data Distribution

It becomes evident that not all institutions adhere to similar patterns. For example, American HEIs tend to have a reduction in the number of publications between October and December, a phenomenon that can be correlated with the midterm elections occurring during that time. Moreover, institutions in the USA consistently demonstrate a higher publication frequency during the month of open house for admitted students in March, as well as during the commencement of activities, typically between May and June. Also in March, other institutions such as the case of the University of Göttingen, EPFL, and the Complutense University of Madrid presented an increased number of postings, similarly in May and June at the end of the academic year.



Figure 4.3: Monthly tweet distribution over the selected period

This variation contrasts with the University of Cambridge, where the highest frequency of

posts was notably concentrated in August of both 2022 and 2023, aligning with key academic milestones such as A-Level Results Day. Such strategic timing was designed to maximize engagement with prospective students, as evidenced by the data distribution. Similarly, UCLA experienced a significant peak in March, driven by interactive tweets that effectively boosted engagement through direct follower interaction. This pattern is consistent with the highest frequencies observed in Table 4.1, which further points out the consistent and high average posting frequency of Harvard throughout the year. The peaks observed around October and March further emphasize the role of strategic communication efforts in maintaining high levels of engagement year-round, particularly during key academic events.

4.1.2 Publication Period by Hour, Day and Month

As shown in Figure 4.4, and further detailed in the appendix (Figure 1), the analysis considered the time zones corresponding to each institution's location, providing insights into their specific timing behaviors. While most institutions tend to publish during traditional working hours, variations in publishing times are evident across different universities. For example, Yale and Manchester prefer morning publications, while Trinity and Leicester opt for the afternoon.

MIT, despite its higher frequency of publications during August, exhibits a unique and narrower publication window, predominantly posting between 11 am and 3 pm on weekdays, with weekend posts occurring only around 11 am. Additionally, both MIT and Stanford show a slight tendency to post between 2 am and 6 am, deviating from the usual working hours of other American institutions and their respective local times.

Publishing strategies during late hours, particularly between 8 pm and 11 pm, also show significant differences. Some institutions, such as Duke, completely refrain from posting during these hours, while others, like Complutense, maintain their peak post rate almost unaltered during this period.

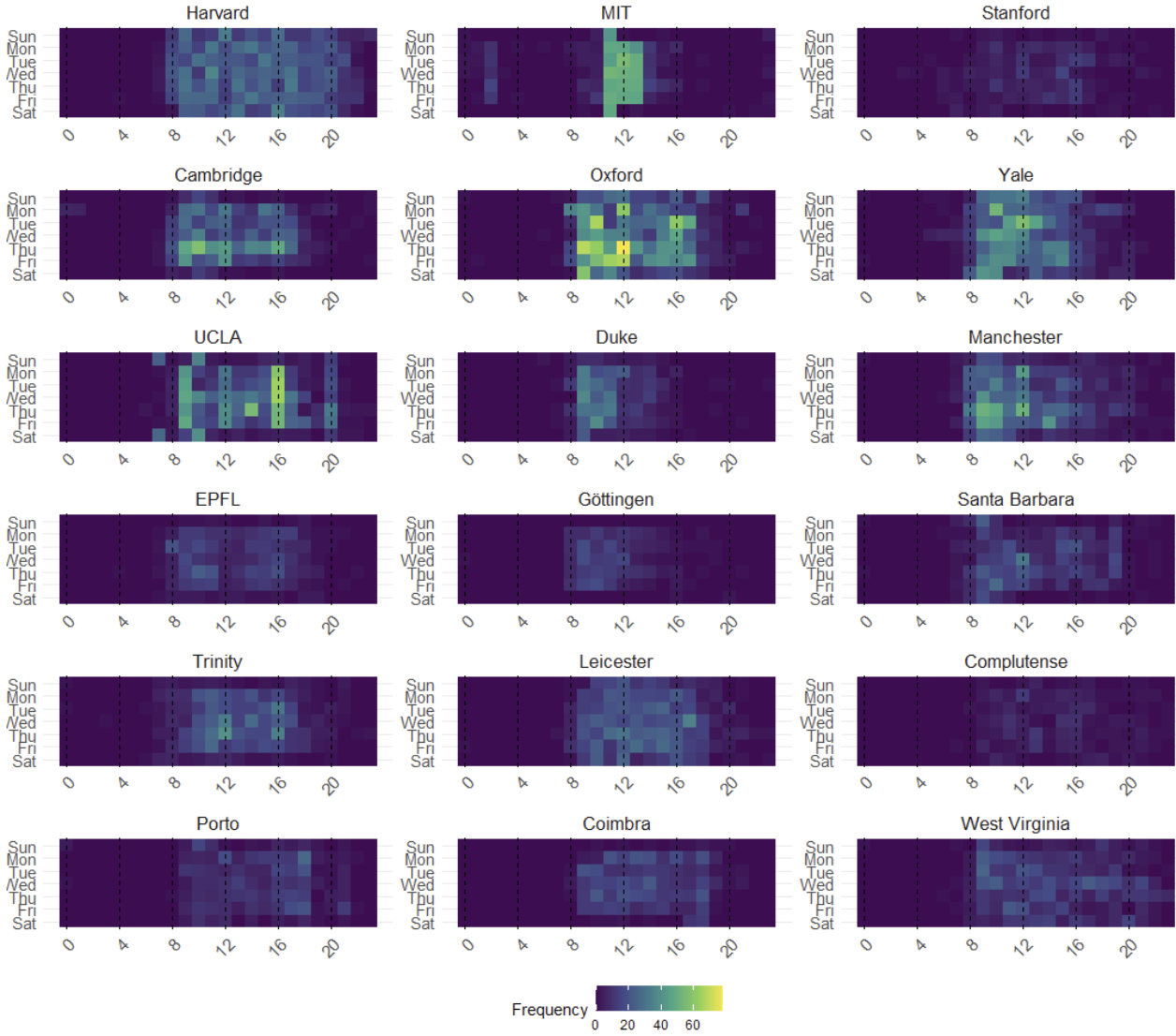


Figure 4.4: Hour versus day of the week of HEIs

4.1.3 Retweet and Favorite

The audience's response to the content was also evaluated. For that purpose, this involved examining the number of retweets and favorites which are the engagement forms the social media offers, enabling us to identify the institution with the greatest success or reach a wider audience. Generally, followers tend to interact with their favorite feature rather than with retweets. As the ranking decreases, the level of interaction decreases as well, although there is some correlation between the retweet and favorite distributions. Harvard, MIT, and Stanford exhibit the highest median counts for both favorites and retweets, with wide Interquartile Range (IQR), indicating a broad distribution of engagement. Regarding favorite feature distribution, except for the top five institutions that show significantly higher engagement, universities generally display a broader range of

favorite counts. Outliers like Complutense and West Virginia do not follow the typical pattern, presenting unique engagement levels despite their lower rankings. Notably, the visualizations highlight some exceptions. For instance, despite being ranked 249th, the Complutense University of Madrid has a retweet distribution similar to a more renowned university, in the top 10. Similarly, regarding favorites and retweets, Complutense University shows comparable behavior to West Virginia. These instances suggest that rank position alone does not always directly correlate with social media engagement metrics.

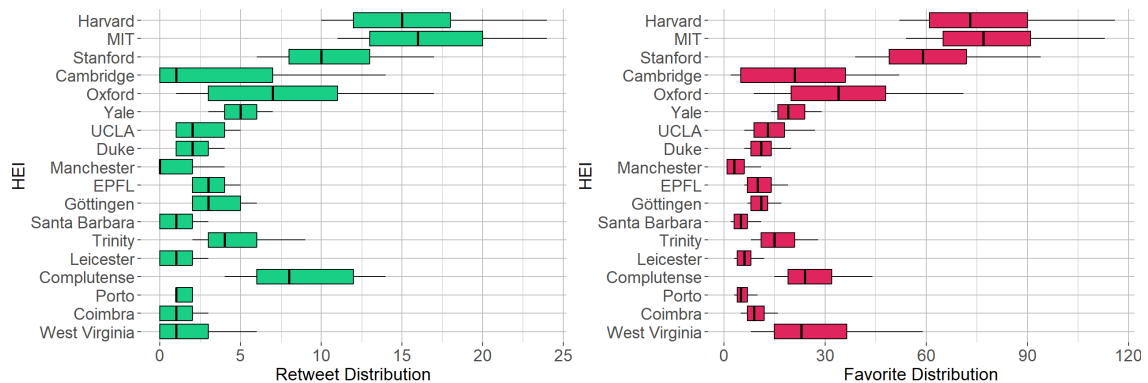


Figure 4.5: Retweet and favorite distributions by HEI

4.1.4 Media Type, Hashtags, and URL

In a detailed examination of digital engagement across various institutions, as shown in Figure 4.6, three categories were analyzed: media types (such as photos, videos, and animated GIFs), hashtags, and URLs. Among these institutions, Leicester emerged as the foremost performer, consistently excelling across all three categories. While Leicester led in overall performance, MIT distinguished itself with a particularly high usage of photos, suggesting a strategic emphasis on visually engaging content.

The analysis also features that URLs are the most commonly used content type across the board, reflecting their critical role in facilitating link-sharing and directing audiences to further resources. Despite this general trend, Manchester diverged significantly, with notably lower usage of URLs, media types, and hashtags than other institutions.

Additionally, Coimbra was identified as an outlier with its remarkably high usage of hashtags, a behavior that stands in contrast to the more uniform practices observed among similarly ranked institutions. In contrast, Trinity and Leicester displayed similar patterns of high engagement across all measured URL categories, reinforcing Leicester's position as the leader in digital engagement.

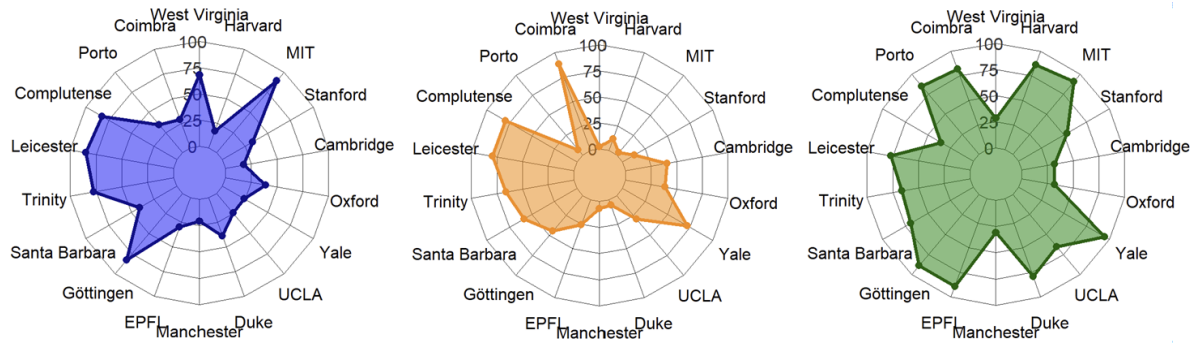


Figure 4.6: Average distribution of Media Type, Hashtag, and URL usage per HEI

4.1.5 Emojis

The incorporation of emojis has emerged as an interesting dimension in digital communication. The HEIs selected in our analysis have embraced emojis for their visual appeal and as a deliberate method of textual expression. The overall number of emojis displayed in each post falls within the interval from 0 to 24, with two specific outliers associated with the University of Manchester. These outliers are correlated posts, consisting of a tweet and a reply on the same day (August 18, 2022), with only a few seconds between them. On this day, the students received the news of their acceptance into the university.

