

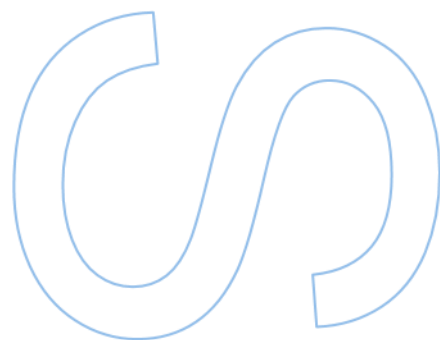
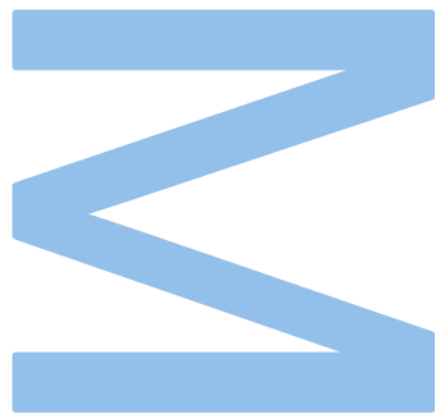
Analysis of Higher Education Institutions' Social Media Posting Using NLP Techniques

Lirielly Vitorugo

Master's Degree in Data Science
Department of Computer Science
2023

Supervisor

Prof. Dr. Álvaro Figueira, Assistant Professor, Faculty of Sciences



“ Life is not easy for any of us. But what of that? We must have perseverance and above all confidence in ourselves. We must believe that we are gifted for something and that this thing must be attained. ”

Marie Curie

Acknowledgements

I would like to thank all those who, directly or indirectly, contributed to the execution of this work.

Firstly, I would like to express my gratitude to Professor Álvaro Figueira for suggesting the dissertation topic, for his unwavering dedication and for his guidance throughout this endeavor. He has always been available to help me with my doubts, in the role of my mentor, he encouraged me to do my best in any circumstance. It is important to recognize that without his guidance none of this would have been possible.

I would also like to thank all my friends, especially Amanda Tavares and Arina Sanches, my partners throughout the entire master's degree journey. Without you girls, this path would have been much more challenging and boring.

To my entire family, especially my father, mother and brother, who have always loved me unconditionally, supported my decisions, celebrated my victories with me, and were by my side when needed.

To Daniela Augustin and her wonderful family, I extend my sincere gratitude for their love and support during the intense last month of this journey. She was by my side, assuring me every day that I would finish writing and submit my dissertation within the stipulated deadline. Thank you for believing in me.

Last but certainly not least, I extend my gratitude to my friend and mentor, Jodelson Sabino, for inspiring me to grow intellectually, including pursuing this master's degree. He never allowed me to give up and was always available to listen and guide me in the most difficult moments, whether academic, professional, or even personal.

Abstract

Social media platforms have emerged as vital tools for Higher Education Institutions (HEIs) to spread information and engage students. This study investigates the intricate interplay between institutional performance, as measured by rankings, and social media engagement strategies. It specifically examines how these strategies can be influenced by their respective ranking positions and how this can impact an institution's brand, thus attracting top professionals and students. To achieve this, we employ Natural Language Processing (NLP) techniques to analyse historical Twitter data from these institutions. Our study found no significant differences between well-known university rankings based on Kendall τ and RBO metrics. Subsequently, a clustering model was conducted to identify institutions with similar publishing strategies and collect valuable insights from segmentation. Our findings revealed that top-ranked institutions tend to exhibit more pronounced sentiments, engage in longer tweets, and maintain higher posting frequencies. Additionally, we performed sentiment, emotion, and topic analyses of tweet content for each HEI. This revealed similarities and common patterns in topic content across the institutions analyzed, providing a deeper understanding of their publishing strategies. Interestingly, some universities showed less prominent strategies that did not align with their ranking positions. Therefore, this study offers, not only a methodological path for this type of research but also insights into the role of social media in HEIs' marketing strategies and its impact on their global rankings.

Keywords - NLP, Higher Education Institutions, Social Media Communication, Twitter, Ranking Analysis, Publishing Strategies.

Resumo

As plataformas de redes sociais surgiram como ferramentas vitais para Instituições de Ensino Superior (IES) na divulgação de informações e no engajamento com os estudantes. Este estudo investiga a interação entre o desempenho institucional, medido por rankings, e as estratégias de envolvimento nas redes sociais. O estudo especificamente examina como essas estratégias podem ser influenciadas pelas respectivas posições nos rankings e como isso pode impactar a imagem de uma instituição, atraindo assim profissionais e estudantes de alto nível. Para alcançar este objetivo, foram utilizadas técnicas de Processamento de Linguagem Natural (PLN) para analisar dados históricos do Twitter dessas instituições. O estudo não encontrou diferenças significativas entre os mais conhecidos rankings universitários com base nas métricas de Kendall τ e RBO. Posteriormente, realizou-se um modelo de agrupamento para identificar instituições com estratégias de publicação semelhantes e obter informações valiosas com base na segmentação. Nossas descobertas revelaram que instituições melhores classificadas tendem a exibir sentimentos mais pronunciados, envolvem-se com tweets mais longos e mantêm suas frequências de publicação mais altas. Além disso, realizamos análises de sentimento, emoção e tópico do conteúdo dos tweets de cada IES. Os resultados revelaram semelhanças e padrões comuns no conteúdo do tópico entre as instituições analisadas, fornecendo uma compreensão mais profunda de suas estratégias de publicação. Curiosamente, algumas universidades exibiram estratégias menos proeminentes que não se alinhavam com suas posições de classificação. Portanto, este estudo oferece não apenas um caminho metodológico para esse tipo de pesquisa, mas também insights sobre o papel das redes sociais nas estratégias de marketing das IES e seu impacto em suas classificações globais.

Palavras-Chave - PLN, Instituições de Ensino Superior, Comunicação em Mídias Sociais, Twitter, Análise de Rankings, Estratégias de Publicação.

Contents

Acknowledgements	iii
Abstract	v
Resumo	v
Contents	vi
List of Figures	ix
List of Tables	xiii
1 Introduction	1
1.1 Context	1
1.2 Motivations	2
1.3 Objectives	2
1.4 Document's Structure and Organization	3
2 Literature Review	5
2.1 Social Analysis of Social Media Publications for HEIs	5
2.2 Predictive Modeling using HEIs Social Media Content	9
3 A System for Automatic Analysis of HEIs' Twitter Posts	13
3.1 Data Retrieval	13
3.1.1 Ranking Systems Analysis	13
3.1.2 Twitter Data Collection	16
3.2 Exploratory Data Analysis	19
3.2.1 HEIs Number of Followers	19
3.2.2 Data Distribution Analysis	20
3.2.3 Publication Period of the Day	24
3.2.4 Word Cloud Analysis	31
3.2.5 Sentiment Analysis	34
3.2.6 Emotion Analysis	38
3.2.7 Topic Analysis	42
4 Discussions and Results	47
4.1 Topic Modeling	47

4.2	Grouping the Strategies	49
4.2.1	The Feature Space Vector Model	49
4.2.2	Clustering the HEIs	50
4.2.3	Analysis of the Clustering Results	53
5	Conclusions	57
5.1	Synthesis	57
5.2	Contributions	57
5.2.1	Objectives	57
5.2.2	Preliminary Publications	59
5.3	Future Work	59
A	Sentiment Evolution	61
B	HEIs Topics Keywords Discrimination	67
	Bibliography	79

List of Figures

3.1	Data recovery period for each HEI	17
3.2	Number of tweets published per each HEI in all the collected period	18
3.3	Number of tweets published per each HEI in the common period	18
3.4	Number of followers, in millions	19
3.5	Retweet boxplot without outliers.	20
3.6	Like boxplot without outliers.	20
3.7	Daily tweet frequency for each HEI	23
3.8	Boxplots for the daily posting frequency for each HEI	24
3.9	Daily distribution posting frequencies for each HEI	25
3.10	Daily publication of HEIs crossed by the day of the week and time hour, first part	27
3.11	Daily publication of HEIs crossed by the day of the week and time hour, second part	28
3.12	Daily publication of HEIs crossed by the day of the week and month of the year in 2022, first part	29
3.13	Daily publication of HEIs crossed by the day of the week and month of the year in 2022, second part	30
3.14	Word cloud considering all the retrieved publications for each HEI	32
3.15	Word cloud considering the common period of retrieved publications for each HEI	35
3.16	Monthly evolution of sentiment for Harvard, MIT and Stanford posts	36
3.17	Daily cumulative sentiment frequency for all HEIs	38
3.18	Proportion of emotions and sentiments found in the overall tweets all the HEI, first part	40
3.19	Proportion of emotions and sentiments found in the overall tweets all the HEIs, second part	41
3.20	Federal University of Minas Gerais best number of topics for LDA based on coherence score	43
3.21	LDA topic model visualization for Federal University of Minas Gerais	44
4.1	Elbow method results	51
4.2	Positioning and clustering for all the HEIs in a 2D projection of the feature space	52
4.3	Distribution for all the features separated by cluster	53
4.4	Distribution of each variable in each cluster highlighting the major changes	55
A.1	Monthly evolution of sentiment for Boston University posts	61
A.2	Monthly evolution of sentiment for California Institute of Technology posts	61

A.3	Monthly evolution of sentiment for Columbia University posts	62
A.4	Monthly evolution of sentiment for Federal University of Minas Gerais posts	62
A.5	Monthly evolution of sentiment for Harvard University posts	62
A.6	Monthly evolution of sentiment for Massachusetts Institute of Technology posts	62
A.7	Monthly evolution of sentiment for Princeton University posts	63
A.8	Monthly evolution of sentiment for Stanford University posts	63
A.9	Monthly evolution of sentiment for University at Buffalo posts	63
A.10	Monthly evolution of sentiment for University of Cambridge posts	63
A.11	Monthly evolution of sentiment for University of Chicago posts	64
A.12	Monthly evolution of sentiment for University of Lisbon posts	64
A.13	Monthly evolution of sentiment for University of Oklahoma posts	64
A.14	Monthly evolution of sentiment for University of Oxford posts	64
A.15	Monthly evolution of sentiment for University of Pennsylvania posts	65
A.16	Monthly evolution of sentiment for University of Porto posts	65
B.1	Boston University topics keywords. Topic 0: University Event and Topic 1: Beginning of the Academic Year	67
B.2	California Institute of Technology topics keywords. Topic 0: Technology, Topic 1: Research and Development and Topic 2: Beginning of the Academic Year	68
B.3	Columbia University topics keywords. Topic 0: Academic Programs, Topic 1: Education, Topic 2: Community and Topic 3: University Event	69
B.4	Federal University of Minas Gerais topics keywords. Topic 0: University Event, Topic 1: Research and Development, Topic 2: Beginning of the Academic Year and Topic 3: Academic Programs	70
B.5	Harvard University topics keywords. Topic 0: Health Care and Medicine, Topic 1: Education, Topic 2: Beginning of the Academic Year and Topic 3: Research and Development	71
B.6	Massachusetts Institute of Technology topics keywords. Topic 0: Beginning of the Academic Year and Topic 1: Research and Development	71
B.7	Princeton University topics keywords. Topic 0: Education, Topic 1: Professional Development and Topic 2: Beginning of the Academic Year	72
B.8	Stanford University topics keywords. Topic 0: Beginning of the Academic Year and Topic 1: Research and Development	72
B.9	University of Cambridge topics keywords. Topic 0: Community, Topic 1: Technology, Topic 2: Research and Development, Topic 3: Collaboration and Accessibility and Topic 4: University Event	73
B.10	University of Chicago topics keywords. Topic 0: Health Care and Medicine, Topic 1: Education, Topic 2: Diversity and Inclusion, Topic 3: University Event and Topic 4: Research and Development	74
B.11	University of Oklahoma topics keywords. Topic 0: Beginning of the Academic Year, Topic 1: University Event and Topic 2: Community	75
B.12	University of Oxford topics keywords. Topic 0: Health Care and Medicine, Topic 1: Research and Development and Topic 2: University Event	76
B.13	University of Pennsylvania topics keywords. Topic 0: Sports, Topic 1: Research and Development, Topic 2: Beginning of the Academic Year and Topic 3: Education	77

B.14 University of Porto topics keywords. Topic 0: Education and Topic 1: Beginning of the Academic Year	77
--	----

List of Tables

3.1	The position of the HEIs on the four rankings chosen in 2022.	15
3.2	The position of the HEIs on the four rankings chosen in 2023.	16
3.3	Posting daily frequency average, maximum and standard deviation (in decreasing order).	21
3.4	Topics found by the Latent Dirichlet Allocation (LDA) model for each HEI. (A)	45
3.5	Topics found by the Latent Dirichlet Allocation (LDA) model for each HEI. (B)	45
4.1	Cluster assignment together with the rankings scores of 2022.	54

Chapter 1

Introduction

1.1 Context

It is a fact that the proliferation of rankings has provided individuals with a wealth of information to assist in their decision-making processes. This trend is particularly evident in the realm of Higher Education Institutions (HEIs), where university rankings have become both increasingly common and diverse. Arguably, the main purpose of these rankings is to quantify and evaluate success in a variety of areas or criteria.

The metrics used to evaluate HEIs are continually being refined and improved, as are the methodologies used to determine them with greater precision. Typically, these institutions are evaluated based on factors such as student success, research volume, funding and awards, internationalization, employment rates, and industry connections, among others. This comprehensive approach allows for a more nuanced understanding of an institution's performance and strengths.

Presently, there are numerous prominent indices for Higher Education Institutions (HEIs). Among the most recognised and widely used are the Centre for World University Rankings - CWUR, QS World University Rankings - QS, CWTS Leiden Ranking - Leiden, Academic Ranking of World Universities - ARWU (also known as the Shanghai ranking), University Ranking by Academic Academic Performance - URAP and U.S. News & World Report - USNews.