In the following scatter plot, Figure 4.7, we depict the top 5 emojis encountered in the considered HEIs, with these being: clapping hands, backhand index pointing right, down arrow, backhand index pointing down, and blue heart. It is noteworthy that the institution that contributed the most to its usage count was the University of Porto, likely due to its prominent overall usage of emojis. However, despite the high volume of emojis, it also exhibits notable diversity in emoji selection as the usage of more than one emoji from the top in the same post, as evidenced by the figure. On the contrary, the University of Oxford displays the highest density of occurrences of the top 5 with the down arrow being the most used, contributing to 87% out of the total of times the emoji had been used by all the universities. This frequent use of this emoji typically serves as a visual tool to direct attention to key information, enhancing data presentation and reader engagement. Additionally, it is important to note that Harvard is not present in this visualization. Although it does use emojis, none of them are within the top 5 overall among the 18 HEIs considered, and it has a relatively low percentage of emoji usage, around 6%.



Figure 4.7: Top 5 emojis through the period selected by month

4.1.6 Counting Characters

X/Twitter is known for its unique constraints on message length, then forcing individuals and organizations to convey their ideas and updates with brevity and precision. Since 2017, users of this social platform have been restricted to a maximum of 280 characters per tweet until February 2023. Following this date, accounts that subscribed to Twitter Blue, exclusively available in the USA, saw their character limit increase to 4000. Then, in March of the same year, the Twitter Blue subscription became available worldwide and has remained so since.

The count of characters on Twitter follows a specific set of rules [74], which dictate how different elements within a tweet contribute to the overall character count. For example, when composing replies, @names that auto-populate at the start of a reply tweet and URLs related to media attached to the post are excluded from the character count. Additionally, emoji sequences composed of multiple combining glyphs count as 2 characters, while all other characters, including those with diacritics, count as 1 character each.

Given the selected timeframe coincides with changes in character limits, a study was conducted to assess how various institutions adapted to these modifications. Overall, the majority of institutions published posts below the 280 characters threshold allowed by the free version. However, there were exceptions, like MIT, in Figure 4.8a, which stood out with a specific outlier. This outlier occurred on July 1, 2023, during the "One-MIT" commencement activities involving undergraduates and graduate students with the new college president. The tweet described the president's robe in detail, highlighting its unique features such as the two cardinal red stoles adorned with 18 infinite symbols, symbolizing MIT's 18th presidency. Alternatively, The Complutense University of Madrid

appears to have subscribed to Twitter Blue in March and continued until the end of the selected period since the threshold of 280 characters was exceeded, visible in Figure 4.8b. A similar behavior was observed at the University of Manchester, but only for the single month of August.

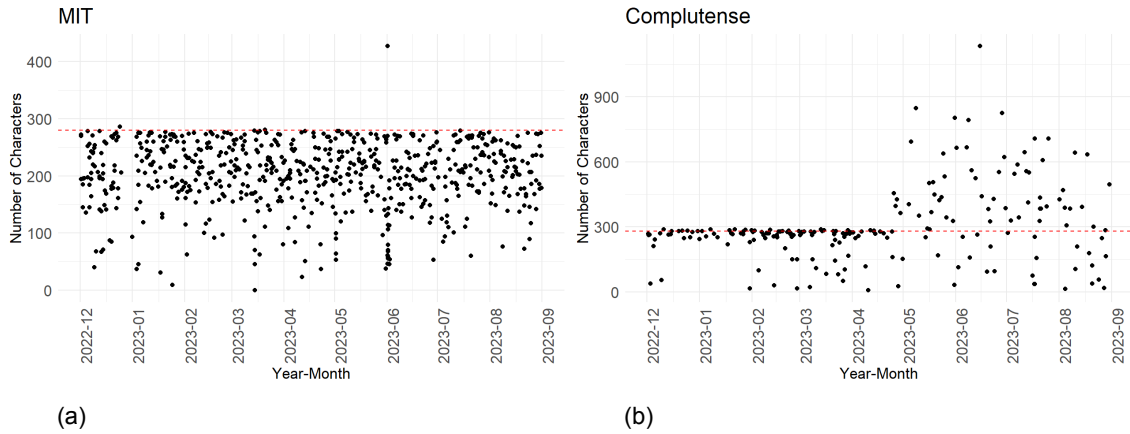


Figure 4.8: Number of characters during the period selected by month

4.1.7 Sentiment and Emotion Analysis

To further explore the context of tweets a sentiment and emotion analysis was conducted for the general tweets for each Higher Education Institution (HEI). To accomplish that, it used library NRCLEX [75]. This package incorporates approximately 27,000 words and draws from the National Research Council Canada (NRC) affect lexicon and NLTK library's WordNet synonym sets. Using this tool, we were able to analyze the sentiments and emotions expressed in a given text based on the words it contains.

Before applying the sentiment analysis algorithm, we conducted text preprocessing. Since the HEIs in our study are located in six different countries, with varying native languages, we translated tweets from non-English institutions into English for those institutions where English is not the native language. For institutions like EPFL, where the content was already in English, no translation was necessary. Additionally, as the text consisted of tweets, we performed several preprocessing steps, including the removal of @names, URLs, punctuation, emojis, stop words, italics, bold formatting, and lemmatization. This allowed us to tokenize the text and accurately analyze each word based on its emotional affection.

The returned emotional effects measured include fear, anger, anticipation, trust, surprise, positive, negative, sadness, disgust, and joy. These emotions and sentiments were then grouped based on their connotation: positive emotions (joy and trust), neutral emotions

(anticipation and surprise), and negative emotions (anger, disgust, fear, and sadness). Additionally, sentiments were classified separately, with the word "positive" corresponding exclusively to positive sentiment and the word "negative" to negative sentiment.

The analysis depicted in Figure 4.9 reveals that the last four HEIs in the study are distinguished by a predominantly positive emotional tone, with minimal counts of negative emotions, suggesting an overall positive sentiment in their communications. Interestingly, Leicester, despite not being among the top 5 institutions in terms of publication volume during the analyzed period, demonstrates a careful choice of words that reinforces its strong positive emotional expression. In contrast, Manchester and Trinity present a more balanced emotional profile, with moderate levels across all emotional categories, indicating a diverse emotional tone in their messaging. On the other hand, Harvard and Oxford display a broad range of emotions, with a noticeable tendency towards negative and neutral emotions. Oxford, in particular, has a strong bias towards negative emotions, which stands out in its overall emotional profile.

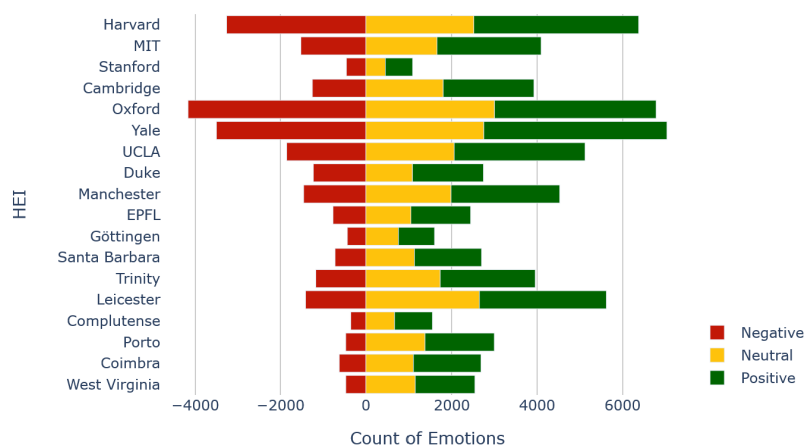


Figure 4.9: Positive, neutral, and negative emotion distribution of HEIs selected

In terms of sentiment analysis, HEIs generally exhibit a stronger positive sentiment compared to negative sentiment, as seen in Figure 4.10. Interestingly, the data suggests that as the percentage of positive sentiment increases, there is a corresponding rise in negative sentiment, indicating a possible correlation between the two. Additionally, institutions ranked below position 20 in the CWUR ranking tend to display a more consistent behavior, with around 80% showing this pattern.

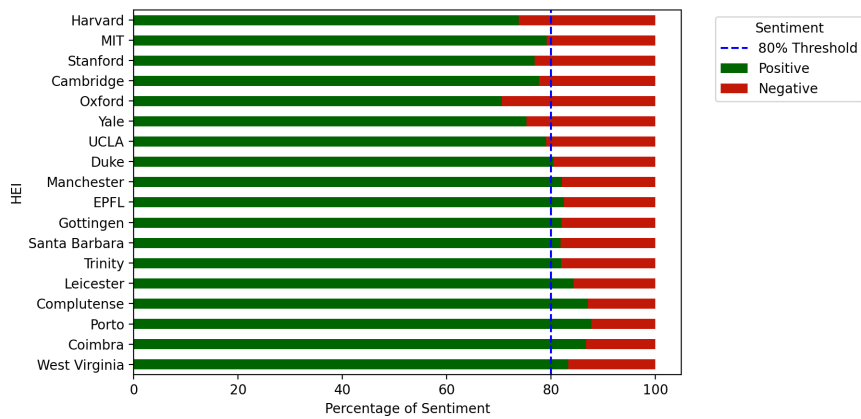


Figure 4.10: Sentiment distribution across HEIs of the period selected

This preliminary analysis has provided a foundational understanding of some emerging trends across various higher education institutions. While the findings offer valuable insights, they represent only a surface-level view of the complex dynamics with the patterns and characteristics that tend to exhibit. A more comprehensive and integrated analysis of these characteristics is necessary. The following Chapter 5, will delve deeper into these insights, providing an in-depth exploration of how these institutions group in terms of characteristics.

5. HEIs Grouping and Analysis

To facilitate the identification and generalization of patterns, we aim to group HEIs and discern the diverse characteristics employed by each university during the chosen period. This grouping would enable us to uncover different strategies and variations in behavior practiced by universities to manage their online presence and relate these to their position in the previously chosen ranking. For this analysis, we identified and created a set of features based on the previous analysis, and additional considerations, that will be used in k-means. This holistic approach allowed us to uncover various aspects such as engagement patterns and content characteristics.

We initially began by calculating the Pearson correlation coefficient, which revealed a strong correlation between the highest frequency and the standard deviation (0.97), and between the mean negative emotions and the mean negative sentiments (0.96). To reduce dimensionality, we excluded the standard deviation and mean of negative emotions from further analysis leaving us with the following 18 features:

- Highest posting rate
- Weekend and Weekday posting ratio
- Morning, Evening, and Night posting ratio
- Average number of publications per day
- Average number of photos, animated gifs, videos, URLs, and hashtag usage
- Average number of characters per post
- Positive and Neutral average emotion
- Positive, Neutral, and Negative average sentiment

To determine the optimal number of clusters for grouping the HEIs, we employed the k-means method with normalized variables, using statistics to better choose the centroids and the Hartigan-Wong algorithm [76]. The "elbow method" was applied to identify the optimal number of clusters by calculating the within-cluster Sum of Square Error (SSE) (also called "distortion") for different numbers of clusters. The point where the distortion

starts to decrease more slowly, forming an "elbow" in the plot, indicates the optimal number of clusters, as it balances the trade-off between minimizing distortion and avoiding overfitting. Thus, this method led us to the result of 4 clusters for our analysis.

Table 5.1: Formed clusters for the chosen HEIs in CWUR.

Higher Education Institution	Cluster
University of Porto	1
West Virginia University	1
Massachusetts Institute of Technology	2
Duke University	2
University of Göttingen	2
École Polytechnique Fédérale de Lausanne	2
University of California, Santa Barbara	2
Harvard University	3
Stanford University	3
University of Cambridge	3
University of Oxford	3
Yale University	3
University of California, Los Angeles	3
University of Manchester	3
Trinity College Dublin	4
University of Leicester	4
Complutense University of Madrid	4
University of Coimbra	4

5.1. Analysis of Bi-plot

A positioning map for each HEI is presented below (5.1) to visually illustrate the clustering outcomes, with colors denoting their assigned clusters. For the creation of this figure, Principal Component Analysis (PCA) was employed to reduce the 18-dimensional data points to two dimensions (the first two principal components).

The principal component analysis of the data reveals distinct influences on PC1 and PC2, accounting for 28.5% and 22.8% of the variance respectively. As it can be seen, through the size and the direction of the vectors present in the image, Principal Component 1 (PC1) is predominantly influenced by variables such as the average positive sentiment (positive_emotion_mean) and night ratio (night_ratio). In contrast, Principal Component 2 (PC2) is impacted by morning ratio (morning_ratio) at the lower end and variables like by the average publication per day (mean_publication_per_day) and weekend ratio (week-end_ratio) towards the top. Together, they represent more than half of the total variance, underscoring that sentiment (positive and negative) and communication strategies (like frequency and timing) are central themes in the dataset.

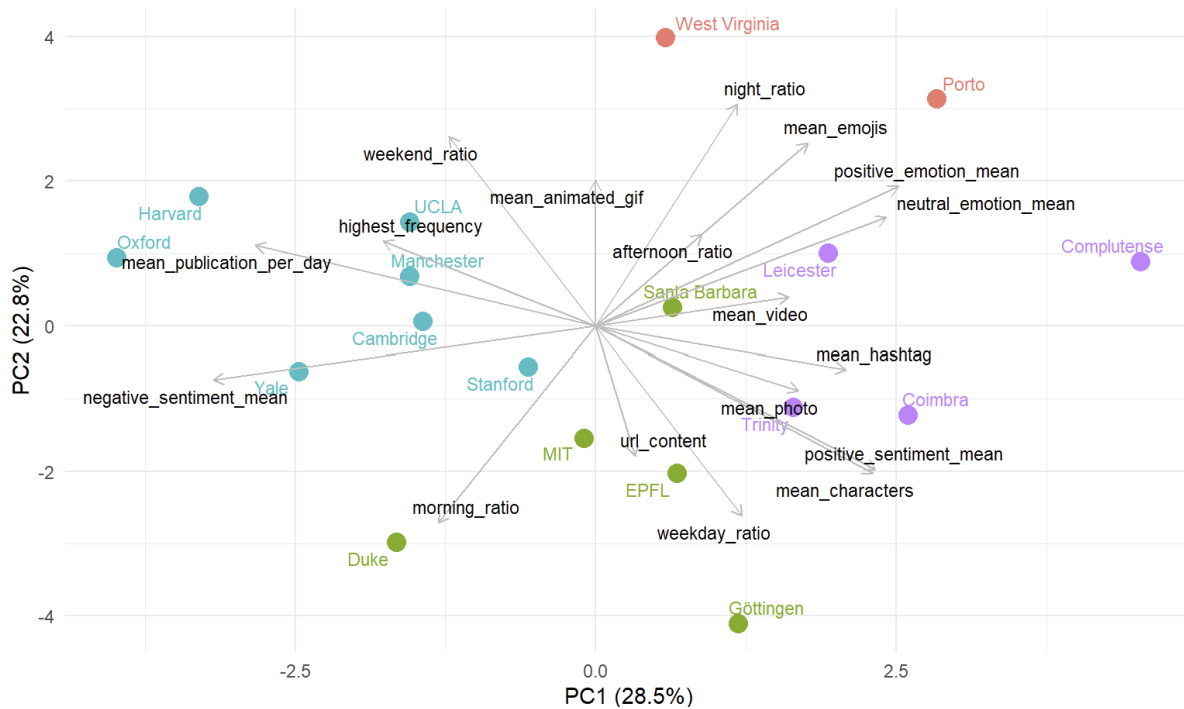


Figure 5.1: Biplot PCA for HEIs selected

Note that, although some of the features may seem correlated on the biplot shown in Figure 5.1, especially in the case of average number of characters per post (`mean_characters`) and the average number of positive sentiment (`positive_sentiment_mean`), this does not necessarily correspond to a direct correlation, as can be assessed by analyzing the Pearson correlation coefficient for these features (0.58 in the formerly mentioned case). This is a consequence of the dimensional reduction applied to create the biplot visualization, which condenses an 18-dimensional space into a 2-dimensional one, inevitably losing nearly half of the total variance that would be considered in the full 18-dimensional space, including important details captured by the Pearson correlation coefficient.

Furthermore, the positioning of West Virginia and Porto far from the origin (center of the plot) suggests that these two observations, which correspond to Cluster 1, are outliers in the context of the first two principal components. This is likely due to the distinct values in the variables that significantly contribute to PC1 and PC2, such as night ratio (`night_ratio`), average positive emotion (`positive_emotion_mean`), and emojis (`mean_emojis`), potentially indicating a unique behavioral pattern. In contrast, the positioning of Harvard, UCLA, Stanford, and other HEIs near the center of the biplot implies the formation of a core cluster with balanced characteristics. This cluster demonstrates a positive correlation across PC1 and PC2, suggesting that these universities generally share similar and moderate

values across the variables contributing to these components. Their central location reflects a shared profile that aligns well with the key factors driving the variance captured by PC1 and PC2.

5.2. Box-plot Analysis

To complement the analysis a visual representation of the distribution of selected features across different clusters, a boxplot is provided in Figure 5.2.

Cluster 1, encompassing institutions such as Porto and West Virginia, displays notably high values in variables like night ratio (`night_ratio`) and mean of positive emotion (`positive_emotion_mean`). These elevated results align with the identification of these institutions as outliers in the PCA biplot. The small number of institutions in this cluster is reflected in the narrow interquartile range (IQR), suggesting that despite their distinct characteristics, the members of this cluster are relatively homogeneous.

Cluster 3 exhibits balanced but slightly upraised values across various variables, with peaks in the average number of publications per day (`mean_publication_per_day`), highest frequency (`highest_frequency`), and a notable lean towards average negative sentiment (`negative_sentiment_mean`). This reaffirms that the institutions in this cluster maintain high levels of activity and a significant presence of negative sentiment.

Cluster 2, similar to Cluster 3, shows moderate values, reflecting its position near the center of the biplot. The variability in this cluster indicates that its institutions are more likely to engage in morning activities (`morning_ratio`) and make use of URL content (`mean_url`).

Finally, Cluster 4 maintains a consistent and positive profile, with strong weekday engagement and a balanced emotional tone leaning towards positivity. Within this cluster, Complutense stands out due to its relatively high dispersion across several variables. This suggests that, while sharing the cluster's general positive engagement profile, Complutense also exhibits greater variability in behavior. This could be due to diverse engagement strategies, a broader emotional spectrum, or less consistent patterns over time. Although this variability might slightly affect its positioning in the PCA biplot, it remains well-aligned with the overall characteristics of the members in Cluster 4.

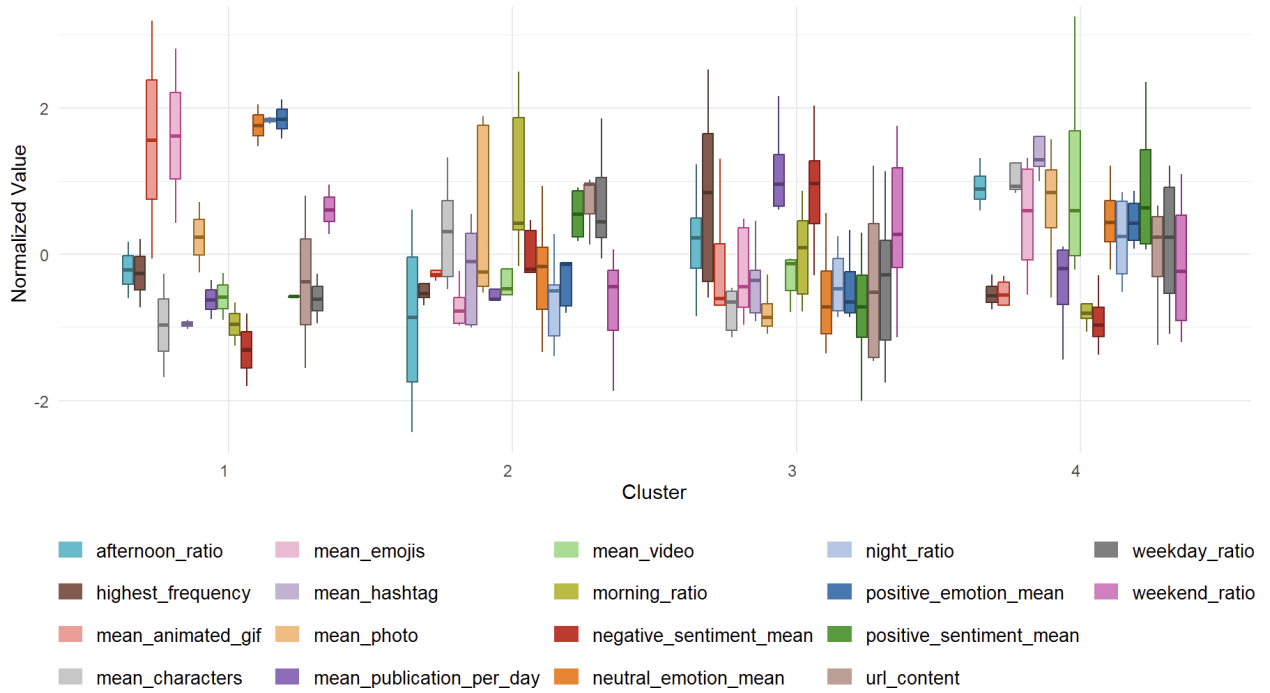


Figure 5.2: Distribution of all the features grouped by cluster.

5.3. Ranking and Posting Strategies

The correlation between the ranking of the HEIs and their behavior with features and clusters reveals distinct behaviors across high, moderate, and low rankings based on the selected batch. The geometric mean of rankings within each cluster further highlights these differences, with lower means indicating high-ranked institutions while higher means lower-ranked ones.

As analyzed in Sections 4 and 5:

- **High-ranking institutions** (Cluster 3) - high levels of content creation and engagement, with frequent publications and strategic use of sentiment to enhance their visibility and impact.
- **Moderate-ranking institutions** (Cluster 2 and 4) - balance a variety of features and strategies, reflecting diversity in their engagement approaches. Cluster 2, which includes institutions like MIT and Duke, is positioned closer to the high-ranking category due to their consistent performance and strategic focus, particularly in specialized areas like morning activities and URL content utilization. These institutions maintain a strong presence, with some variation, indicative of their solid academic standing. In contrast, Cluster 4, which includes institutions like Complutense and Coimbra, is positioned toward the lower end of the moderate range. The institutions

in this cluster often adopt a broader range of engagement strategies, resulting in greater variability in their outcomes.