In fact, research has indicated a robust correlation between these indices over time [1]. Despite small variations in the indices, which can be attributed to the unique evaluation criteria employed by each, the overall picture of an institution's performance and position

remains consistent across different indices. This consistency underscores the reliability of these indices as measures of institutional performance in the global academic community.

Numerous studies, including [2], [3], and [4], have explored the comparison of university rankings and the inherent challenges associated with such comparisons. However, our work takes a different approach by comparing Higher Education Institution (HEI) rankings with their respective Twitter posting strategies. The main objective of this analysis is to examine the degree of variation in the external communication strategies employed by different HEIs, rather than to focus on the ranking itself. This new approach provides a unique perspective on the interaction between institutional performance, measured by rankings, and social media engagement strategies.

1.2 Motivations

The motivation behind this analysis is rooted in the fundamental importance of external communication in recruiting prospective students, eminent scholars, and financial resources, as recognized in previous academic work [5]. Given the increasing prominence of the image projected by Higher Education Institutions (HEIs), it becomes imperative to subject their performance and strategies to meticulous scrutiny within popular academic and non-academic platforms, namely Twitter, Facebook, Instagram, TikTok, and LinkedIn.

1.3 Objectives

The objective of this study is therefore to discern whether rankings reflect differences in how HEIs communicate their messages. This involves a comprehensive review of their social media strategies, especially on Twitter, which has become an indispensable tool in academia for information dissemination and engagement. By understanding the correlation between an institution's ranking and its communication strategy, we can gain valuable insights into the effectiveness of different approaches and their impact on an institution's perceived value and reputation.

To accomplish the principal objective, the study was subdivided into several research goals, as delineated below:

1. Identify a ranking that can be consistent with the existing ones.

2. Identify publication characteristics of HEIs.
3. Establish categories for the classification of HEIs according to their strategies.
4. Compare these categories with the ranking positions.

1.4 Document's Structure and Organization

This dissertation is organized as follows:

- Starting with **Chapter 1** where the introduction of the current study was present.
- **Chapter 2** provides an overview of the Social Analysis of Social Media Publications for Higher Education Institutions (HEIs) and discusses the Predictive Modeling using HEIs' Social Media Content.
- **Chapter 3** conducts an analysis of the collected data, including exploratory data analysis, sentiment analysis, emotion analysis, and topic analysis.
- **Chapter 4** presents the discussions and results segregated by topic modeling. It also uses a vector space model and clustering to compare all HEIs and analyze the results.
- Finally, **Chapter 5** summarises the research process and presents our final conclusions.

This structure ensures a logical progression of ideas and findings, facilitating a comprehensive understanding of the research conducted.

Chapter 2

Literature Review

In this chapter, we present a comprehensive literature review that focuses on the analysis of social media publications within Higher Education Institutions (HEIs). This review not only investigates the current state of the art in this domain but also explores the use of predictive modeling regarding HEIs' social media publishing content.

2.1 Social Analysis of Social Media Publications for HEIs

Previous studies have explored the analysis of social media publications in HEIs, including works such as Oliveira and Álvaro Figueira [6] that explore the challenges organizations face when trying to measure the return on investment (ROI) and the efficiency of their social media efforts, especially when these efforts lack a clear strategic design. It focuses on the higher education sector and uses clustering techniques and social media (Facebook) key performance indicators (KPIs) to analyze various content strategies. The research highlights the significance of content strategies centered on an organization's identity and image for maximizing audience engagement. These insights offer valuable guidance to organizations, particularly in higher education, aiming to enhance their social media strategies and achieve better ROI.

Peruta and Shields [7] examines the social media strategies of 66 top U.S. colleges and universities on Facebook. It investigates the content and posting frequency of their social media campaigns and analyzes the engagement generated by their posts. The research reveals variations in engagement levels among different institution types and highlights

the influence of media type and posting frequency on engagement. The study offers insights for university social media managers on how to optimize their content and posting strategies to enhance engagement and expand their online reach.

Assimakopoulos et al. [8] presented the use of university Facebook (FB) groups and pages by undergraduate students to obtain departmental information and explores their potential for student acquisition. Two universities, the University of Novi Sad (Serbia) and the Technological Educational Institute of Western Macedonia (Greece), are compared using a structured questionnaire with 343 and 300 student respondents, respectively. The key findings include the introduction of an improved model of FB-related technology acceptance, insights into student demographics and behavioral patterns within the FB group, and the identification of factors that can increase engagement within these groups. The paper offers marketing strategies for higher education institutions to effectively utilize FB and similar social networking sites for marketing and recruitment. A limitation is the relatively small sample sizes, suggesting the need for larger samples in future research. However, the article provides valuable insights into the role of FB as a marketing tool in higher education.

Irfan et al. [9] highlights the importance of marketing on social media, particularly through platforms such as Facebook, in the context of higher education institutions (HEIs). It emphasizes that universities are increasingly focusing on online marketing to attract international students and improve their rankings. The study explores how students frequently use social media to select their universities and draws attention to the importance of a strong online presence for HEIs. The research investigates the impact of social networking sites, such as Facebook, on university rankings and Facebook engagement ratings. It recognizes that young generations and potential students prioritize social media when making decisions about their education. The article concludes that social media plays a vital role in disseminating information about universities and highlights the importance of online marketing for HEIs, particularly in the context of Malaysian universities.

Coelho and Figueira [10] discussed the challenges faced by organizations, including Higher Educational Institutions (HEIs), as they adopt social media platforms. It emphasizes the importance of understanding and assessing the current social media strategy to identify opportunities and threats, evaluate strategies, and make necessary adjustments for enhanced efficiency. To address these challenges, the authors propose an automated

procedure for evaluating the posting behavior and strategy identification of HEIs. They conducted their study using a sample of the top 10 globally ranked educational institutions. Over the course of an entire school year, they collected data from the official Facebook pages of these institutions. The study primarily focused on the frequency and intensity of publications by universities, analyzing factors such as the number of responses (shares) to posts. Additionally, the content of the posts was examined to identify the predominant topics and editorial areas that each university was most active in.

Coelho and Figueira [11] also investigates the impact of the Covid-19 pandemic on the Twitter content strategies of 12 top-ranked Higher Education Institutions (HEIs) based on the 2019 Center for World University Rankings (CWUR). It analyzes over 18,000 tweets and employs topic modeling to identify content topics. The results reveal strategic shifts in response to the pandemic, with health and events topics gaining prominence. Despite varying publication frequencies, engagement levels do not necessarily correlate with frequency. The study proposes a methodology for assessing HEIs social media content strategies and suggests future research directions, including sentiment analysis during the pandemic.

María del Rocío Bonilla Quijada and Olmo-Arriaga [12] studied the ability of the official Instagram accounts of the world's top 5 universities to create engagement among their users. It involves encoding and analyzing all Instagram posts made by these universities during an academic year (a total of 758 posts). The objective is to identify the variables that most contribute to user engagement and provide strategic recommendations for marketing management on this social network. The findings aim to assist social network managers at higher education institutions in understanding publishing patterns and communication strategies that can promote greater engagement and more meaningful relationships with stakeholders.

In 2023 Sörensen et al. [13] represented the first comprehensive examination of social media communication by all 42 Higher Education Institutions (HEIs) in Switzerland over an extended 18-year period. It covers three major social media platforms (Twitter, Instagram, and Facebook) and combines automated analysis with manual content review. Although Swiss HEIs use multiple social media channels, there was no significant overall increase in communication intensity in recent years. Research universities preferred Twitter, universities of applied sciences were active on Instagram and Facebook, while universities of teacher education were less active across all platforms. User engagement

varied across platforms, with Instagram generating considerably more reactions to HEIs posts compared to Facebook and Twitter. The study found limited impact of the Covid-19 pandemic on the intensity, content, or engagement of social media communication by Swiss HEI, except for higher engagement with research university posts in 2020. Overall, the study suggests that Swiss HEI have effectively adopted social media as a complementary channel to traditional communication formats. They focus on topics and audiences distinct from news media, with an emphasis on self-referential and community-building content. While social media use is stable, opportunities for enhanced engagement on platforms like Instagram are noted.

Yalalem Assefa and Kumar [14] focuses on assessing the role of social media in mediating the relationship between student engagement and academic achievement among undergraduate university students. It employs a correlation study design with data collected from 395 participants using a simple random sampling technique. Structural equation modeling was used for data analysis. The findings indicate that both social media usage and student engagement have a significant direct effect on academic achievement. However, when social media is introduced as a mediator variable between student engagement and academic achievement, its mediating effect becomes non-significant, suggesting that social media plays a direct but not a mediating role. This may be due to social media allowing for students' minimal engagement in their learning.

Our approach distinguishes itself by adopting a longitudinal perspective and examining a broader selection of Higher Education Institutions (HEIs), embracing not only those occupying the upper echelons of rankings but also institutions further down the hierarchy. This comprehensive approach anticipates the possibility of detecting variations and trends as we extend our analysis beyond the top-ranked institutions. Furthermore, in addition to differentiating our study through the analysis of HEI's publication strategies.

In the previously published work [15, 16], an investigation was conducted into how Higher Education Institutions (HEIs) utilize social media, particularly Twitter, as a marketing channel and its potential impact on their global rankings. The study assesses the publishing strategies of the selected HEIs and compares them with their rankings. While the analysis based on Kendall τ and RBO metrics does not reveal significant differences in the rankings, the CWUR ranking was used to guide all the analysis. The previous work also presented a sentimental, emotion, and topic analysis using the data collected until

February 2023. Finally, an unsupervised model was applied to try to cluster the universities regarding some statistical metrics that were created to find some segmentation and a good separation aiming to explain the relationship between publication strategies.

2.2 Predictive Modeling using HEIs Social Media Content

Symeonidis et al. [17] explored pre-processing techniques for text classification, particularly in the context of sentiment analysis on Twitter datasets. The research assesses 16 common pre-processing methods and their impact on classification accuracy when applied to machine learning algorithms (Linear Support Vector Classification, Naïve Bayes, Logistic Regression and Convolutional Neural Networks). Techniques such as lemmatization, contractions replacement, and number removal prove effective, while others like punctuation removal show mixed results. The study also investigates the interactions between these techniques when used in combination. Overall, the findings recommend specific pre-processing steps. The recommended techniques are lemmatization, replacing repetitions of punctuation, replacing contractions, and removing numbers, for improved sentiment analysis on Twitter data. The results vary depending on the chosen classifier, and future studies may explore these techniques in diverse domains and with emotion labels.

Álvaro Figueira [18] explores the significance of social media communication strategies for Higher Education Institutions (HEIs). It focuses on Twitter posts from the top 12 globally ranked HEIs during the academic year from September 2019 to October 2020. The objectives were to understand and compare the communication strategies used by these institutions during this period.

Their findings in the publication frequency was that Imperial College had the highest number of Twitter posts, while Stanford University had the lowest. Through topic modeling, the study identified five primary themes in the tweets: university events, public health (Covid-19), research, education, and public image. Their study observed changes in topic distribution before and during the Covid-19 pandemic. Public interaction with HEIs publications, especially in the areas of Education and Health, increased during the pandemic. The authors also attempted to predict future dominant topics based on past posts but showed limited accuracy, indicating that other factors influence the content of new posts. Overall, the work highlights the growing importance of social media communication strategies for HEIs and the impact of external events like the Covid-19 pandemic

on their online engagement. However, predicting the specific topics of future posts remains a challenging task.

Figueira [19] highlights the importance of aligning social media strategies with an organization's overall strategic goals. It focuses on Higher Educational Institutions (HEIs) and proposes a three-step data-mining procedure to assess their social media behavior, understand the communication strategy, compare HEIs through created metrics based on reactions, shares, and comments, and finally, predictive analysis using Random Forests, Naïve Bays, Support Vector Machine and Neural Networks. The research used data from the 5 top-ranked universities in 2017. The model shows promising results in predicting effective social media strategies for HEIs, the accuracy of the classifier was slightly above 80% and the F1-measure was 84%, both values were computed as the average. However, it relies on historical data, and future work aims to develop a self-feedback system to continuously improve strategy recommendations.

Brohi et al. [20] compared the accuracy of machine learning algorithms for predictive analytics in higher education. Traditional algorithms like Naïve Bayes and Random Forest, along with ensemble methods such as Stochastic, Linear Discriminant Analysis (LDA), Tree model (C5.0), Bagged CART (tree-bag), and K Nearest Neighbors (KNN) were used. The study aimed to classify students' performance and identify the most influential independent variables. Results showed that using 11 features with Random Forest achieved the highest accuracy (of 0.7333). The experiment was further refined with ensemble algorithms to reduce variance, and bias, and improve prediction accuracy. Bagging Random Forest outperformed other methods with an accuracy of 0.7959. Behavioral features were considered to be the most crucial, followed by academic background and demographic features. The study highlights the importance of feature selection in improving algorithm accuracy and suggests further exploration of additional features for future research.

Son et al. [21] investigates the dissemination of disaster-related information on Twitter, focusing on the challenges posed by the platform's character limit and the resulting uncertainty in tweets. The researchers introduce entropy as a measure of a tweet's uncertainty and propose that higher entropy leads to lower retweet counts, aligning with Uncertainty Reduction Theory (URT). The study employs statistical and predictive analyses to validate the use of entropy as a reliable measure of tweet uncertainty. It contributes

to our understanding of how information spreads on Twitter during disasters and introduces entropy as a novel variable for assessing tweet uncertainty. From a practical perspective, the research provides guidelines for crafting effective disaster messages on Twitter, recognizing the role of entropy, URLs, and emoticons in conveying information. This study can also help improve the timely dissemination of disaster-related information on social media platforms like Twitter.

Kennedy et al. [22] research focuses on the importance of social networks in marketing and communication strategies, with the global reach of platforms like Facebook. It emphasizes the challenges of effective social media marketing and proposes predictive modeling as a solution. The study demonstrates the benefits of predictive modeling using Facebook data from a Division I football team over two years, with the goal of predicting postseason engagement during an off-season marketing campaign. The results highlight the potential of predictive analytics as a marketing measurement technique that offers insights beyond traditional methods, providing valuable tools for evaluating the success of marketing campaigns in the dynamic world of social media.