- **Low-ranking institutions** (Cluster 1) - demonstrate high variability and adopt experimental approaches, reflecting the influence of geographic and strategic diversity. These institutions are exploring a variety of communication strategies designed to foster engagement and encourage positive interactions.

Throughout the analyses conducted in Chapters 4 and 5, we have explored the diverse behavioral patterns of HEIs across different clusters. As a result, each cluster exhibits distinct characteristics in terms of engagement strategies and emotional connotations used in their communications. Given these characteristics, it was possible to group and analyze distinct behaviors associated with various positions within the ranking.

Moreover, our findings reveal clear engagement strategies and underscore the potential for using predictive analytics on social media content. In the next chapter, 6, we will explore predictive modeling of post contents. This involves initially categorizing the types of content, a critical step for developing a strategy that can be learned computationally. We aim to uncover and utilize the patterns observed in the selected batch of HEIs during the study period, allowing us to predict future posting behaviors and enhance engagement strategies.

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. While previous analysis, in Chapter 4 and 5, has provided a clearer understanding of communication strategies and their correlation with institutional rankings, 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, with the purpose of classifying their publishing strategies but also identifying and predicting emerging trends and patterns. This categorization is grounded in the methodologies and findings detailed in research conducted by Oliveria et al. [77, 78], and Coelho et al. [79], 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.

BERTopic involves five key steps: embedding, dimensionality reduction, clustering, tokenization, and weighting, with optional representation tuning. Despite the modularity of BERTopic, the default steps were mostly retained for this analysis. Given some of the HEIs in the selected batch are located in different countries, to maintain consistency and accuracy in this analysis, the non-English publications were translated into English. Since the documents consisted of tweets, it was crucial to preserve the full context for accurate embeddings. Therefore, certain preprocessing steps were applied, including the removal of @names, URLs, hashtags, and formatting elements like italics and bold text.

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 [80] and Word2Vec [32], 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 [81], 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 transformed-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. However, 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 [82] 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, informed by the analysis previously conducted in Section 4:

- **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, engagement, 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.

6.1.1 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, inferring 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 6.1. UCLA presents a sharp peak in March, as noted in Section 4, 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.

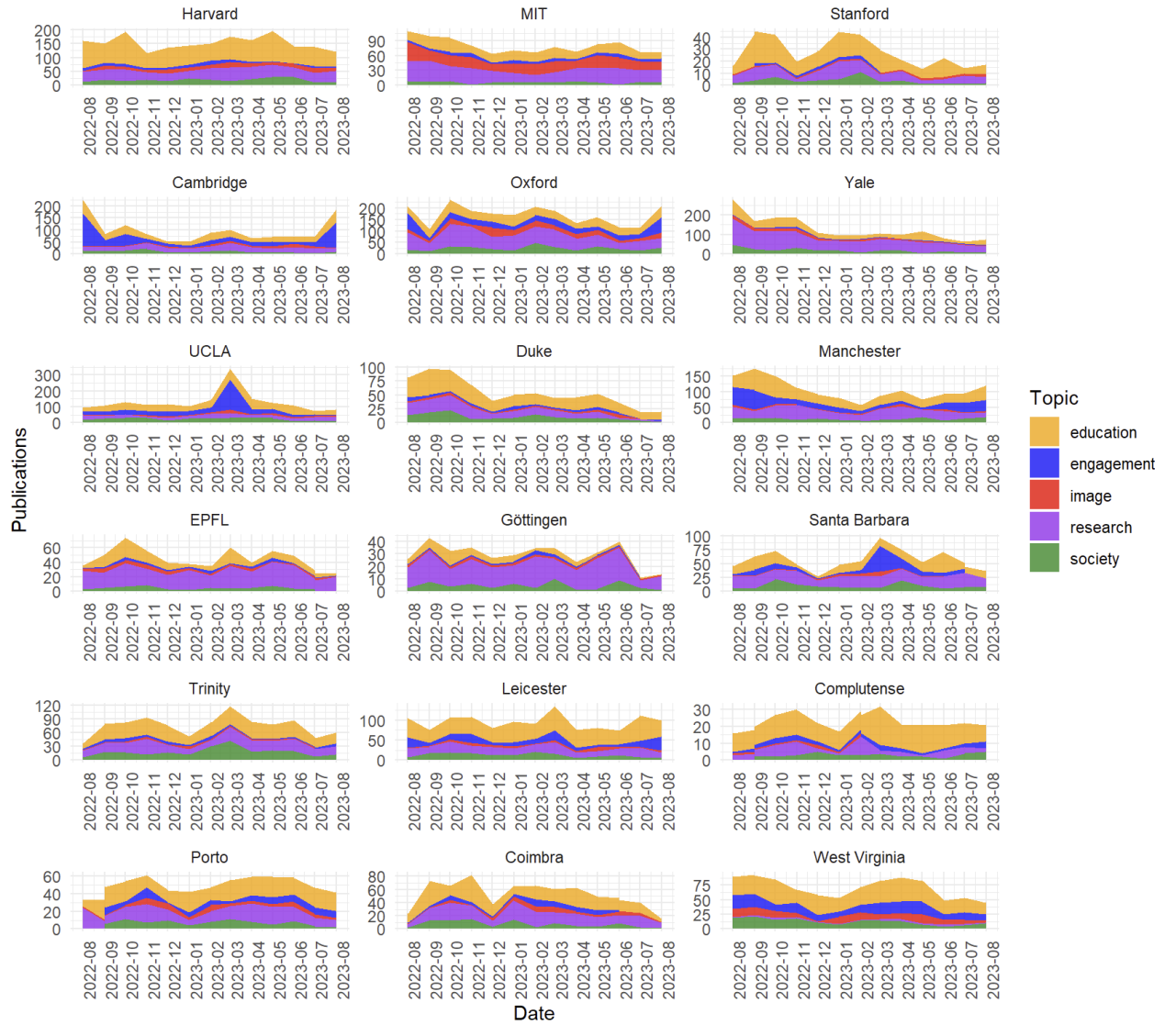


Figure 6.1: Temporal distribution of publication frequencies by topic

To delve deeper into the analysis, the PrefixSpan algorithm [83] 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 patterns results a period of 2 weeks was selected, as illustrated in Figure 6.2. 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 them 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.

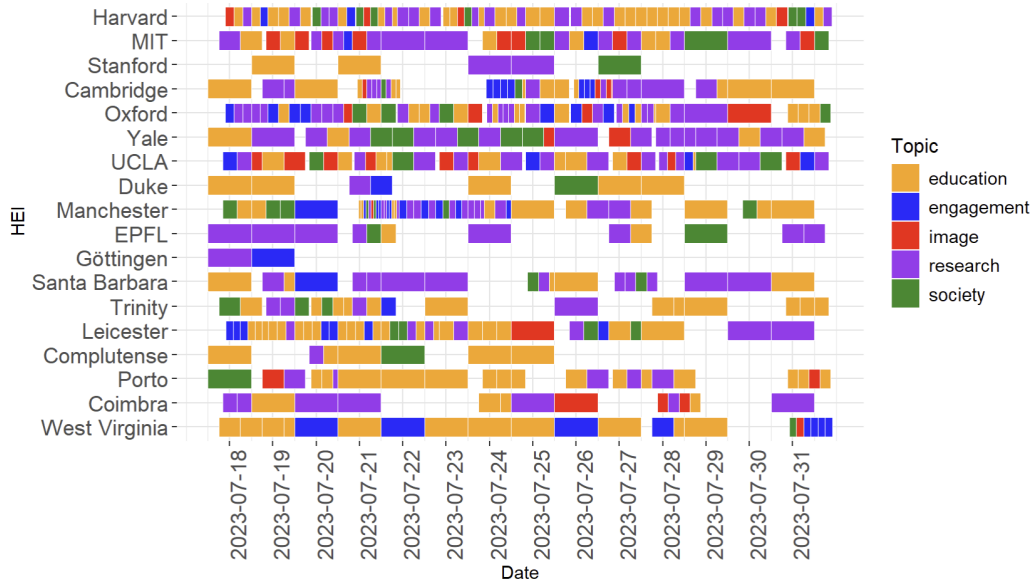


Figure 6.2: Topic distribution per day for each HEI during July 2023

6.2. Predictive Modeling

The rapid growth of social media has created both opportunities and challenges for Higher Education Institutions (HEIs) in extracting meaningful insights and predicting trends [84]. By analyzing the categorized themes of posts, particularly during the last five days of each month, it is possible to anticipate shifts in communication strategies. To accomplish this, a LSTM model [41] was employed for predictive analysis.

The text of the posts published underwent a second round of pre-processing, incorporating additional steps to ensure cleaner and more uniform data. This included the removal of URLs, @names, hashtags, italics, bold formatting, and emojis. Additionally, all text was converted to lowercase, and stop words, digits, punctuation, and any unusual characters (such as "—") 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 topic, 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 target period, ensuring the most relevant examples were selected. Those closer to the target period 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, temporal features such as the month and day of the week are included to account for seasonal publication patterns, along with the average cosine similarity score which determined the topic assignment for each post. 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, which means the entire dataset is passed through the model 10 times, allowing it to progressively learn and adjust its parameters with each pass.

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, and 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. An example of this process is illustrated in Figure 6.3.

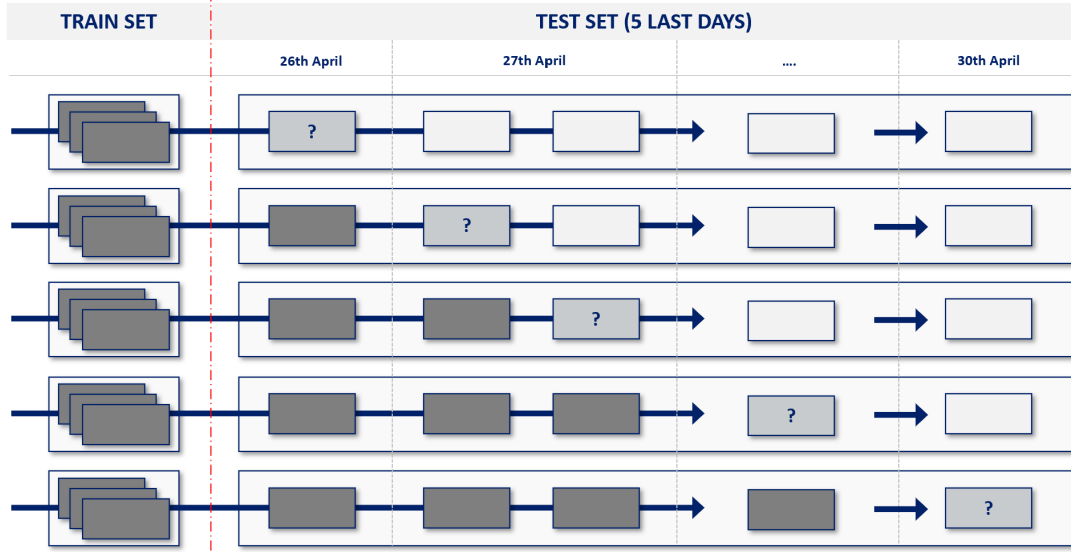


Figure 6.3: Prediction example for the month of April (30 days)

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

$$\text{Accuracy}_t = \frac{1}{t} \sum_{i=1}^t \mathbb{I}(\hat{y}_i = y_i) \quad (6.1)$$

where y_i represent the true label for the i -th post, and $\hat{y}_i$ represent the predicted label for the t -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:

$$\text{Final Accuracy} = \frac{1}{T} \sum_{t=1}^T \text{Accuracy}_t \quad (6.2)$$

6.2.1 End-of-Month Prediction Insights

The results of the LSTM are presented through Figures 6.4 and 6.5. Figure 6.4 illustrates the final accuracy achieved for each month of the academic year with the size of the points indicating the number of publications shared in the last five days of the respective month.