Palacios et al. [23] research focuses on applying data mining and machine learning algorithms to predict student retention in higher education institutions (HEIs) using higher education data. Various machine learning algorithms were used and the list was decision trees, k-nearest neighbors, logistic regression, naive Bayes, random forest, and support vector machines to create predictive models, in this way random forest performed better with an accuracy exceeding 80% in most scenarios. The study highlights the importance of secondary education outcomes and community poverty rate as previously unreported variables in predicting student dropout. The research offers practical implications for educational institutions, emphasizing the importance of identifying at-risk students early and implementing targeted retention strategies. Recommendations for future work include the implementation of an information system, a data-monitoring plan, and support programs for at-risk students. Additionally, the study suggests exploring ensemble learning methods and improving classifier explainability for future research.

Soares et al. [24] investigates the determinants of consumer engagement with content generated on Facebook in the context of Higher Education Institutions (HEIs) located in the southern region of Brazil. The study analyzes data from 16 Brazilian HEIs and examines factors like media type, emotional content, and posting timing to predict engagement metrics such as likes, comments, and shares. The findings suggest that emotional content

positively influences likes and shares, while the type of media and posting timing have a less significant impact. The research provides practical knowledge for HEIs that use social media in their marketing strategies and contributes an empirical predictive model for consumer engagement on Facebook in the HEI context, for example, it can perform a crucial role for HEIs by creating networks, improving reputation, enriching relationships and differentiating themselves from the competition.

The literature review highlights a significant gap in studies related to predicting publication strategies in Higher Education Institutions (HEIs) using social media data. Although there have been valuable research efforts in this field, none have adopted a similar approach or covered the comprehensive analyses that will be presented in this study. This research aims to address this gap by providing relevant contributions and insights.

Chapter 3

A System for Automatic Analysis of HEIs' Twitter Posts

In this chapter, we provide a detailed review of the entire process carried out in this study. It will explain the logic behind the choice of a ranking system. We present the data collection methods, which include the development of a dedicated system for retrieving historical tweets from each HEI. Additionally, in the chapter we dive deeper into our exploratory data analysis, covering a variety of analyses such as post distributions, follower counts, and heatmaps to discern temporal patterns, and an examination of the influence of days of the week and months on HEI strategies. In addition, we will detail our sentiment, emotion, and topic analyses for each HEI.

3.1 Data Retrieval

3.1.1 Ranking Systems Analysis

In this work, we chose to investigate possible subtle discrepancies between four widely recognized ranking systems: CWUR, Shanghai, USNews, and QS. This exploration was undertaken despite our knowledge of the findings presented in reference Olcay and Bulu [1].

To facilitate our comparative analysis of these rankings, we employ Kendall's distance and Kendall's correlation coefficient, commonly referred to as "Kendall's τ " [25] and [26]. Kendall distance measures the dissimilarity between two top-k lists and produces a value of 0 for completely identical lists and 1 for entirely dissimilar ones. On the other hand, Kendall τ serves as a metric to evaluate the level of correspondence between two

rankings. A Kendall τ value close to 1 signifies robust agreement, while a value near -1 indicates substantial disparity.

Additionally, we apply the Rank Biased Overlap metric, called "RBO," which offers a different perspective when comparing ranked lists. In this context, an RBO score of 1 means an identical ranking, while a score of 0 means completely disjoint lists. In particular, RBO is particularly adept at handling the weighting of items of highest ranking [27].

Our primary objective was to assess whether any of the rankings exhibited notable disparities in comparison to the others. Notably, in all instances of ranking comparisons, the Kendall distance metric yielded a result closest to 1 than -1, indicating an absence of dissimilarity among the rankings.

Regarding Kendall's τ and RBO metrics, the results were as follows:

- **CWUR vs. Shanghai**, Kendall τ resulted in 0.64, while RBO yielded a value of 0.95.
- **CWUR vs. USNews**, the Kendall τ and RBO scores were 0.63 and 1.00, respectively.
- **CWUR vs. QS**, Kendall τ produced a result of 0.47, while RBO indicated a score of 0.05.

Although there was no similarity between the CWUR and QS rankings, the general conclusion drawn from these findings was that there were no significant variations between the rankings CWUR, Shanghai and USNews. Consequently, we proceeded with the CWUR ranking as the basis of guidance for our further analysis.

For our data collection, we collected posts from Higher Education Institutions (HEIs) occupying positions ranging from 1 to 10 in the CWUR ranking, as well as those positioned at 100, 200, 300, 400, and 500. This deliberate inclusion of a wide spectrum of ranking positions serves a dual purpose: firstly, it offers insights into the performance of top-ranked HEIs, and secondly, it allows us to explore potential variations across a diverse range of ranking positions.

Table 3.1, provides a comprehensive overview of these positions and their corresponding rankings across the four prominent indices. A careful examination of the table reveals that, for the HEIs we specifically selected, by using CWUR, the disparities in rankings did not have significant consequences within the scope of our investigation.

Some modifications were necessary in the selection of Higher Education Institutions (HEIs) for our analysis. Instead of opting for position 100, we deliberately chose position

99. This decision was made due to Keio University (position 100) discontinuing its activity on Twitter after April 2020. Additionally, we introduced the inclusion of the University of Porto, motivated by author affiliation and curiosity.

TABLE 3.1: The position of the HEIs on the four rankings chosen in 2022.

Higher Education Institution	CWUR	Shanghai	USNews	QS
Harvard University	1	1	1	5
Massachusetts Institute of Technology	2	3	2	1
Stanford University	3	2	3	3
University of Cambridge	4	4	8	3
University of Oxford	5	7	5	2
Princeton University	6	6	16	20
University of Chicago	7	10	15	10
Columbia University	8	8	6	19
University of Pennsylvania	9	15	13	13
California Institute of Technology	10	9	9	6
Boston University	99	101-150	65	112
University of Lisbon	200	201-300	197	356
University at Buffalo	300	301-400	280	338
University of Porto	308	201-300	255	295
University of Oklahoma, Norman	400	501-600	425	651-700
Federal University of Minas Gerais	500	401-500	256	651-700

To demonstrate that the variations in ranking positions remain relatively stable and to provide insights into the changes introduced in the new 2023 version, we have chosen to present Table 3.2. This table showcases an updated and revised set of data taking into account the most recent iteration of the rankings of universities in the world. The aim is to offer a comprehensive overview of the evolving landscape and to shed light on the changes and developments that have taken place in this latest edition.

Upon careful analysis of the most substantial changes noted in the CWUR ranking, one notable shift was the California Institute of Technology (Caltech), which moved from the previous 10th place to the 11th, with Yale University moving up to claim 10th spot. Additionally, in this new version of the ranking, for the HEIs that had not previously held the top 10 in the ranking, they apart experienced noteworthy alterations.

In a positive turn of events, Boston University made a significant jump, securing the 90th position, reflecting a notable improvement in its ranking. On the other hand, some HEIs, including University of Lisbon, University of Buffalo, University of Porto, University of Oklahoma, and Federal University of Minas Gerais, registered slight declines in their positions in the CWUR ranking. However, it's worth noting that these negative

TABLE 3.2: The position of the HEIs on the four rankings chosen in 2023.

Higher Education Institution	CWUR	Shanghai	USNews	QS
Harvard University	1	1	1	4
Massachusetts Institute of Technology	2	3	2	1
Stanford University	3	2	3	5
University of Cambridge	4	4	8	2
University of Oxford	5	7	5	3
Princeton University	6	6	16	17
University of Chicago	7	10	22	11
Columbia University	8	8	7	23
University of Pennsylvania	9	14	15	12
California Institute of Technology	11	9	9	15
Boston University	90	101-150	70	93
University of Lisbon	202	201-300	202	266
University at Buffalo	305	301-400	286	491
University of Porto	309	201-300	247	253
University of Oklahoma, Norman	405	701-800	425	641-650
Federal University of Minas Gerais	503	401-500	468	691-700

variations ranged from 2 to 5 positions, and their overall impact on the rankings remained relatively modest.

When assessing the Shanghai, USNews, and QS rankings, we observed that, in general, there were subtle changes in the positions of the analyzed universities. However, it's important to highlight that some significant and impressive changes were also evident.

For example, in the USNews ranking, the Federal University of Minas Gerais suffered a substantial drop, losing 212 positions in the new version. This significant change has notable implications for the university in the context of this ranking.

Similarly, in the QS ranking, the University of Buffalo witnessed a significant negative change, dropping by 153 positions. This sharp decline represents a substantial change in its position in the QS ranking.

On a positive note, both Portuguese universities, the University of Lisbon and the University of Porto, demonstrated notable improvements in their rankings. The University of Lisbon surged ahead by 90 positions, while the University of Porto made a significant gain, climbing 42 positions. These positive changes reflect the improvement in their positions in the QS ranking.

3.1.2 Twitter Data Collection

In the initial phase of our data collection process, we implemented an internal tweet collector, using the Twitter API v2 in Python, with the primary objective of obtaining the

most recent 2,500 tweets for each Higher Education Institution (HEI) under investigation. Our data collection approach entailed establishing a post-termination cutoff date of July 31, 2022. Subsequently, we executed two distinct rounds of tweet extraction, one conducted on August 5, 2022, and the other on August 17, 2022. It is important to note that for the University of Lisbon, only 1,583 tweets were successfully retrieved, and for the University of Buffalo, the count reached 1,235 tweets. This discrepancy arose because the Twitter API did not furnish the complete 2,500 tweets for these particular HEIs. Additionally, we incorporated a specific data collection strategy that excluded retweets from the process for all institutions.

To increase the size of our dataset, we continued the collection of recent tweets on a monthly basis, starting at the beginning of each month. This extended data collection effort persisted until May 2023, allowing us to capture tweets published after the initial analysis period.

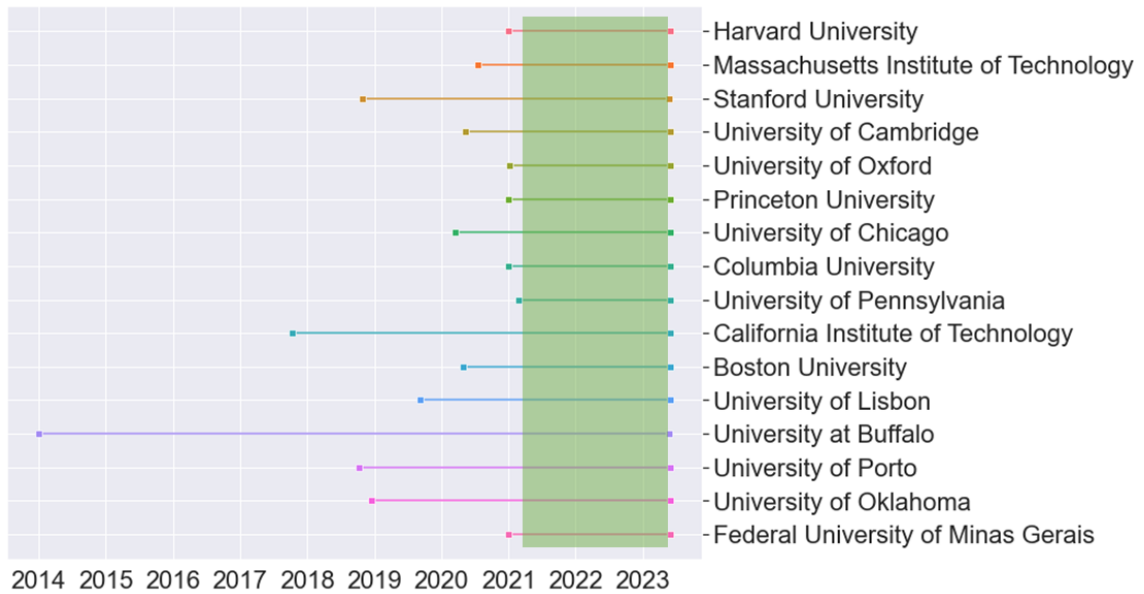


FIGURE 3.1: Data recovery period for each HEI and its common intersection period.

It is crucial to recognize that discrepancies in publication frequencies among HEIs led to variations in the temporal coverage of retrieved tweets for each respective institution. As exemplified in Figure 3.1, we present a graphical representation delineating the overlapping time frame during which tweets were disseminated across all HEIs. The emphasis on the most expansive common period, extending from July 2021 to May 2023, a time interval guided by University of Pennsylvania, which presented the largest number of publications.

In Figure 3.2, presented below, we illustrate the tweet count per Higher Education Institution (HEI) over the entire data collection period. In contrast, Figure 3.3 provides the same information but restricts the analysis to a common period.

Upon closer examination of these two figures, it becomes evident that the University of Pennsylvania consistently exhibits a higher frequency of tweet publication, whereas the University at Buffalo appears to have a notably lower rate of tweet activity. It is worth noting that while there are some fluctuations in tweet counts among various universities, these variations remain relatively minor.

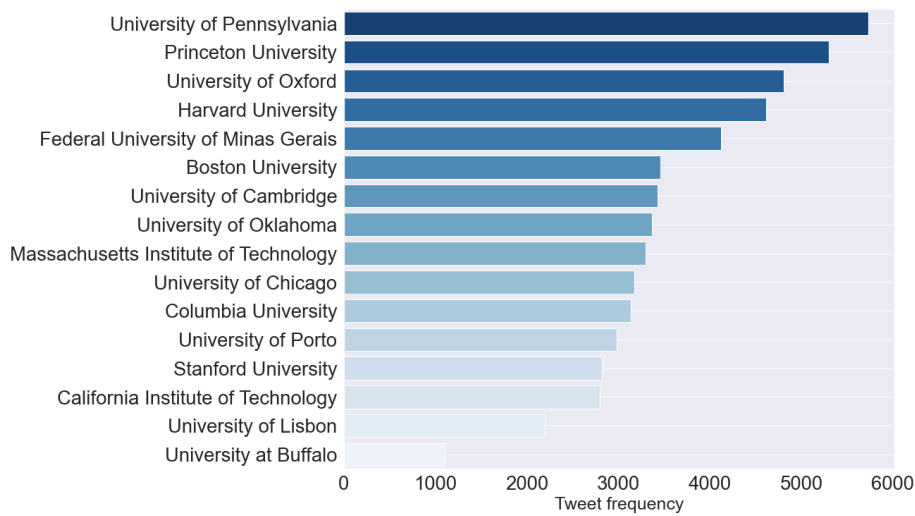


FIGURE 3.2: Number of tweets published per each HEI in all the collected period.