Among the 18 Higher Education Institutions (HEIs) analyzed, Duke presents the worst results falling below the 50% threshold, indicating significant challenges in maintaining consistent accuracy. In contrast, Stanford exhibits the widest range of accuracy, fluctuating dramatically from 0 in August 2023 to 100% in July 2023. This high volatility suggests that Stanford may be particularly sensitive to changes in publication volume or shifts in topics during this period.

When considering the volume of data published in the last five days of each month, Yale stands out, particularly in the first 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, and a tendency to reduce the number of publications over time, which could further impact the model's performance. Similarly, Stanford, EPFL, and Coimbra experience ups and downs in their accuracy trends, but with a general trend towards improvement, indicating potential adaptation or refinement in their models over time.

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%, Göttingen's performance remains strong and reliable despite the fluctuations.

Overall, MIT and Oxford show both consistent dot sizes and stable accuracy, implying that their publication processes are likely more uniform and effectively managed, minimizing any significant impact from end-of-month publications on accuracy.

Meanwhile, Figure 6.5 displays the distribution of accuracy across the five desired topics, showcasing the data for each month. The accuracy for each topic was calculated using the following expression:

$$\text{Accuracy Topic}(k) = \frac{C_k}{N_k} \quad (6.3)$$

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.

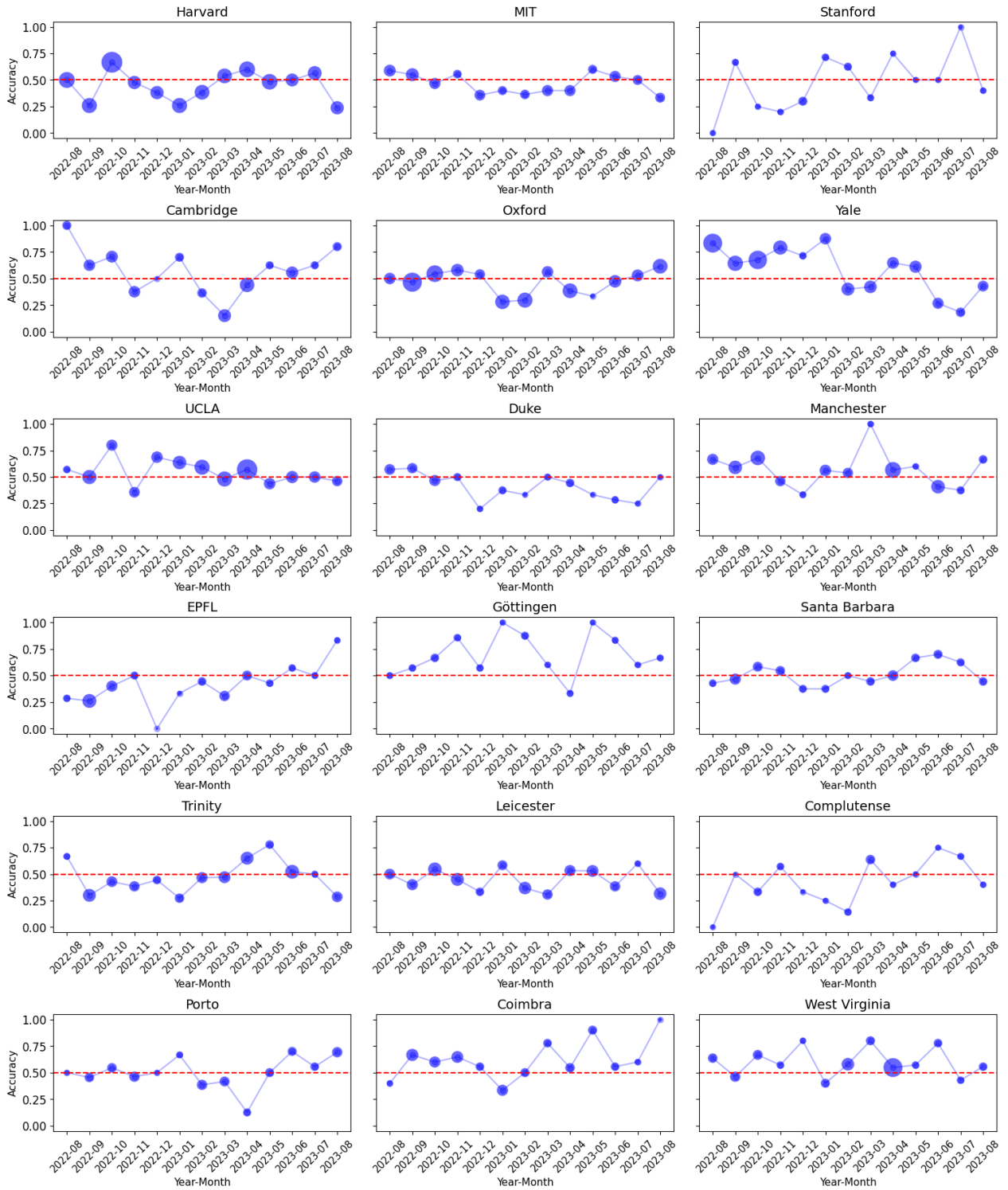


Figure 6.4: Prediction accuracy for posts: last 5 days of each month

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 boxplots 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. Given the distinct characteristics of each HEI as previously noted, improvements in model accuracy should be pursued independently, with tailored strategies that address the specific needs and features of each institution.

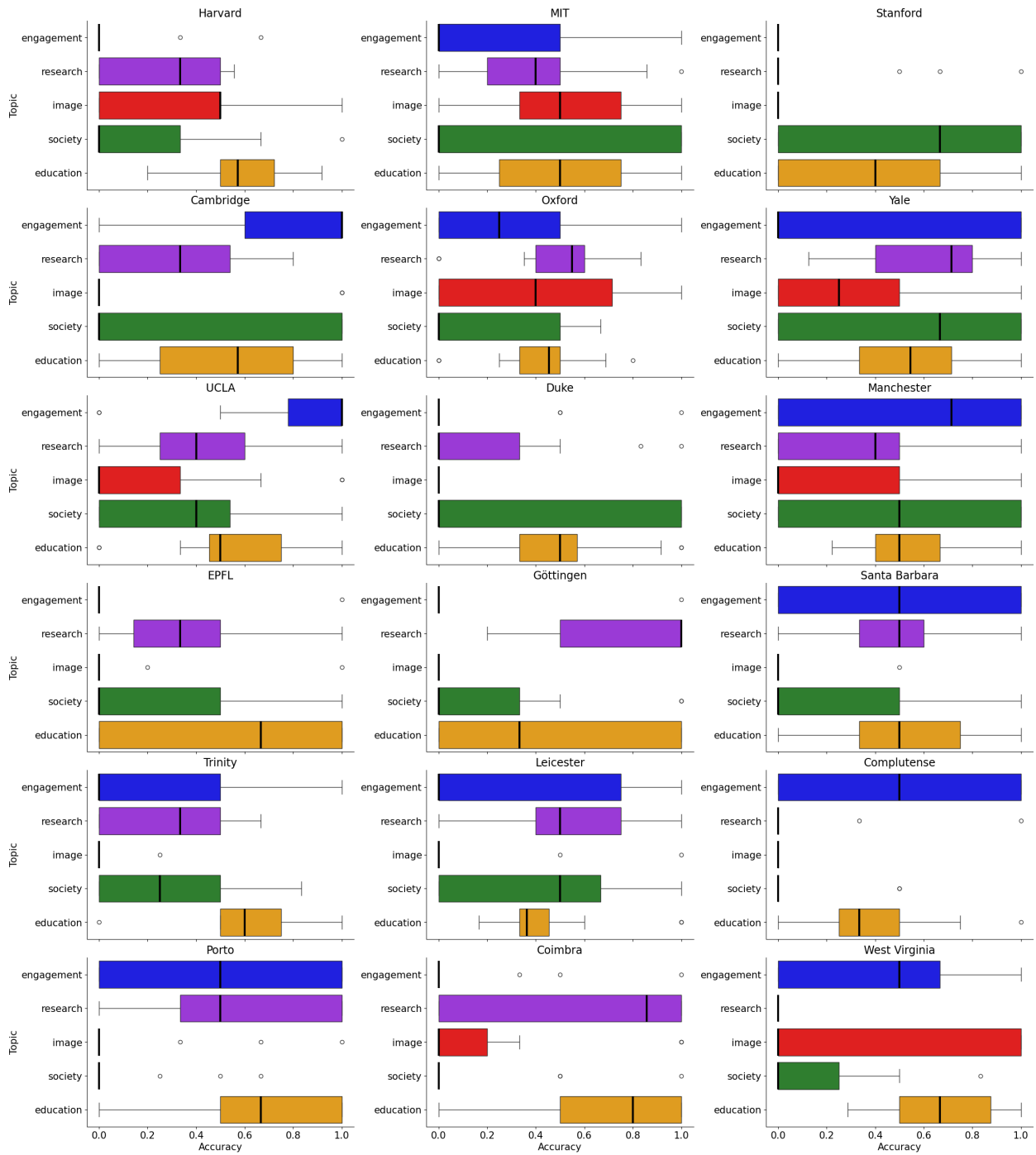


Figure 6.5: Topic-wise accuracy distribution across each HEI

7. Interpretation and Evaluation of Results

In this research, we successfully categorized the 18 HEIs into four distinct clusters, which we labeled 1 to 4. These clusters gathered HEIs with similar posting patterns, which we aim to analyze. These clusters were then assigned to a ranking layer, indicating their constituent universities' overall results in the chosen consolidated ranking system. We now aim to find clear evidence of a correlation between the layer of each cluster and the observed posting patterns, indicating that posting strategies may affect the success of universities or vice-versa.

7.1. Posting Strategies

The low-ranking layer was assigned to Cluster 1, composed of the universities of Porto and West Virginia. As discussed, the members in this cluster adopt more experimental strategies for their publications. This can be seen, for example, in the more unusual and diverse use of emojis, which accentuate the message transmitted to the public in Figure 4.7. One would also expect that institutions in lower-ranking positions would have a stronger sense of identity, as higher academic prestige is often accompanied by a globalization of each university's ideals and traditions. This was observed in our analysis, with institutions like Porto having a significant increase in activity around the traditional celebrations of the "Queima das Fitas", which attracts past, present, and future generations of members of the University of Porto's academic community.

The moderate-ranking level is the only one that has been assigned to multiple clusters, namely 2 and 4. These clusters belong in the intermediate region between the low and high-ranking levels. Due to this, the observed behaviors vary more distinctly based on whether the institutions are at the top or bottom end of the layer. For example, Cluster 2 is closer to the top end of the spectrum, including prestigious institutions such as Duke and MIT (top 20), but its average position is lowered by others such as EPFL and Santa Barbara (top 100). Nevertheless, institutions in this cluster share significant similarities, such as a more consistent posting schedule around the middle of the day, as seen in high-ranking HEIs like MIT. Universities in this cluster also have a more pronounced use of URLs, which often lead to the corresponding institutional website, allowing for more detailed dissemination of information beyond Twitter's character limit. This type of behavior

is expected from research-focused institutions like MIT and EPFL despite the apparent gap in their ranking positions [85].

The situation is different for Cluster 4, including for example Complutense and Coimbra, hence being located nearer to the bottom end of the moderate-ranking level. This means that there are some noticeable similarities to the low-ranking layer's posting patterns, such as the broader range of engagement strategies and the consequent wider variability of their corresponding outcomes. However, unlike in Cluster 1, elements of Cluster 4 have a stronger weekly engagement and a more balanced emotional tone, as seen in Figure 5.1.

The high-ranking layer was assigned to Cluster 3, including institutions like Harvard and Cambridge. All except two of the HEIs in this cluster are in the top 10 of the consolidated ranking, meaning that they are all considered the standard in the higher education landscape. As a result, their public image is important, leading to the use of sentiment strategically to boost visibility and public impact. This is reflected in their stronger social media presence, with higher posting frequency and a larger retweets and favorites count, as shown in Figure 4.5.

7.2. Topic Distribution

Having analyzed the overall behavior of each cluster and its positioning in the corresponding ranking layer, we then focus on the interpretation of the results from the social media post categorization. As mentioned earlier, each post was classified into one of five categories: Engagement, Research, Image, Society, or Education.

The low-ranking, Cluster 1 (Porto, West Virginia), showed a clear tendency towards Education, with smaller contributions from Engagement and Society topics. This was confirmed by the sequence pattern mining results from the PrefixSpan algorithm, which found Education to be the most frequent pattern in these institutions' daily publications. Nevertheless, we should add that the formerly discussed correlation could be present in the form of the secondary focus on Engagement and Society, which are associated with content focused on current students, with less attention given to research-focused topics in Cluster 1.

Cluster 2 (MIT, Duke, EPFL, Göttingen, and Santa Barbara), being part of the moderate-ranking layer but with some proximity to the top layer, shows a strong focus on Research. For example, Göttingen has 63% of its posts related to Research, a trend confirmed by

sequence pattern mining, which highlights the research orientation across these institutions. MIT also shows a pattern that emphasizes the institutional image, though this is less notorious.

On the other hand Cluster 4 (Trinity, Leicester, Complutense, and Coimbra), in the same layer but with more resemblance to the low-ranking institutions, exhibits a lower volume of publications compared to the previous cluster, making it harder to detect significant patterns in their social media presence. However, the publication trends in this cluster are relatively stable, apart from occasional peaks in concentrated periods of activity. Some members of the cluster, particularly Trinity, indicate that there is still a considerable amount of research activity, although it remains secondary to the globally dominant trend of education-related posts. Overall, both Clusters 2 and 4 show a greater tendency for research-focused content when compared to the remaining clusters.

Finally, Cluster 3 (Harvard, Cambridge, Stanford, Oxford, Yale, UCLA, Manchester), which was assigned to the high-ranking layer, is composed of HEIs with a significant social media presence, thus leading to a more diverse partition between the different topics. In addition to the prevalent Education topic, Engagement plays a prominent role in these HEIs' publication strategies, particularly in the university pair of Cambridge and Oxford, although its significance can fluctuate over time instead of being homogeneously present throughout the year. This cluster also experiences a peak in the first three months of 2023, followed by a gradual decline. Notably, Harvard follows complex patterns in its social media presence by combining Education and Research in multiple sequences. This approach demonstrates a strategic effort to engage its audience with educational content over several days, alongside its research activities, suggesting how its academic and investigative endeavors are interconnected.

7.3. Publication Prediction Analysis

Our analysis reveals different patterns in the accuracy of topic predictions across the arranged clusters. Naturally, this prediction depends on the number of posts available to train the model and the predictability of each institution's behavior, such as their adherence to defined strategies, which may not always be observed. In this way, the accuracy metric corresponds to a robust metric for testing the hypothesis that each HEI employs distinct strategies implemented in their social media presence, whether intended or unconscious.

Cluster 1 contains the smallest set of data analyzed, which leads to consistently moderate accuracy over time, suggesting that the models are stable with accurate results. A narrower focus on specific topics, such as Education and Engagement, is evident. Even though Engagement has a broad IQR, its median indicates that most predictions do not exceed 50% accuracy, which points to a balanced but not consistently high performance. Education has a smaller IQR, with all institutions maintaining general accuracy levels between 0.5 and 0.9, which reflects the overall stable performance in this topic. In contrast, Society despite not being the topic with the lowest percentage of entries has a lower accuracy level, marked by outliers in its inconsistent predictions, suggesting that the model could not detect a clear pattern. Altogether, the institutions in Cluster 1 provide significantly positive results for two out of five topics, while the quality considerably drops for the remaining ones. This reason is attributed to these topics having fewer entries in the dataset, which diminishes the model's ability to learn effectively.

Moving on to the moderate-ranking clusters, we see that Cluster 2 has small to medium dot sizes, which reflects its consistent number of posts throughout the academic year. Although most HEIs in this cluster show stable accuracy with only minor deviations, they generally fall below a 50% accuracy threshold for the selected period. Göttingen stands out with the best predictive results, except for a negative outcome in April. The result was significantly influenced by the predominant Research topic, which constituted 63% of the data with a IQR between 0.5 and 1 and a median accuracy of 100%. Thus, making it an outlier compared to the general trend. In Cluster 4, Coimbra exhibits a similar pattern with a more even distribution of data across topics, achieving remarkable results specifically in Research and Education. However, despite these successes, the overall cluster performance does not reach an average of 50% accuracy. This behavior reflects a mixed performance across different topics, with only some achieving positive accuracy results for the moderate-ranking layer.

Finally, when analyzing the high-ranking Cluster 3, we find high variability in prediction accuracy, reflected by the large number of posts and the intricate range of topics handled by these institutions. In the case of Stanford, we see that it shows some improvement over time, suggesting that the model progresses positively as it moves through the year. However, with the increase in the data volume desire to predict, the model finds it challenging to consistently maintain accuracy. Despite these challenges and being more difficult to predict as the low-ranking layer, the model still achieves an overall positive prediction for the period under examination, with the average result above the baseline. Moreover, HEIs

in Cluster 3 tend to outperform those in the moderate-ranking Clusters, often achieving more favorable results, as illustrated in Figure 6.5, indicating a global trend of success in this task. Society reveals a significant oscillation in results as evidenced by a wide IQR. This variability suggests that while some institutions may occasionally perform well, they generally struggle to achieve consistently positive outcomes. The negative median result underscores the challenges and the complexity of the topic, not being able to encounter a pattern.

8. Conclusion

In the conclusion of this study, we explored the relationship between university rankings and social media strategies among higher education institutions. We analyzed how these rankings correlate with dynamic social media engagement and its impact on engagement, recruitment, and funding. Our findings underscore how HEIs are integrating digital outreach with their academic and operational goals, emphasizing the use of modern communication strategies to boost institutional performance.

This chapter contains a general summary of the research carried out, addressing the research questions elaborated in Chapter 1 with a brief explanation of the methodology elaborated and conclusions obtained.

RQ1. Which ranking provides the most consensus in institutional positioning compared to the other major ranking systems?

In Chapter 3, we analyzed five major ranking systems: THE, QS, CWUR, ARWU, and WEB to determine which best represents the general positioning of the institutions. The similarity measures chosen correspond to the overlap size, Spearman's footrule, and M measure. The results obtained for the top 10, 100, and 200 indicated that ARWU had the best result based on the three measures. However since the ARWU does not individually classify all institutions, which significantly influenced the results, CWUR was chosen as the more suitable ranking system to pursue.

RQ2. How do different online strategies and behaviors of HEIs correlate with their positioning in academic rankings?

In Chapter 4, we analyzed patterns such as publication periods and social media engagement metrics (retweets and favorites) of HEIs. To further explore these patterns those institutions are grouped into four different clusters in Chapter 5 using the k-means algorithm. Then each cluster was assigned to one of three ranking layers according to the HEI's position. We observed that institutions in different layers display distinct preferences, such as the choice of morning activities by moderate-ranking institutions and the adoption of diverse strategies by those in the high and low tiers.

RQ3. What patterns can be identified in the content of the publications across different higher education institutions?

To get a more thorough insight into these strategies, the posts were categorized into five predefined topics - Engagement, Research, Image, Society, and Education - in Chapter 6.1. The categorization consisted of attributing one of the categories mentioned to each post during the period selected to conduct the research with the aid of BERTopic. Education stands out as a dominant topic in the overall 18 HEIs with Harvard presenting more recurrent patterns with Education->Education and Education->Research. These patterns underscore the emphasis on educational themes in their public communications, reflecting their priorities and possibly influencing their engagement with their audiences.

RQ4. How accurately can their activity be predicted using a sequential training approach?

After categorizing the posts into the five predefined topics, the LSTM model was employed to predict the last 5 days of each month based on the data from earlier in the month. As discussed in Chapter 6.2, the model can effectively predict HEI social media activity for some institutions, such as UCLA and Coimbra, according to the analysis presented. However, the accuracy of the predictions varied across different HEIs, influenced by factors such as the consistency and volume of available data. Given the distinct characteristics of each HEI as previously noted, improvements in model accuracy should be pursued independently, with tailored strategies that address the specific needs and features of each institution.

RQ5. Is there a significant correlation between social media publishing strategies and university rankings?

As depicted in Chapter 7, our research explored the relationship between social media strategies employed by HEIs and their placement in global rankings. We have found that low-ranking institutions often adopt experimental social media strategies presence, focusing on local community engagement and educational content. Despite the experimental nature of their approaches, these institutions exhibit a strong inclination towards education-related posts, aligning with their focus on internal community and cultural events. While distinct, the variability in their strategies often leads to the best results, emphasizing the effectiveness of tailored, community-focused content.

Institutions with moderate-ranking placement may adopt elements from both high and low-ranking strategies. They maintain a consistent posting schedule around the middle of the day and pronounced use of URLs, typically leading to the corresponding institutional

online platform, thus allowing for the dissemination of larger amounts of information in limited-character tweets. The focus tends to be on research, particularly notorious for those closer to the top of this layer and that therefore typically distribute their attention away from the education aspect, as seen in the examples of MIT and Duke. Despite varied emphases across this group, the predictive accuracy of their social media impact generally remains under 50%.

Higher-ranking layer institutions tend to have more sophisticated and consistent social media strategies. By using these online platforms to reinforce their typically globally recognized brand and public image, engaging with their far-spreading audiences, and highlighting their often world-leading academic achievements and research breakthroughs, as seen by their significantly high volume of retweets and favorites, these universities can boost their reputation and thus continue to attract top students and researchers, which is vital to their high placement in the rankings.

This analysis suggests a pattern where social media strategies are aligned with institutional rankings, highlighting tailored approaches that resonate with their respective audiences. It is important to note that these conclusions are based on examining 18 selected HEIs, which may not provide a complete picture of the broad landscape of institutional social media accounts. This suggests that the findings may not apply to all institutions, indicating a potential direction for future research based on this work.