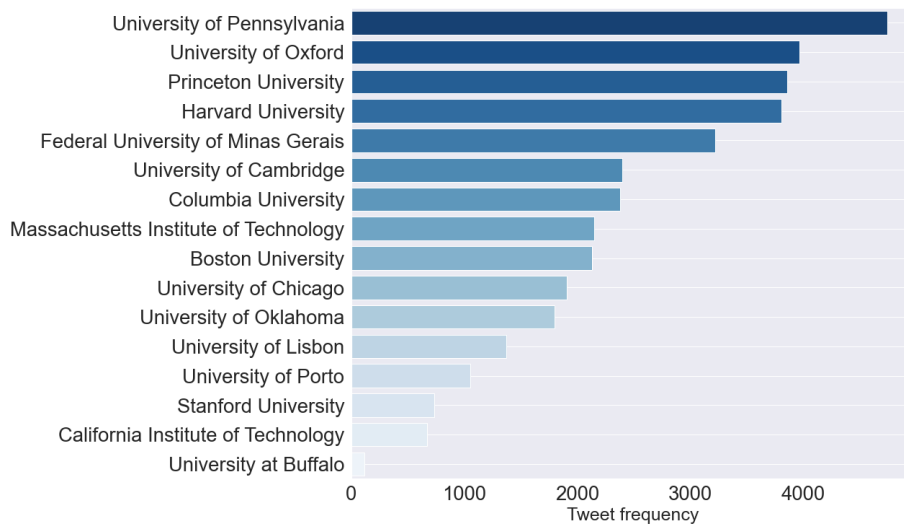


FIGURE 3.3: Number of tweets published per each HEI in the common period.

In the next section, we delve into a comprehensive examination of the collected data, where we conduct an in-depth analysis focusing on publishing timestamps and content.

3.2 Exploratory Data Analysis

3.2.1 HEIs Number of Followers

Examining the data presented in Figure 3.4, it becomes evident that Harvard University has the largest number of followers, surpassing the 1.56 million mark. Closely behind, the Massachusetts Institute of Technology (MIT) commands a substantial number of followers, exceeding 1.27 million. Stanford University and the University of Oxford maintain a comparable level of prominence, with more than 928,000 followers each. Likewise, the University of Cambridge has a substantial following of over 776,000.

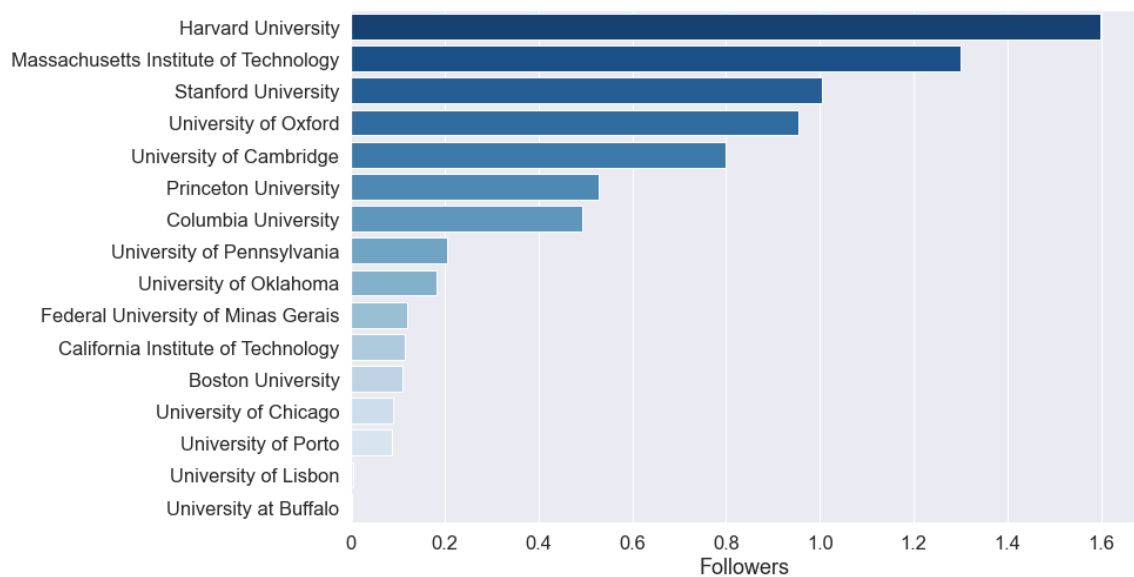


FIGURE 3.4: Number of followers, in millions, as of June 2023.

In descending order of follower count, Princeton University and Columbia University each command the attention of 520,000 and 486,000 followers, respectively. Going even further down the scale, the University of Pennsylvania and the University of Oklahoma attract the attention of more than 201,000 and 179,000 followers, respectively. The Federal University of Minas Gerais, the California Institute of Technology, and Boston University each garner a significant number of followers, boasting more than 119,000, 114,000, and 109,000 followers, respectively.

Continuing with the analysis, the University of Porto and the University of Chicago maintain a presence with more than 86 thousand followers each. Bringing up the rear in this ranking are the University of Lisbon and the University at Buffalo, each commanding a comparatively modest number of followers, with less than 11,000 followers, in the specified sequence.

In Figure 3.5, a boxplot representation displays the distribution of retweet counts, while Figure 3.6 shows a boxplot representing like counts. Notably, MIT, Harvard, Oxford, Cambridge, and Stanford emerge as prominent universities in terms of both median values and upper-bound outliers. On the other hand, institutions such as Lisbon, Buffalo, and Porto have comparatively lower retweet and like counts.

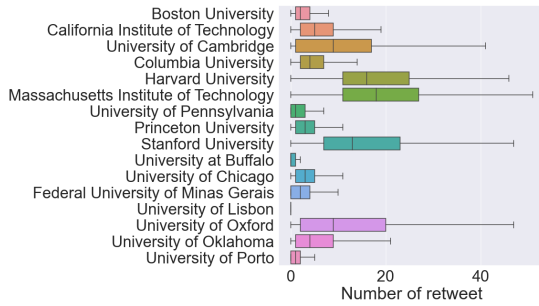


FIGURE 3.5: Retweet boxplot without outliers.

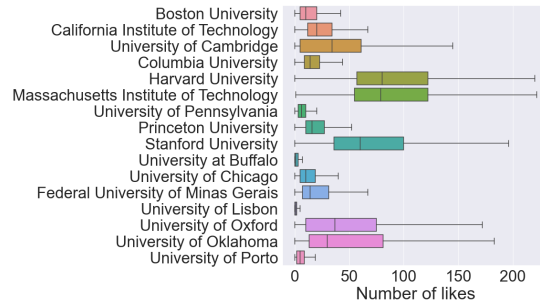


FIGURE 3.6: Like boxplot without outliers.

This visualizations underscores the level of engagement that certain higher education institutions have compared to others. In this context, institutions with significant engagement occupy the top positions in the rankings. The other way round, those with lower engagement do not appear within the top 10 ranked institutions.

3.2.2 Data Distribution Analysis

Table 3.3, presented below, provides an overview of the average, maximum and standard deviation of the daily tweet frequency for all Higher Education Institutions (HEIs). We have observed that the California Institute of Technology has the lowest standard deviation in its posting frequency, indicating a more consistent posting strategy, with an average of approximate two posts per day. On the other hand, the University of Lisbon has the highest standard deviation in its posting frequency, which implies that there are occasional days when this university's account publishes a considerably higher number of tweets. It was found that the University of Lisbon registered a notably high number of posts, specifically 355, on September 25th, 2021. This increase in its activity can be attributed to the university's response to the approval of new students during that specific period.

At the same time, it is noteworthy that the University of Chicago, the University of Oklahoma-Norman, Stanford University, the University of Porto, the University at Buffalo, and the California Institute of Technology consistently maintain a posting frequency

TABLE 3.3: Posting daily frequency average, maximum and standard deviation (in decreasing order).

Rank	Higher Education Institution	Average	Maximum	Std.
9	University of Pennsylvania	7.23	18	3.00
500	Federal University of Minas Gerais	6.39	22	3.49
6	Princeton University	6.16	42	3.91
5	University of Oxford	5.59	88	5.78
1	Harvard University	5.26	19	1.71
8	Columbia University	3.99	21	2.81
200	University of Lisbon	3.77	355	22.64
4	University of Cambridge	3.46	100	5.13
2	Massachusetts Institute of Technology	3.24	14	1.76
99	Boston University	3.17	40	2.29
7	University of Chicago	2.84	20	2.11
400	University of Oklahoma, Norman	2.73	19	1.82
3	Stanford University	2.51	40	2.52
308	University of Porto	2.23	22	1.61
300	University at Buffalo	2.12	40	3.28
10	California Institute of Technology	2.03	19	1.60

within the range of two to three posts per day, accompanied by standard deviations that fall between 2 and 3. This observation highlights the regularity and consistency in posting frequencies from these universities, indicating a stable and well-maintained posting pattern.

Continuing our analysis of Table 3.3, we turn our attention to Columbia University, the University of Lisbon, the University of Cambridge, the Massachusetts Institute of Technology, and Boston University. These institutions present a daily variation in tweet frequency which varies between 3 to 4 tweets per day, accompanied by standard deviations that vary from 2 to 23. However, it is important to note that the University of Lisbon's data includes some outliers in their posting frequency, as previously discussed.

If we exclude the outlier value from the University of Lisbon's data set, the standard deviation decreases considerably, falling within a range of 2 to 5. This adjustment demonstrates that, in addition to occasional outliers, these institutions present a more consistent and regular posting pattern, with variations in daily tweet frequency clustered mainly around 3 to 4 tweets per day. This perception underscores the propensity of these universities to maintain a relatively stable posting frequency during the examined period.

Concluding our analysis of Table 3.3, we turn our attention to the University of Pennsylvania, the Federal University of Minas Gerais, Princeton University, the University of

Oxford, and Harvard University. These institutions stand out for their high daily frequency of tweets, varying between 5 and 7 posts per day. The associated standard deviations range from 2 to 6, indicating varying levels of consistency in their posting patterns.

These universities, with their robust and frequent posting activities, are evidently focused on building and maintaining their online presence, which is likely to improve their image and attract resources. However, it is worth noting that, particularly in the case of the University of Oxford, there are occasional days marked by even greater variations in its posting frequency, as evidenced by the higher standard deviation within this group. This suggests that, although these institutions generally maintain a high level of publishing activity, there are periods of greater variability in their posting patterns, potentially reflecting unique events or initiatives that encourage massive online engagement.

To support the analysis presented in Table 3.3, we also present the daily frequency results for all Higher Education Institutions (HEIs), 3.7. These frequencies are presented for the entire period of data retrieval, allowing for visual confirmation of all the information mentioned above.

Figure 3.8 provides a visual representation of the boxplot depicting the daily tweet frequency for each university in our analysis. Notably, Harvard University and the Federal University of Minas Gerais (UFMG) present daily tweet frequency distributions that approximate a normal distribution pattern. However, it is important to recognize the existence of a modest number of outliers in these distributions, 2 for Harvard and 5 for UFMG, all of which manifest beyond the upper threshold.

The Massachusetts Institute of Technology (MIT) and the University of Pennsylvania indeed display strikingly similar patterns in their boxplots, Figure 3.8, and the corresponding distribution graphs, as depicted in Figure 3.9. This similarity can be attributed to the presence of bimodal distributions in both cases, denoting the existence of two distinct modes or means in their data. Consequently, the medians in the boxplots for these institutions are visible and positioned above the center of the boxes, a clear manifestation of this bimodal behavior in their distributions. This phenomenon highlights the presence of two dominant modes in the daily tweet frequency data for MIT and the University of Pennsylvania, signifying distinct posting patterns or periods of intense activity.