8.1. Contributions

In addressing these research questions, this study offered the following contributions:

- A comprehensive analysis of five major institutional ranking systems to determine which most accurately reflects the general positions of the HEIs.
- A grouping of HEIs that reflect their posting patterns in which concerns frequency, posting time and day, length of post, sentiment, and emotions used.
- An automatic classification, based on BERT, of the posts' content based on pre-determined categories.
- An analysis of the sequencing of the posting categories using LSTM and the creation of a methodology for the prediction of the category of the next post, for each HEI.

Part of the research conducted for this dissertation resulted in a paper that was accepted for presentation at the 2024 International Symposium on Foundations and Applications

of Big Data Analytics (FAB), held in conjunction with the 16th International Conference on Advances in Social Networks Analysis and Mining (ASONAM). The paper "Comparing Higher Education Rankings with Social Media Posting Strategies" is awaiting publication by Springer.

8.2. Limitations and Future Work

One of the limitations of this study is the reduced engagement metrics on our dataset, which restricted our ability to explore the publishing process from the perspective of responses to the publications. A deeper analysis could have enabled us to more precisely identify and differentiate the various types of feedback and associated sentiments present in the publications held. Furthermore, expanding our analysis to include image and video content in these posts could significantly enrich our understanding, offering a more holistic view of the data. At last, exploring the expectations and preferences of the stakeholders of higher education institutions as well as extending the analysis to include multiple academic years could also have enhanced the robustness of this study.

References

- [1] T. Rafique, M. U. Awan, M. Shafiq, and K. Mahmood, "Exploring the role of ranking systems towards university performance improvement: A focus group-based study," *Heliyon*, vol. 9, no. 10, 2023. [Cited on page 1.]
- [2] A.-P. Pavel, "Global university rankings-a comparative analysis," *Procedia economics and finance*, vol. 26, pp. 54–63, 2015.
- [3] E. Hazelkorn, "How rankings are reshaping higher education," 2013. [Cited on page 1.]
- [4] M. A. Peters, "Global university rankings: Metrics, performance, governance," *Educational Philosophy and Theory*, vol. 51, no. 1, pp. 5–13, 2019. [Cited on page 1.]
- [5] E. Constantinides and M. C. Zinck Stagno, "Potential of the social media as instruments of higher education marketing: A segmentation study," *Journal of marketing for higher education*, vol. 21, no. 1, pp. 7–24, 2011. [Cited on page 1.]
- [6] L. Oliveira and Á. Figueira, "Social media content analysis in the higher education sector: From content to strategy," *International Journal of Web Portals (IJWP)*, vol. 7, no. 2, pp. 16–32, 2015. [Cited on page 1.]
- [7] Á. Figueira, "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*, 2018, pp. 923–929. [Cited on page 1.]
- [8] A. Figueira and L. Nascimento, "An nlp approach to understand the top ranked higher education institutions' social media communication strategy," in *International Conference on Web Information Systems and Technologies*. Springer, 2022, pp. 166–185. [Cited on page 3.]
- [9] Á. Figueira, "Uncovering social media content strategies for worldwide top-ranked universities," *Procedia Computer Science*, vol. 138, pp. 663–670, 2018. [Cited on page 3.]
- [10] D. Evans, *Social media marketing: the next generation of business engagement*. John Wiley & Sons, 2010. [Cited on page 5.]

- [11] L. Oliveira and Á. Figueira, "Improving the benchmarking of social media content strategies using clustering and kpi," *Procedia Computer Science*, vol. 121, pp. 826–834, 2017. [Cited on page 5.]
- [12] A. Peruta and A. B. Shields, "Social media in higher education: understanding how colleges and universities use facebook," *Journal of marketing for Higher Education*, vol. 27, no. 1, pp. 131–143, 2017. [Cited on page 5.]
- [13] S. Sengupta and A. Vaish, "A study on social networking platforms in higher education institutions," *Academy of Marketing Studies Journal*, vol. 26, no. 3, 2022. [Cited on page 5.]
- [14] C. L. Santos, P. Rita, and J. Guerreiro, "Improving international attractiveness of higher education institutions based on text mining and sentiment analysis," *International Journal of Educational Management*, vol. 32, no. 3, pp. 431–447, 2018. [Cited on page 6.]
- [15] P. Capriotti, A. Oliveira, and C. Carretón, "A model for assessing the active presence of institutions on social media: application to universities worldwide," *Journal of Marketing for Higher Education*, pp. 1–21, 2023. [Cited on page 6.]
- [16] I. Shehatta and K. Mahmood, "Correlation among top 100 universities in the major six global rankings: policy implications," *Scientometrics*, vol. 109, pp. 1231–1254, 2016. [Cited on page 6.]
- [17] M. P. Çakır, C. Acartürk, O. Alaşehir, and C. Çilingir, "A comparative analysis of global and national university ranking systems," *Scientometrics*, vol. 103, pp. 813–848, 2015. [Cited on page 6.]
- [18] I. Aguillo, J. Bar-Ilan, M. Levene, and J. Ortega, "Comparing university rankings," *Scientometrics*, vol. 85, no. 1, pp. 243–256, 2010. [Cited on pages 6 and 15.]
- [19] A. Figueira and L. V. Nascimento, "Do top higher education institutions' social media communication differ depending on their rank?" *Technology*, vol. 2, no. 3, p. 2, 2022. [Cited on page 6.]
- [20] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010. [Cited on page 7.]
- [21] J. A. Hartigan, M. A. Wong *et al.*, "A k-means clustering algorithm," *Applied statistics*, vol. 28, no. 1, pp. 100–108, 1979. [Cited on page 7.]

- [22] S. KS, R. Darsan, and N. Sampathila, "Examination of the relationship between social media usage, anxiety, and depression: a cluster analysis," *Trends in Psychology*, pp. 1–14, 2023. [Cited on page 7.]
- [23] K. Crockett, D. Mclean, A. Latham, and N. Alnajran, "Cluster analysis of twitter data: A review of algorithms," in *Proceedings of the 9th International Conference on Agents and Artificial Intelligence*, vol. 2. Science and Technology Publications (SCITEPRESS)/Springer Books, 2017, pp. 239–249. [Cited on page 7.]
- [24] J. H. E. Guzman, R. A. Zuluaga-Ortiz, L. E. G. Donado, E. J. Delahoz-Dominguez, A. Marquez-Castillo, and M. Suarez-Sánchez, "Cluster analysis in higher education institutions' knowledge identification and production processes," *Procedia Computer Science*, vol. 203, pp. 570–574, 2022. [Cited on page 7.]
- [25] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003. [Cited on page 8.]
- [26] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in *Proceedings of the first workshop on social media analytics*, 2010, pp. 80–88. [Cited on page 9.]
- [27] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*, 2010, pp. 261–270. [Cited on page 9.]
- [28] I. F. Strydom, J. Grobler, and E. Vermeulen, "Investigating the use of topic modeling for social media market research: A south african case study," in *International Conference on Computational Science and Its Applications*. Springer, 2023, pp. 305–320. [Cited on page 9.]
- [29] V. Maphosa and M. Maphosa, "Artificial intelligence in higher education: a bibliometric analysis and topic modeling approach," *Applied Artificial Intelligence*, vol. 37, no. 1, p. 2261730, 2023. [Cited on page 9.]
- [30] C. D. P. Laureate, W. Buntine, and H. Linger, "A systematic review of the use of topic models for short text social media analysis," *Artificial Intelligence Review*, vol. 56, no. 12, pp. 14 223–14 255, 2023. [Cited on page 9.]
- [31] A. Figueira and L. Nascimento, "An nlp approach to understand the top ranked higher education institutions' social media communication strategy," in *Web Information*

- Systems and Technologies*, M. Marchiori, F. J. Domínguez Mayo, and J. Filipe, Eds. Cham: Springer Nature Switzerland, 2023, pp. 166–185. [Cited on page 9.]
- [32] T. Mikolov, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013. [Cited on pages 9 and 39.]
- [33] M. Grootendorst, “Bertopic: Neural topic modeling with a class-based tf-idf procedure,” *arXiv preprint arXiv:2203.05794*, 2022. [Cited on page 9.]
- [34] M. Mendonça and Á. Figueira, “Topic extraction: Bertopic’s insight into the 117th congress’s twitterverse,” in *Informatics*, vol. 11, no. 1. MDPI, 2024, p. 8. [Cited on page 10.]
- [35] M. de Groot, M. Aliannejadi, and M. R. Haas, “Experiments on generalizability of bertopic on multi-domain short text,” *arXiv preprint arXiv:2212.08459*, 2022. [Cited on page 10.]
- [36] R. Egger and J. Yu, “A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts,” *Frontiers in sociology*, vol. 7, p. 886498, 2022. [Cited on page 10.]
- [37] T. Fütterer, C. Fischer, A. Alekseeva, X. Chen, T. Tate, M. Warschauer, and P. Gerjets, “Chatgpt in education: global reactions to ai innovations,” *Scientific reports*, vol. 13, no. 1, p. 15310, 2023. [Cited on page 10.]
- [38] D.-N. Grigore and I. Pintilie, “Transformer-based topic modeling to measure the severity of eating disorder symptoms.” in *CLEF (Working Notes)*, 2023, pp. 684–692. [Cited on page 10.]
- [39] N. Schneider, S. Shouei, S. Ghantous, and E. Feldman, “Hate speech targets detection in parler using bert,” *arXiv preprint arXiv:2304.01179*, 2023. [Cited on page 10.]
- [40] D. Antypas, A. Ushio, J. Camacho-Collados, L. Neves, V. Silva, and F. Barberi, “Twitter topic classification,” *arXiv preprint arXiv:2209.09824*, 2022. [Cited on page 10.]
- [41] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Cited on pages 11 and 43.]
- [42] I. Lasri, A. Riadsolh, and M. Elbelkacemi, “Self-attention-based bi-lstm model for sentiment analysis on tweets about distance learning in higher education.” *Int. J. Emerg. Technol. Learn.*, vol. 18, no. 12, pp. 119–141, 2023. [Cited on page 11.]

- [43] R. Pandey and J. P. Singh, "Bert-lstm model for sarcasm detection in code-mixed social media post," *Journal of Intelligent Information Systems*, vol. 60, no. 1, pp. 235–254, 2023. [Cited on page 11.]
- [44] Q. Zhou, "Innovation of educational management paths in higher education based on lstm deep learning model," *Applied Mathematics and Nonlinear Sciences*, 2023. [Cited on page 11.]
- [45] Y. Zhang, Y. Xiao, J. Wu, and X. Lu, "Comprehensive world university ranking based on ranking aggregation," *Computational Statistics*, vol. 36, pp. 1139–1152, 2021. [Cited on page 12.]
- [46] "World University Rankings," Sep. 2023. [Online]. Available: <https://www.timeshighereducation.com/world-university-rankings/2024/world-ranking> [Cited on page 12.]
- [47] "QS World University Rankings 2024." [Online]. Available: <https://www.topuniversities.com/world-university-rankings> [Cited on page 12.]
- [48] "World University Rankings 2023 | Global 2000 List | CWUR." [Online]. Available: <https://cwur.org/2023.php> [Cited on page 12.]
- [49] "ShanghaiRanking's Academic Ranking of World Universities." [Online]. Available: <https://www.shanghairanking.com/rankings/arwu/2023> [Cited on page 12.]
- [50] "World | ranking web of Universities |: Webometrics ranks." [Online]. Available: <https://www.webometrics.info/en/world> [Cited on page 12.]
- [51] J. Bar-Ilan, M. Levene, and A. Lin, "Some measures for comparing citation databases," *Journal of Informetrics*, vol. 1, no. 1, pp. 26–34, 2007. [Cited on pages 13 and 15.]
- [52] P. Diaconis and R. L. Graham, "Spearman's footrule as a measure of disarray," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 39, no. 2, pp. 262–268, 1977. [Cited on page 14.]
- [53] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank aggregation methods for the web," in *Proceedings of the 10th international conference on World Wide Web*, 2001, pp. 613–622. [Cited on page 14.]

- [54] R. Fagin, R. Kumar, and D. Sivakumar, "Comparing top k lists," *SIAM Journal on discrete mathematics*, vol. 17, no. 1, pp. 134–160, 2003. [Cited on page 14.]
- [55] École Polytechnique Fédérale de Lausanne, "Epfl summer research program," 2023, accessed: 2024-09-07. [Online]. Available: <https://www.epfl.ch/schools/sv/education/summer-research-program/> [Cited on page 20.]
- [56] Harvard University, "Academic text calendar 2022-23," 2022, accessed: 2024-01-30. [Online]. Available: <https://www.gsd.harvard.edu/wp-content/uploads/2022/07/Academic-Text-Calendar-2022-23-072522.pdf> [Cited on page 21.]
- [57] Massachusetts Institute of Technology, "Key catalog and schedule dates 2022-2023," 2022, accessed: 2024-01-30. [Online]. Available: https://registrar.mit.edu/sites/default/files/2022-09/Key_catalog_and_schedule_dates_2022-2023.pdf
- [58] Stanford University, "Stanford academic calendar 2022-23," 2022, accessed: 2024-01-30. [Online]. Available: <https://studentservices.stanford.edu/stanford-academic-calendar-2022-23#autumn-quarter-2022-23>
- [59] University of Cambridge, "Term dates and calendars," 2022, accessed: 2024-01-30. [Online]. Available: <https://www.cambridge-news.co.uk/news/cambridge-news/cambridge-university-202223-term-dates-24773402>
- [60] Daily Info, Oxford, "Oxford university term dates," 2022, accessed: 2024-01-30. [Online]. Available: <https://www.dailyinfo.co.uk/oxford/guide/university-term-dates>
- [61] Yale University, "Yale college calendar 2022-2023," 2022, accessed: 2024-01-30. [Online]. Available: https://registrar.yale.edu/sites/default/files/files/2022-2023_Yale%20College%20Calendar%20with%20Pertinent%20Deadlines%20_%20Yale%20University.pdf
- [62] University of California, Los Angeles, "Ucla academic calendar 2022-23," 2022, accessed: 2024-01-30. [Online]. Available: <https://registrar.ucla.edu/archives/academic-calendar-archive/academic-calendar-2022-23>
- [63] Duke University, "Duke academic calendar 2022-2023," 2022, accessed: 2024-01-30. [Online]. Available: <https://registrar.duke.edu/2022-2023-academic-calendar/>
- [64] University of Manchester, "Academic calendar 2022-2023," 2022, accessed: 2024-01-30. [Online]. Available: https://studentnet.cs.manchester.ac.uk/ugt/2023/timetable/cts_calendar.pdf

- [65] EPFL, "Epfl information about dates and deadlines 2022-2023," 2022, accessed: 2024-01-30. [Online]. Available: <https://www.epfl.ch/campus/services/internal-trainings/language-centre/practical-information/information-about-dates-and-deadlines-registration-and-courses/>
- [66] University of Göttingen, "Academic calendar 2022-2023," 2022, accessed: 2024-01-30. [Online]. Available: <https://www.uni-goettingen.de/en/academic%2Bcalendar%2B%28including%2Bfuture%2Band%2Bpast%2Bsemesters%29/24440.html>
- [67] University of California, Santa Barbara, "Ucsb academic calendar 2022-2023," 2022, accessed: 2024-01-30. [Online]. Available: <https://registrar.sa.ucsb.edu/calendars/calendars-deadlines/academic-calendars/academic-calendar-for-2022-2023>
- [68] Trinity College Dublin, "Academic year structure 2022-23," 2022, accessed: 2024-01-30. [Online]. Available: <https://www.tcd.ie/calendar/academic-year-structure/2022-23/academic-year-structure.pdf>
- [69] University of Leicester, "Term and semester dates 2022-2023," 2022, accessed: 2024-01-30. [Online]. Available: <https://le.ac.uk/about/info/term-semester-dates/closure-days>
- [70] Complutense University of Madrid, "Complutense university of madrid academic calendar 2022-2023," 2022, accessed: 2024-01-30. [Online]. Available: <https://trabajosocial.ucm.es/file/calendario-ingles-2022-23?ver>
- [71] University of Porto, "Academic calendar 2022-2023," 2022, accessed: 2024-01-30. [Online]. Available: <https://www.up.pt/portal/en/study/academic-information/academic-calendar/>
- [72] University of Coimbra, "University of coimbra academic calendar 2022-2023," 2022, accessed: 2024-01-30. [Online]. Available: <https://www.uc.pt/en/academicos/regulamentos/calendario>
- [73] West Virginia University, "Wvu academic calendar 2022-2023," 2022, accessed: 2024-01-30. [Online]. Available: <https://registrar.wvu.edu/calendars> [Cited on page 21.]
- [74] "Counting Characters." [Online]. Available: <https://developer.twitter.com/en/docs/counting-characters> [Cited on page 28.]

- [75] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon," *Computational intelligence*, vol. 29, no. 3, pp. 436–465, 2013. [Cited on page 29.]
- [76] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979. [Cited on page 32.]
- [77] L. Oliveira and Á. Figueira, "Benchmarking analysis of social media strategies in the higher education sector," *Procedia Computer Science*, vol. 64, pp. 779–786, 2015. [Cited on page 38.]
- [78] L. Oliveira and Á. Figueira, "Social media content analysis in the higher education sector: From content to strategy," *International Journal of Web Portals (IJWP)*, vol. 7, no. 2, pp. 16–32, 2015. [Cited on page 38.]
- [79] T. Coelho and Á. Figueira, "Analysis of top-ranked hei publications' strategy on twitter," in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 2021, pp. 5875–5877. [Cited on page 38.]
- [80] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017, to appear. [Cited on page 39.]
- [81] N. Reimers and I. Gurevych, "Sentence-transformers: Pretrained models," 2023, accessed: 2024-06-02. [Online]. Available: https://www.sbert.net/docs/sentence_transformer/pretrained_models.html [Cited on page 39.]
- [82] M. Grootendorst, "Bertopic faq," 2023, accessed: 2024-06-02. [Online]. Available: <https://maartengr.github.io/BERTopic/faq.html#i-am-facing-memory-issues-help> [Cited on page 40.]
- [83] J. Han, J. Pei, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu, "Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth," in *proceedings of the 17th international conference on data engineering*. IEEE Piscataway, NJ, USA, 2001, pp. 215–224. [Cited on page 42.]
- [84] S. Gaftandzhieva, S. Hussain, S. Hilcenko, R. Doneva, and K. Boykova, "Data-driven decision making in higher education institutions: State-of-play," *International Journal*

of Advanced Computer Science and Applications, vol. 14, no. 6, pp. 397–405, 2023.

[Cited on page 43.]

- [85] M. Benner, “Are some technical universities better than others?” *Technical Universities: Past, present and future*, pp. 27–43, 2020. [Cited on page 51.]

Appendix

Appendix A: Supplementary Figures

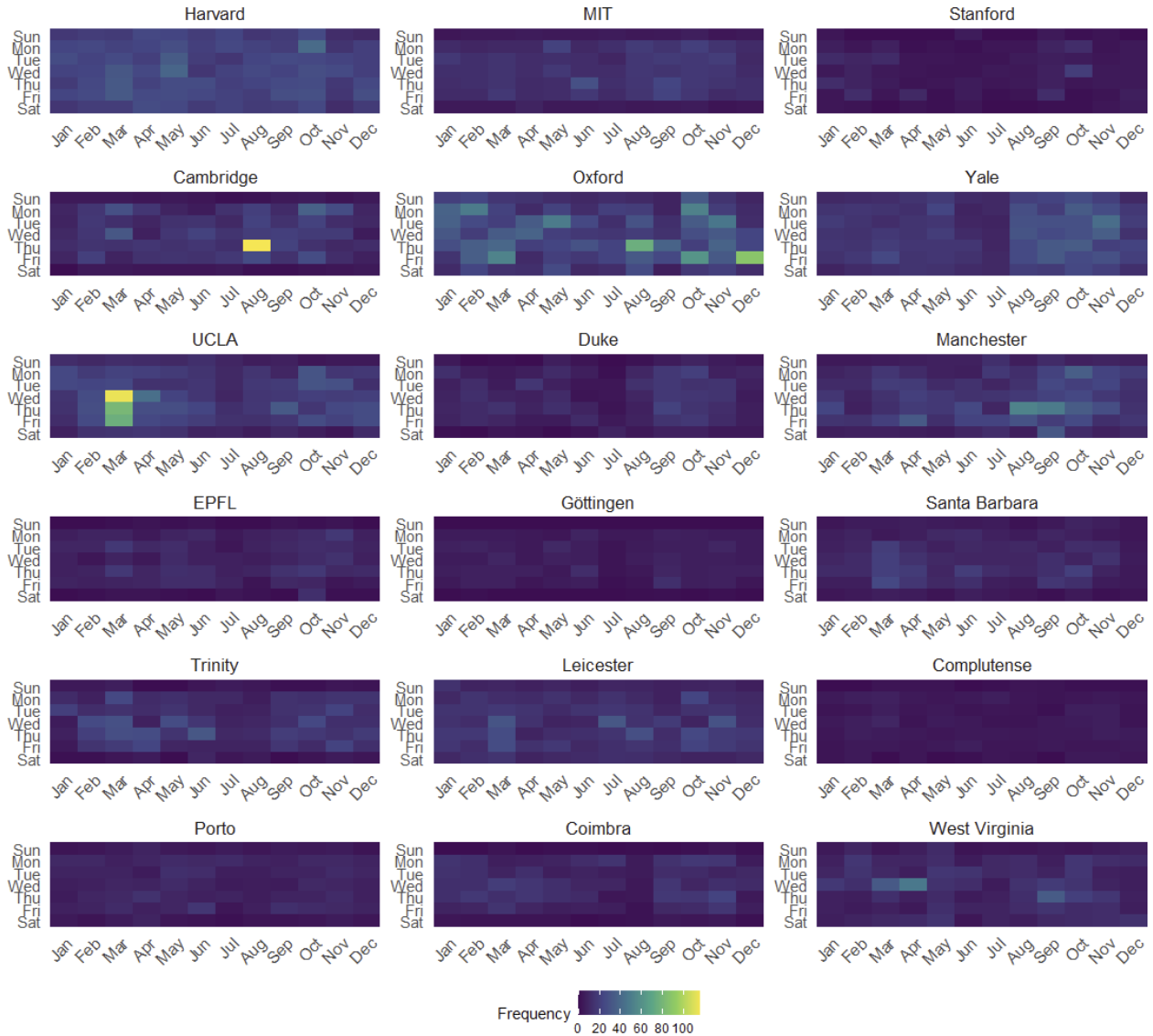


Figure 1: Month versus day of the week of HEIs