Concluding our analysis of the boxplot in Figure 3.8 and the distribution graphs in Figure 3.9, a similar pattern emerges among all the remaining Higher Education Institutions (HEIs). In the boxplot representation, this behavior is characterized by a flatted box

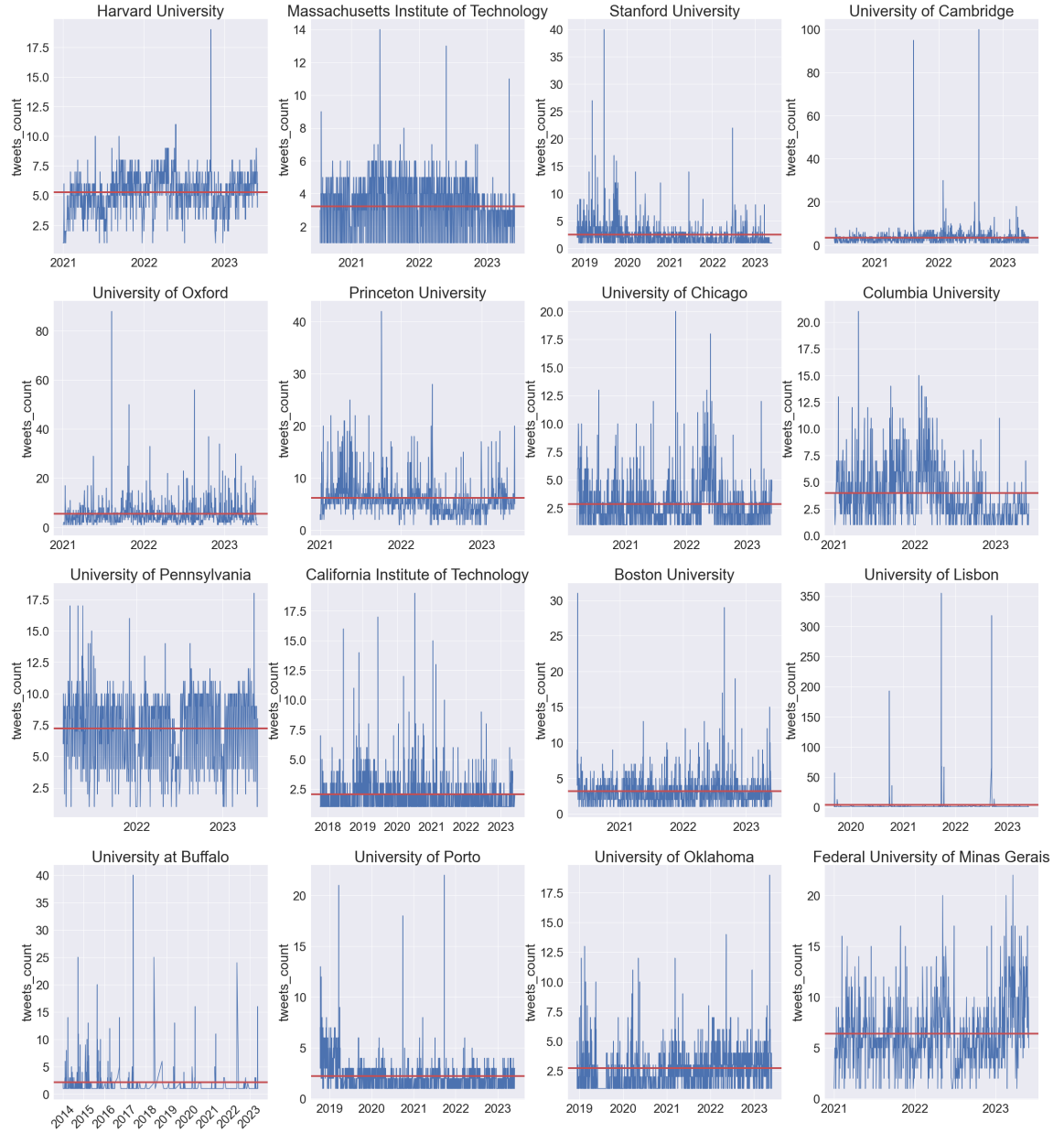


FIGURE 3.7: Daily tweet frequency for each HEI.

of relatively low height with a notable prevalence of outliers. At the same time, in the distribution graphs, this uniformity is evident through a skewed curve that highlights the distribution's amplitude while recognizing the presence of these outliers.

To improve the visualization of this shared pattern, the decision was made to introduce upper and lower bounds to all the distribution graphs. These bounds are determined by a range of three standard deviations from the mean, providing a more comprehensive perspective on the data's distribution and emphasizing the significance of the outliers present in the dataset. This approach allows for a more nuanced understanding of

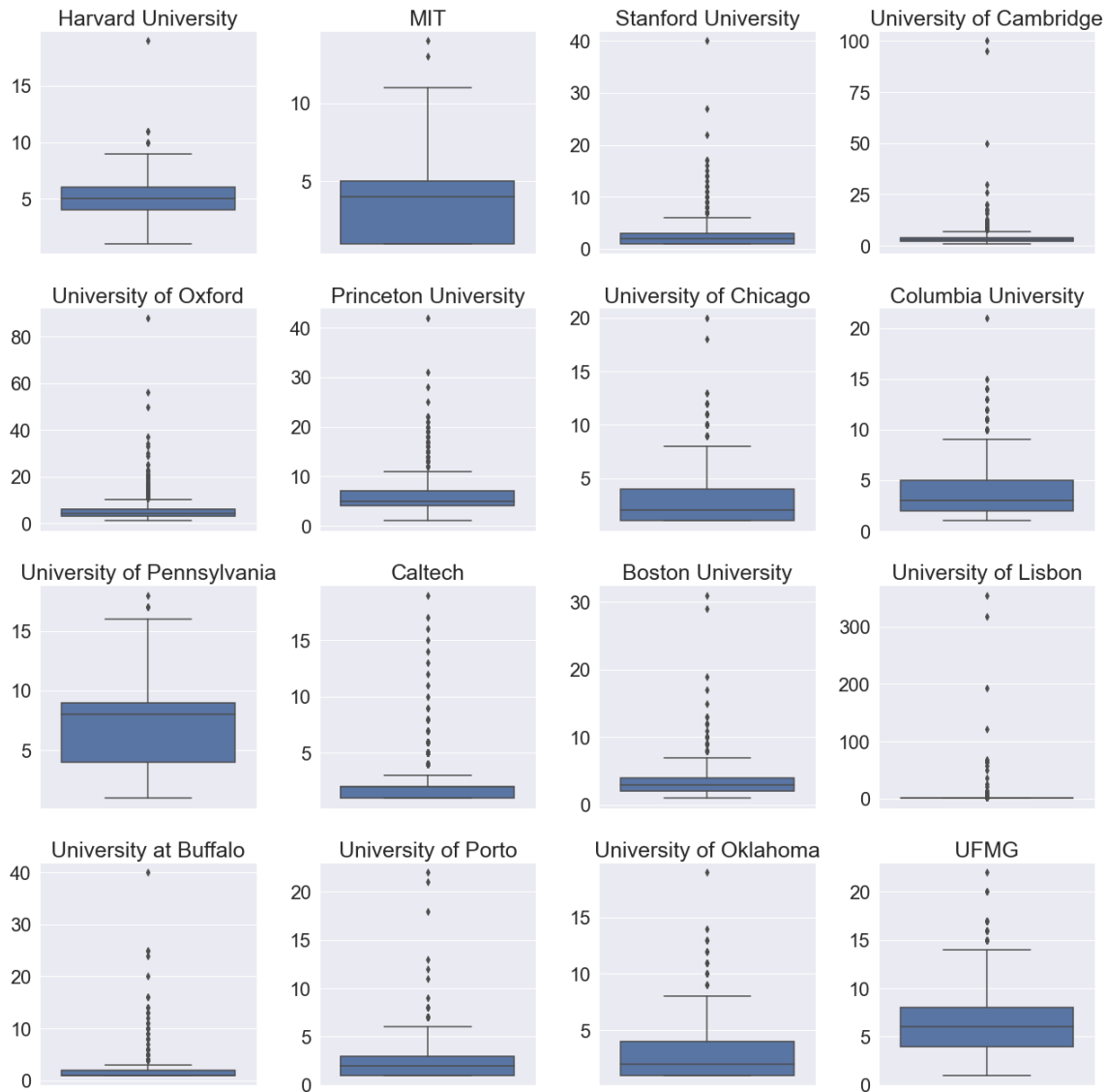


FIGURE 3.8: Boxplots for the daily posting frequency for each HEI.

the tweeting behavior exhibited by these HEIs, which, although consistently distributed around a central tendency, also exhibit notable variations, and a skewed curve, marked by the presence of outliers.

3.2.3 Publication Period of the Day

To thoroughly examine the posting frequency across different days of the week and times, we formulated a tweet frequency table covering all Higher Education Institutions (HEIs) within the shared time period. This table crossed the day of the week with the posting time hour, resulting in the heatmaps shown in Figure 3.10 and Figure 3.11.

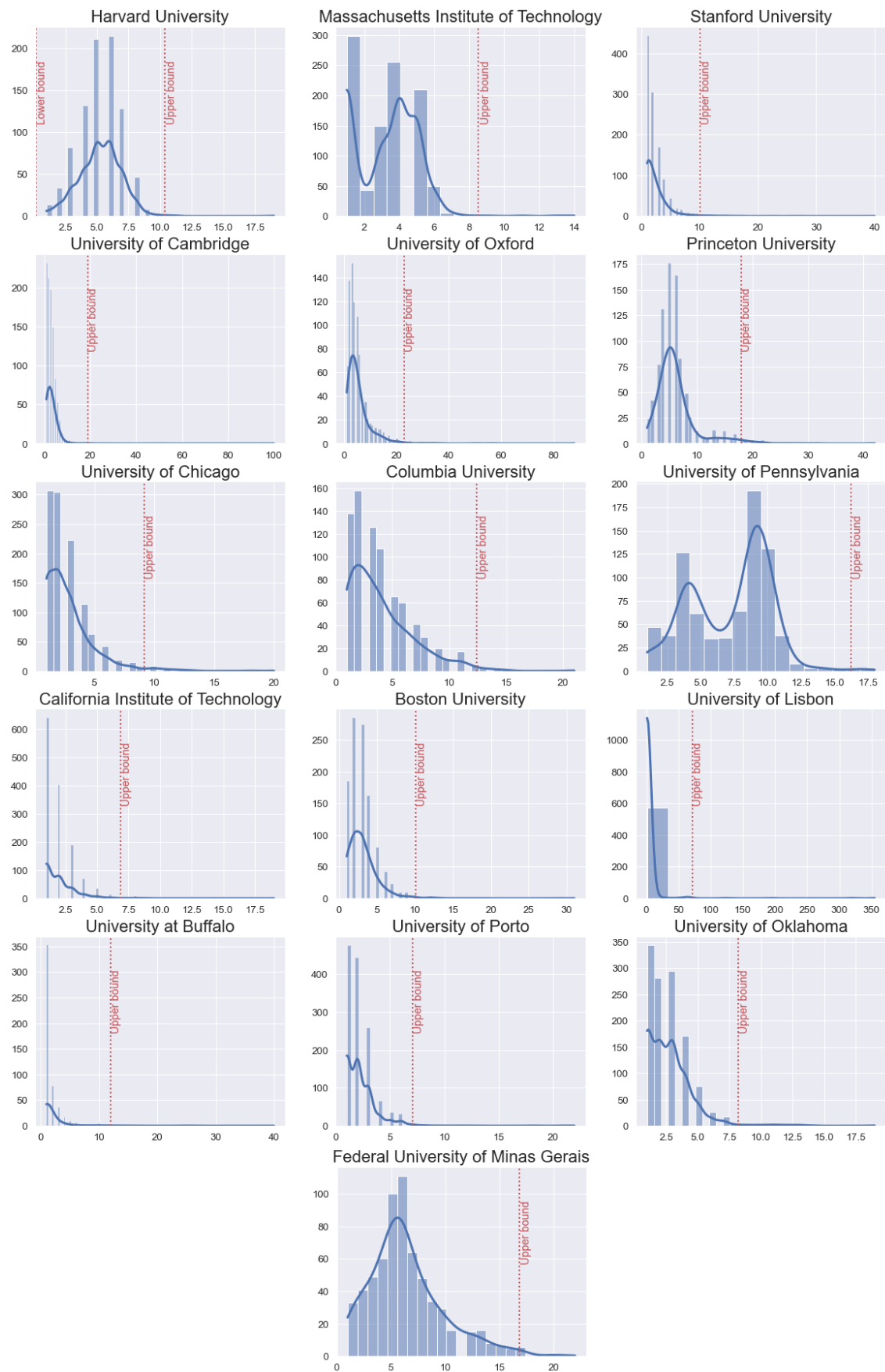


FIGURE 3.9: Daily distribution posting frequencies for each HEI with upper and lower boundaries.

A comprehensive examination of the heatmaps revealed several notable observations. In particular, the Universities of Pennsylvania and Oklahoma exhibited a concentration of posts predominantly between 2 p.m. to 9 p.m. on weekdays, while Federal University of Minas Gerais and University of Porto show a concentration of post mostly between 9 a.m. to 20 p.m. On the other hand, institutions such as Harvard, Princeton, Stanford and Chicago demonstrated a continuous posting pattern throughout the week and weekend, although mostly during normal working hours.

A particularly intriguing pattern emerged as we delved deeper into the data. Specifically, we noted that MIT, the University of Pennsylvania, the University of Oklahoma, the California Institute of Technology and Boston University displayed a high-frequency posting strategy that was distinctly concentrated within a brief time frame on weekdays. This pattern hints at a deliberate and systematic approach to external communication, suggesting a concerted editorial strategy.

On the other hand, institutions such as the University of Cambridge, University of Oxford, Federal University of Minas Gerais, University of Porto, and Boston University exhibit a considerable median frequency of activity during the weekend. However, this activity is predominantly confined to standard working hours. This observation suggests that the timing of posts may be less significant than their consistency.

In conclusion, our analysis of Figure 3.10 and Figure 3.11 reveals that the University of Lisbon and the University of Buffalo do not present a distinct pattern in their publishing behavior. This absence of a discernible pattern can likely be attributed to the low frequency of tweets posted by these institutions.

The irregularity in posting frequency and the absence of a clear posting pattern may result from various factors, including a lower volume of content generation, a less structured social media strategy, or differences in audience engagement. Understanding these patterns, or lack thereof, can be valuable for these institutions as they assess their social media engagement strategies and seek opportunities for improvement.

Figure 3.12 and Figure 3.13 provides heatmaps illustrating daily tweet counts by Higher Education Institutions (HEI) based on the day of the week (y-axis) and month of the year (x-axis) for the year 2022. This visual representation reveals several key insights.

Harvard, MIT, Pennsylvania, Oklahoma, and Porto display "hot" heatmaps, meaning consistently high tweet activity throughout the year. Notably, Harvard displays two distinct peaks — one in April, on a Friday, and another in October, on a Monday.

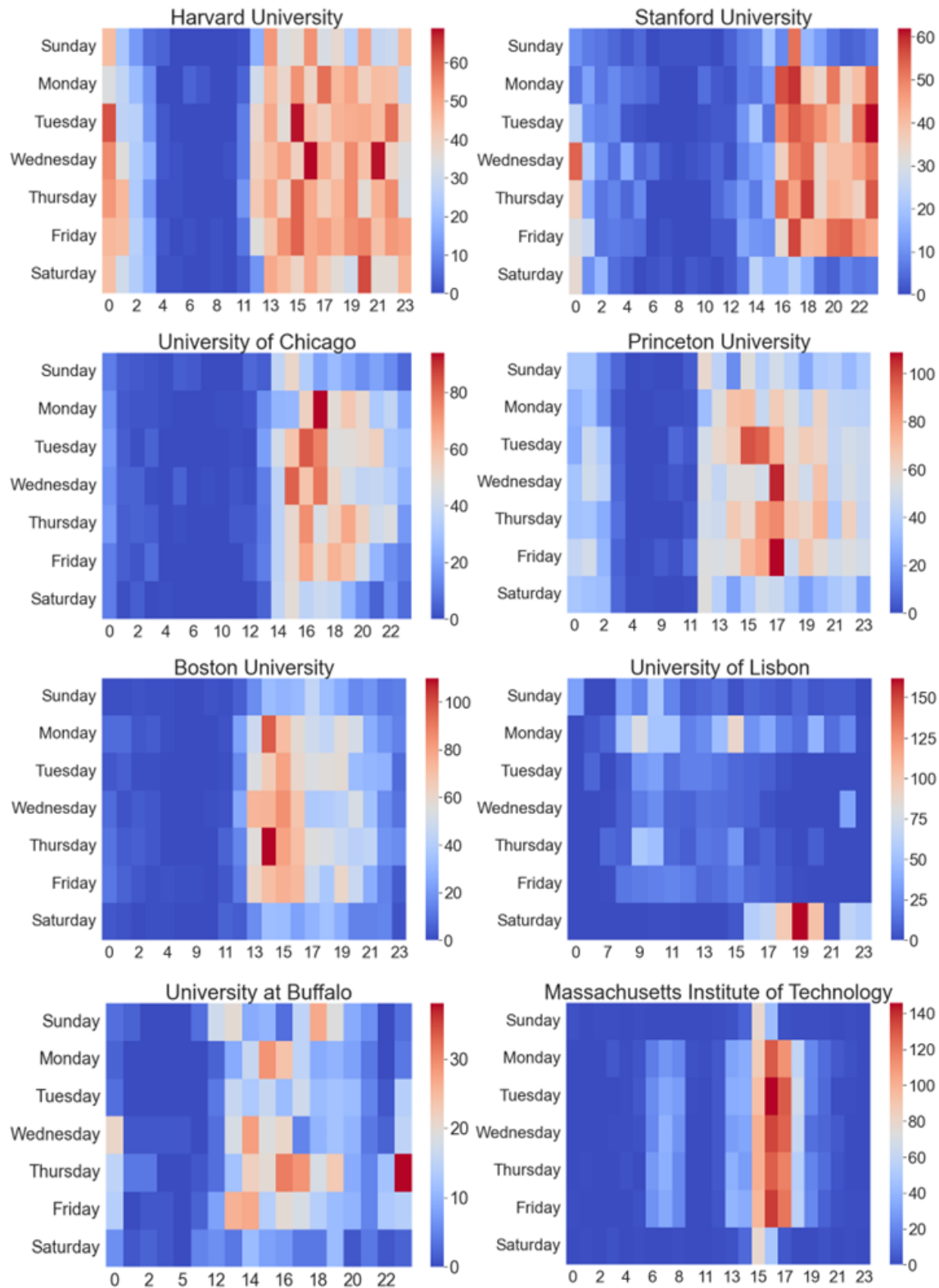


FIGURE 3.11: Daily publication of HEIs crossed by the day of the week and time hour, second part.

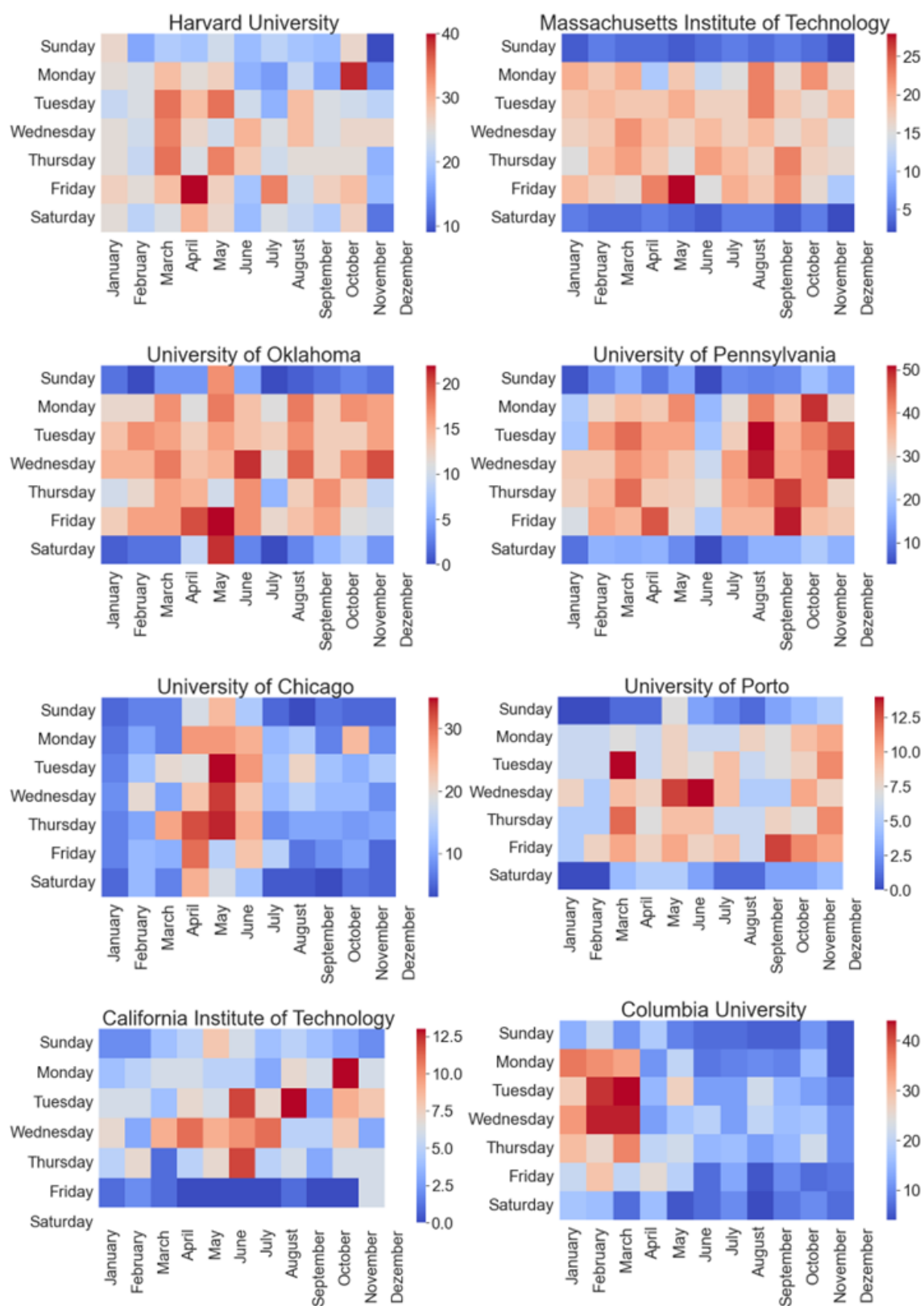


FIGURE 3.12: Daily publication of HEIs crossed by the day of the week and month of the year in 2022, first part.

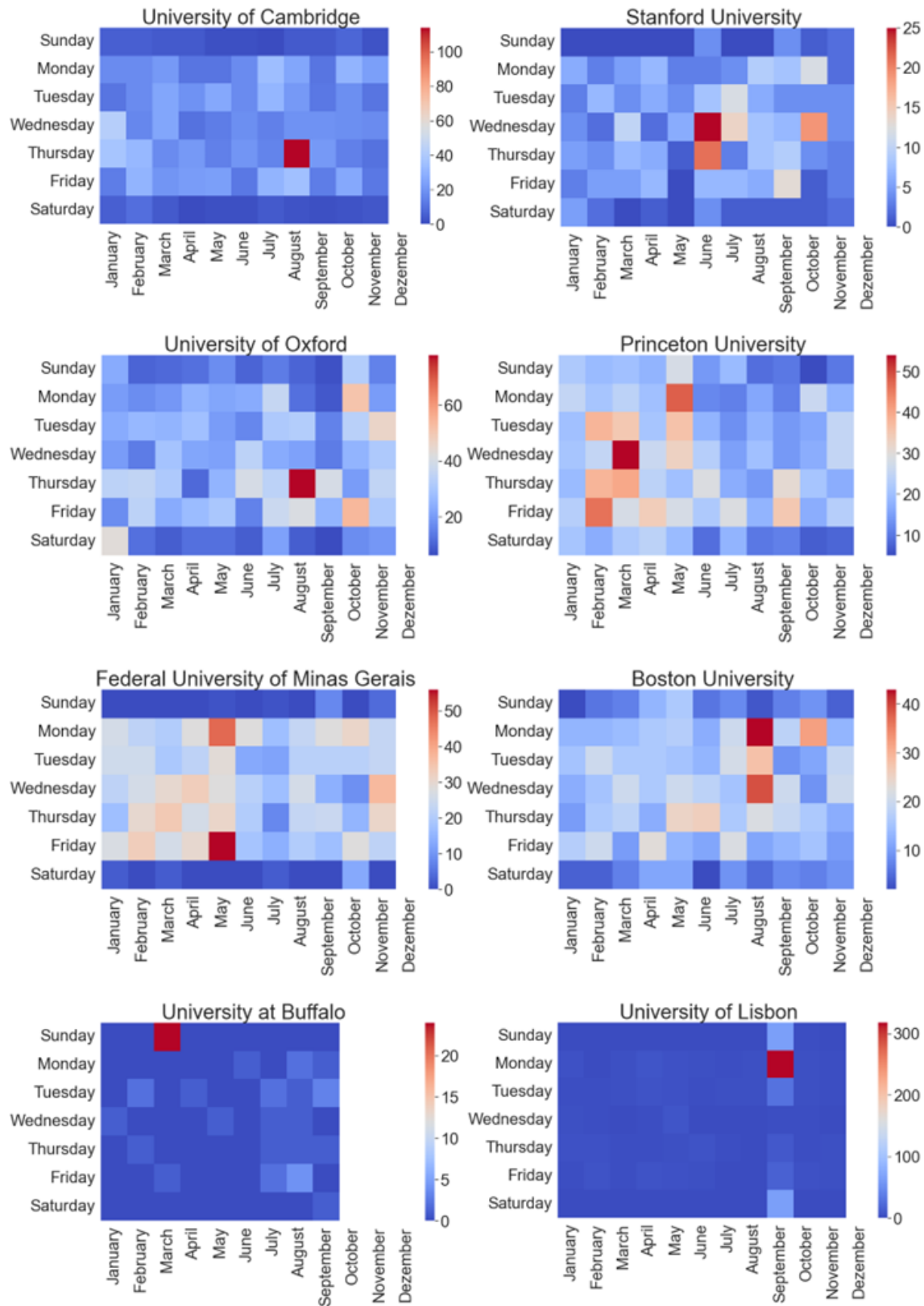


FIGURE 3.13: Daily publication of HEIs crossed by the day of the week and month of the year in 2022, second part.

On the other hand, institutions such as Stanford, Cambridge, Lisbon, and Buffalo appear to have "cold" heatmaps, indicating lower tweet frequency throughout the year.

Other institutions display a mix of "cold" and "hot" pixels, highlighting the fluctuating nature of their tweet posting frequency throughout the year.

Indeed, the observation of "cold" pixels consistently occurring on Saturdays and Sundays for all institutions across all months of the year is quite significant. This pattern aligns with a common expectation in social media strategy, where the engagement and attention of students and other interested parties tend to decrease during the weekends, as the literature suggests that the time of day can influence user engagement [28, 29].

This phenomenon underscores the importance of timing in social media posting strategies for Higher Education Institutions (HEIs). Recognizing that weekends may not yield optimal engagement, HEIs can strategically plan their content publication to align with periods when their target audience is more likely to be active and receptive. Such insights into posting patterns can contribute to more effective communication and outreach efforts.

3.2.4 Word Cloud Analysis

In our research, we have also generated a series of word clouds corresponding to each Higher Education Institution (HEI) created from the entire set of posts during the shared posting period. To achieve this objective, we employed the Python WordCloud library. As illustrated in Figure 3.14, these word clouds provide a visual representation of the most frequently occurring terms within the posts of each HEI.

Before applying the Python WordCloud library, the data underwent preprocessing. The punctuation, special characters, numbers, and links were removed. After that, lemmatization was applied.

Our observations reveal a recurring trend in which the majority of HEIs give priority to projecting their institutional identity, as evidenced by the prominent usage of their own names as the most prevalent term in their publications, always appearing as the biggest word. However, a notable departure from this pattern is discerned in the case of Boston, Lisbon, and Oklahoma, where these institutions exhibit a more diverse lexicon, with the utilization of alternative terms being comparatively more frequent.

Furthermore, our analysis of the word clouds underscores the pervasive presence of terms such as "new", "student", "year", "congratulation" and "research" across almost

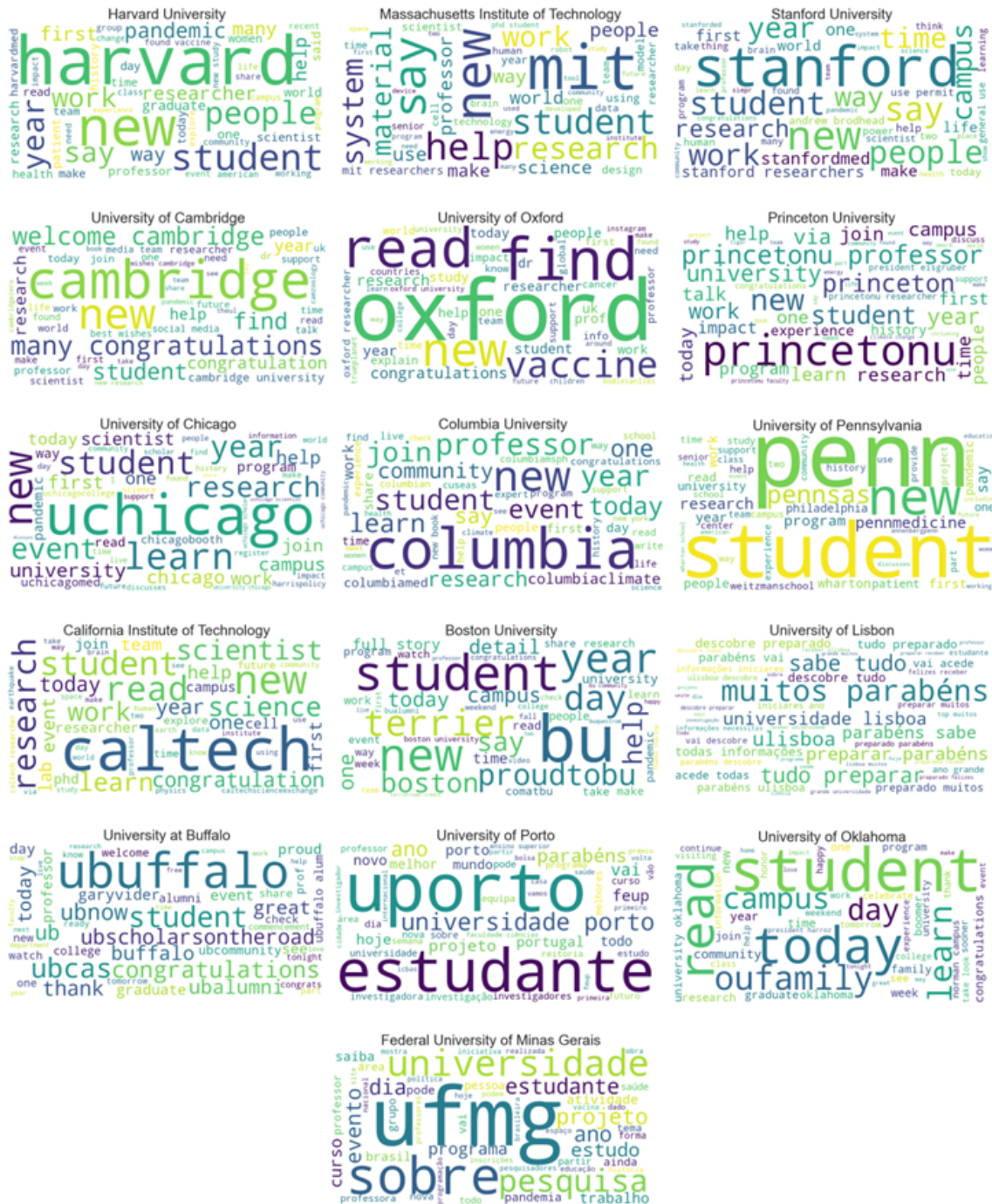


FIGURE 3.14: Word cloud considering all the retrieved publications for each HEI.

all HEIs. This ubiquity highlights the consistent emphasis placed by these institutions on topics related to student engagement and research efforts, i.e., their communication efforts are targeted and directed at specific reader segments.

A notable observation that emerges from our analysis is the recurring prominence of terms such as "research," "scientist," "researchers," and "work" across all the word clouds representing the Higher Education Institutions (HEIs). This consistent thematic emphasis underscores the concerted efforts undertaken by these institutions to actively engage with and convey their commitment to research-related efforts through their social media presence.

The widespread presence of these terms serves as a convicting indication of the primary importance attributed to research activities within the HEI community. It is clear that these institutions highly value research as a central aspect of their identity and mission, as evidenced by their consistent communication on this topic via social media platforms.

This deliberate focus on research-related topics underscores the profound impact and importance of research in academic settings. It reflects the commitment of HEIs to promoting a culture of research, fostering scientific research, and showing their contributions to knowledge generation and innovation. In doing so, these institutions not only engage their immediate community but also convey their dedication to advancing academic pursuits and addressing complex challenges of societal relevance.

It is important to highlight that the University of Lisbon places a relatively lower degree of emphasis on terms such as "student" and "research" compared to its counterparts. On the other hand, the University of Porto and the Federal University of Minas Gerais (UFMG) opt for their Portuguese equivalents, "estudante" and "pesquisa," respectively, in their communication.

Moreover, our analysis reveals a common trend among most HEIs, in which engagement initiatives aimed at new members of their respective communities prevail. This engagement often takes the form of congratulatory messages employing terms like "welcome", "first," "year," "new", "student" and "congratulation". Such initiatives serve as a welcoming gesture to newcomers and promote a feeling of inclusion.

In addition, it is worth mentioning that terms such as "pandemic" and "vaccine" continue to be highlighted in posts by Harvard, Chicago, UFMG, and Oxford. This divergence in editorial focus, distinct from other HEIs, suggests a notable difference in the

thematic priorities of these institutions, possibly motivated by unique circumstances or research efforts.

When we narrow our focus to the common publishing period, as illustrated in Figure 3.15, only two small changes are observed: (a) an increase in engagement actions on Columbia's posts compared to other terms and (b) a reduction in the prominence of branding and projecting institutional image at the University of Porto. These nuances provide additional information about the evolution of these HEIs' communication strategies during the common posting period.

In summary, our comprehensive word cloud analysis indicates that although the different Higher Education Institutions (HEIs) publish content at different intervals and frequencies, there is a notable similarity in their regarding concerning textual content. This similarity extends to the themes, keywords, and engagement strategies used on these institutions' social media platforms.

However, an intriguing aspect emerges from this observation. It raises questions about whether there exists a direct correlation or mapping between the ranking lists of these HEIs and the publishing strategies and patterns they employ in their social media efforts. Of course, the alignment of social media strategies with institutional ranking is not a straightforward one-to-one correspondence.

The complexities and divergences found in university ranking methodologies and the multifaceted nature of social media engagement make it challenging to establish a clear causal relationship between the two. While both ranking and social media presence are integral aspects of an institution's identity and reach, several factors contribute to its rankings, including academic performance, research excellence, faculty, and infrastructure.

Thus, our analysis underscores the need for a more nuanced examination to unravel the complex interplay between institutional ranking and social media strategies. It highlights the multifaceted nature of university communications and underscores that a holistic understanding of an institution's performance and reputation goes beyond its social media presence to encompass a range of factors that influence its position in the academic landscape.

3.2.5 Sentiment Analysis

To delve deeper into the content of the tweets, we conducted a sentiment analysis of each post across all Higher Education Institutions (HEIs). To facilitate this analysis, we utilized

the TextBlob library (version 0.16.0) within a Python implementation. This library categorizes each tweet into one of three categories: positive, neutral, or negative, providing a quantitative representation of the sentiment contained in the text.

As an illustrative example, we provide a representation of the sentiment analysis conducted on the tweets from Harvard University, MIT, and Stanford University along a temporal continuum, as showcased in Figure 3.16. In this representation, tweets were aggregated by month to provide a comprehensive view of the sentiment dynamics within the Twitter activity of these institutions.

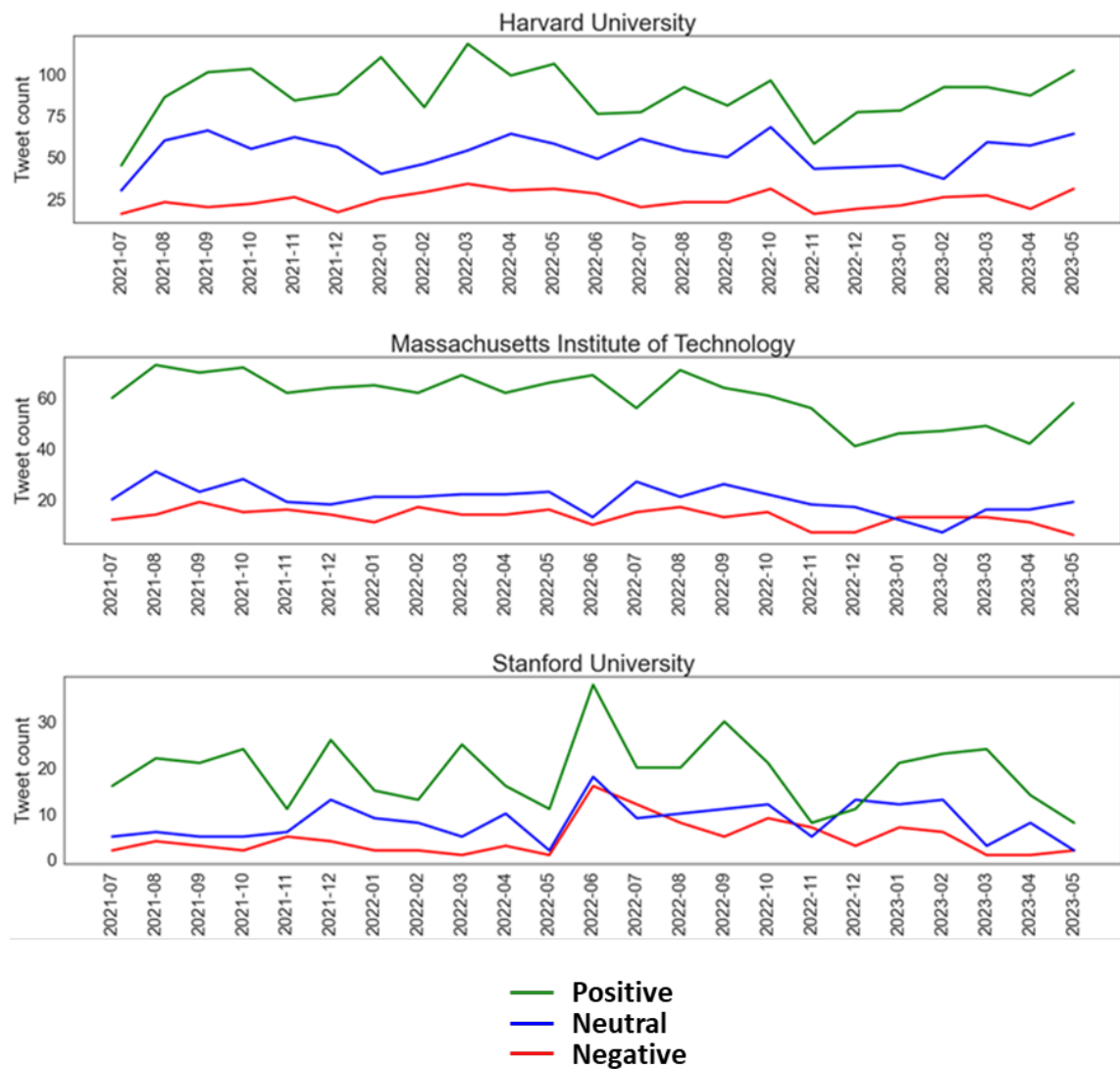


FIGURE 3.16: Monthly evolution of sentiment for Harvard, MIT and Stanford posts. Negative sentiment in red, neutral in blue and positive in green.

This sentiment analysis not only provides us insights into the emotional tone of tweets

but also serves as a valuable tool for evaluating the effectiveness of communication strategies and the public response to the content disseminated by HEIs. Such analyses offer a differentiated understanding of the impact and resonance of posts on social media, contributing to informed decision-making regarding communication and engagement strategies.

The comprehensive set of charts for all Higher Education Institutions (HEIs) is not included in this section. However, they are available for reference in Appendix A. However, as illustrated in Figure 3.17, a consistent pattern emerges throughout the analysis period. The positive sentiment in the tweets dominates, followed by neutral sentiments and lastly, negative sentiments.

It should be noted that, on certain occasions, the count of negative sentiment tweets exceeds that of neutral sentiment tweets, as exemplified by the data point for November 2022 for Stanford University. This particular increase in negative sentiment can be attributed to the contextual factors surrounding the midterm elections and democracy day, both of which appeared prominently featured in tweets during that period. Such fluctuations in sentiment counts can often be attributed to specific events or occurrences during that period, which are reflected in the institution's tweet publications. These changes underline the dynamic nature of sentiment analysis and the potential impact of external factors on an institution's social media sentiment.

An alternative approach to visualizing the sentiment data is depicted in Figure 3.17, where we present the cumulative sentiment frequency on a daily basis. To build this graph, we implemented the following methodology: for each Higher Education Institution (HEI), we initialized a countable variable defined as zero. For each day of the analysis period, we increased the countable variable if the tweets were associated with a positive sentiment, decreased it in the case of a negative sentiment, and left it unchanged if the sentiment was assessed as neutral.

These cumulative sentiment frequency graphs provide a dynamic representation of how sentiment evolves over time for each HEI. It offers a more granular perspective, allowing the observation of daily fluctuations in sentiment and potentially revealing temporal patterns or significant events that influence sentiment trends. Such an approach facilitates a differentiated understanding of the interaction between daily events and the feelings expressed in the discourse of an institution's social network.

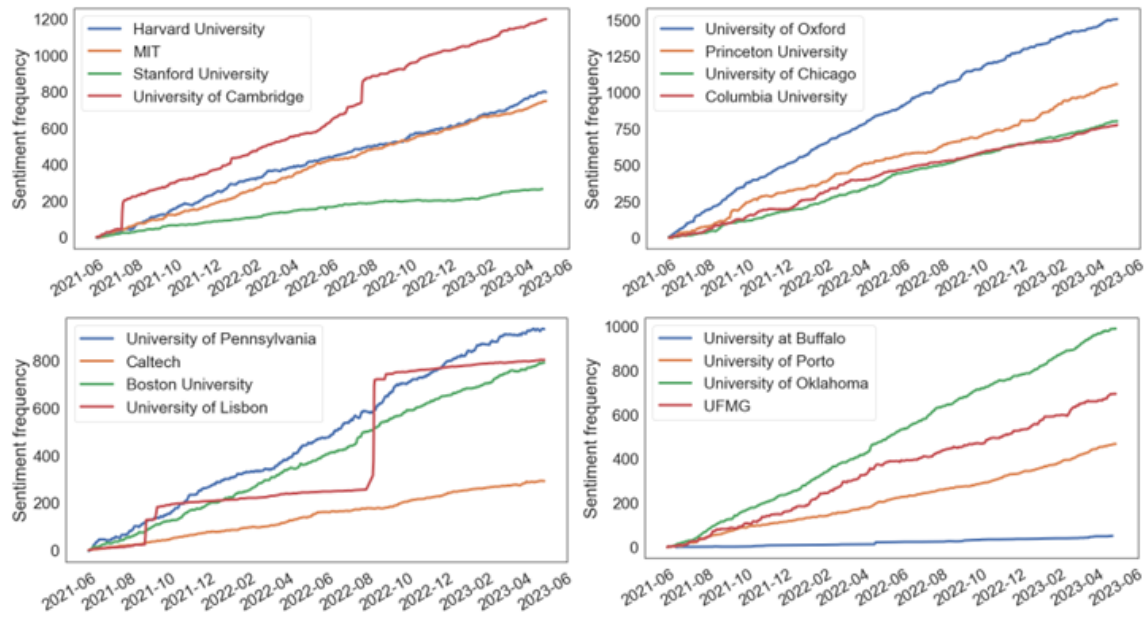


FIGURE 3.17: Daily cumulative sentiment frequency for all HEIs.

Figure 3.17 provides a comprehensive perspective on the cumulative frequency of sentiment for all the universities examined. Notably, a predominant trend emerges in the sentiment of almost all universities' daily tweets as it exhibits a discernible slope towards increasing positively, illustrated by the consistent upward trajectory of the sentiment lines over time.

Furthermore, it is worth highlighting that the growth rates of the sentiment lines for almost all the universities appear to be quite similar, with only subtle deviations in their slopes. In contrast, the University of Lisbon sentiment line demonstrates a relatively unstable pattern, characterized by some inconsistent increase in positive sentiment over time. This distinction can be attributed to the observation that the former institutions tend to have a greater number of positive tweets compared to negative ones, whereas Lisbon's tweet sentiment distribution appears to be more evenly unbalanced between positive and negative sentiments, as depicted in Figure 3.17.

Although our presentation has focused on sentiment graphs for all the universities, the computed sentiment values will serve as a valuable basis for comparative analysis across all Higher Education Institutions (HEIs) in the subsequent section of our study.

3.2.6 Emotion Analysis

To extract insights into the emotional content conveyed by the general tweets of each Higher Education Institution (HEI), we leveraged the capabilities of the NRClex python

library. This library was used to predict sentiments and specific categories of emotional affect present in the text. Emotional affect categories span a spectrum of sentiments and emotions, including anger, anticipation, disgust, fear, joy, negativity, positivity, surprise, sadness, and trust.

In our visual representations, we seek to enhance the accessibility of emotional categories through the attribution of distinct colors. Specifically, we employ a green tone to represent positive emotions, red to denote negative emotions and orange to represent neutral emotions. This color-coded approach makes it easy to visualize the emotional patterns and text distribution in each HEI's tweets.

Figure 3.18 and Figure 3.19 provide a comprehensive representation of the overall proportions of emotion for all the universities, revealing remarkable similarity in their emotional content distribution. Notably, the combined proportions of positive and negative sentiments sum to 1, mirroring the equivalence observed in the combined proportions of trust, joy, anticipation, surprise, anger, fear, sadness, and disgust. This balance means a harmonious emotional spectrum in the tweets analyzed for these universities.

These emotion charts shown in Figure 3.18 and Figure 3.19 provide a comprehensive overview of the emotional landscape and trends in the tweets of all Higher Education Institutions (HEIs), offering detailed insights into the emotional content conveyed through their social media communications. It can be seen that the positive bar varies between 0.7 to 0.95. When examining the spectrum of all HEIs, it is clear that the most prevalent sentiment in the tweets from all these HEIs during our study period is positive rather than negative.

When examining the other group of emotions, which includes trust and joy as positive emotions, anticipation and surprise as neutral emotions, and anger, fear, sadness, and disgust as negative emotions, it becomes evident that positive and neutral emotions have a greater significant presence in the tweets than negative ones.

This observation meets expectations, as these institutions are aware of their reputation and strive to maintain a positive presence on social media. Thus, their goal is to create a positive impact on their community, which can translate into high social media engagement (likes, retweets, and views). This strategy aims to reach as many individuals as possible to maximize the potential return on their social media efforts.

The goal of our analysis was to discern any anomalous or distinct patterns within the emotional content, but our examination did not highlight any deviations from the

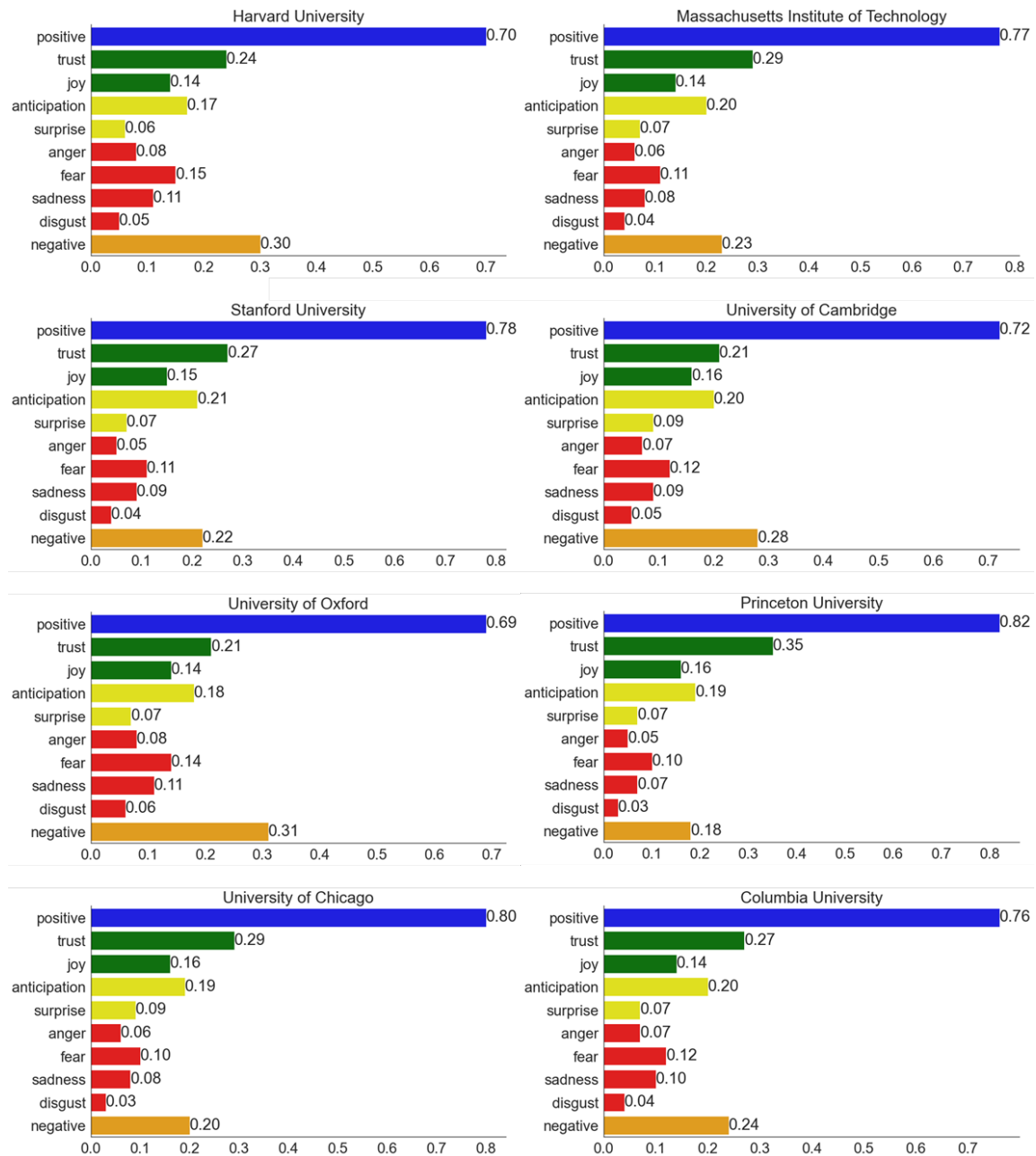


FIGURE 3.18: Proportion of emotions and sentiments found in the overall tweets for all the HEI, first part.

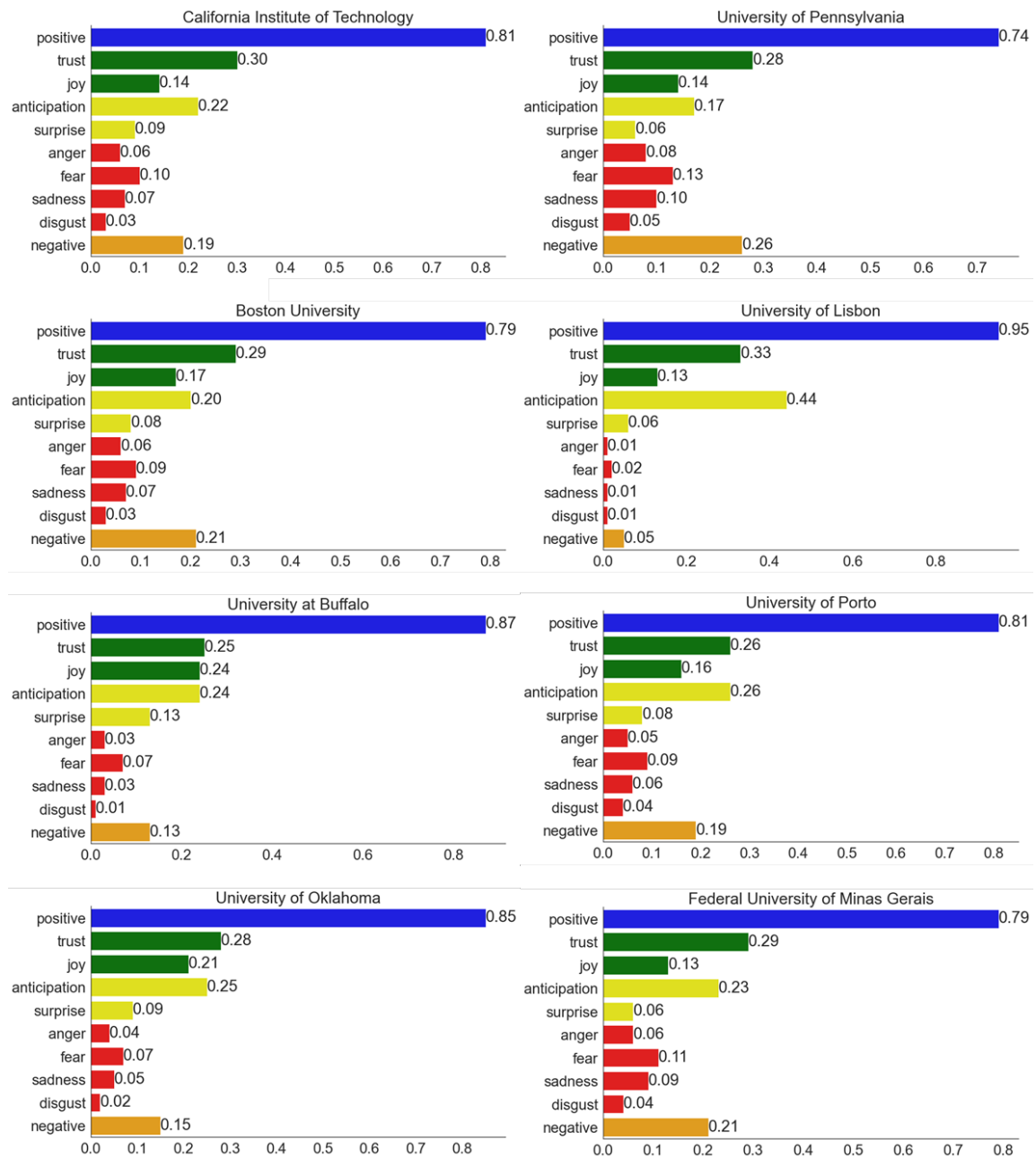


FIGURE 3.19: Proportion of emotions and sentiments found in the overall tweets for all the HEI, second part.

expected emotional distributions. The uniformity in emotional proportions suggests a balanced and consistent expression of emotions in the tweets of all these HEIs'. This finding contributes to a broader understanding of the emotional character of social media discourse within the context of higher education institutions.

3.2.7 Topic Analysis

To delve deeper into the thematic content covered in the general tweets of each Higher Education Institution (HEI) throughout the analysis period, we performed a topic modeling analysis. This analysis was facilitated through the use of the Gensim python library, an important tool for modelling topics, combined with the application of the topic coherence measure [30]. The Topic Coherence Metric played a crucial role in order to identify the most appropriate number of topics inherent to the text corpus, thus increasing the accuracy and interpretability of the topic modeling results.

Before implementing the Latent Dirichlet Allocation (LDA) model for topic modeling [31], a preprocessing procedure was performed [32]. This procedure involved the following steps:

1. **Removal of punctuation and special characters:** all punctuation marks, special characters, and non-alphabetic characters were eliminated from the textual corpus.
2. **Deletion of numbers and links:** numeric values and hyperlinks have been removed from the text to streamline the content for topic modeling.
3. **Capitalization standardization:** capitalization was standardized, ensuring all text was in lowercase to promote uniformity.
4. **Stop words removal:** stop words that have no relevance in the context of topic modeling were removed from the textual corpus. These words typically include common, non-informative terms such as "and," "the," "is," and so on.
5. **Lemmatization:** a lemmatization function was applied to convert each word into its base or dictionary form. This process helps consolidate words with similar meanings.

Employing the coherence score metric as a reference, we optimized the number of topics for each Higher Education Institution (HEI). To achieve this, we executed the Latent

Dirichlet Allocation (LDA) model on a variety of different numbers of topics, finally selecting the one that yielded the highest coherence value. As exemplified in Figure B.2, we present the results for the Federal University of Minas Gerais. It is worth mentioning that this procedure was conducted for all HEIs in the study.

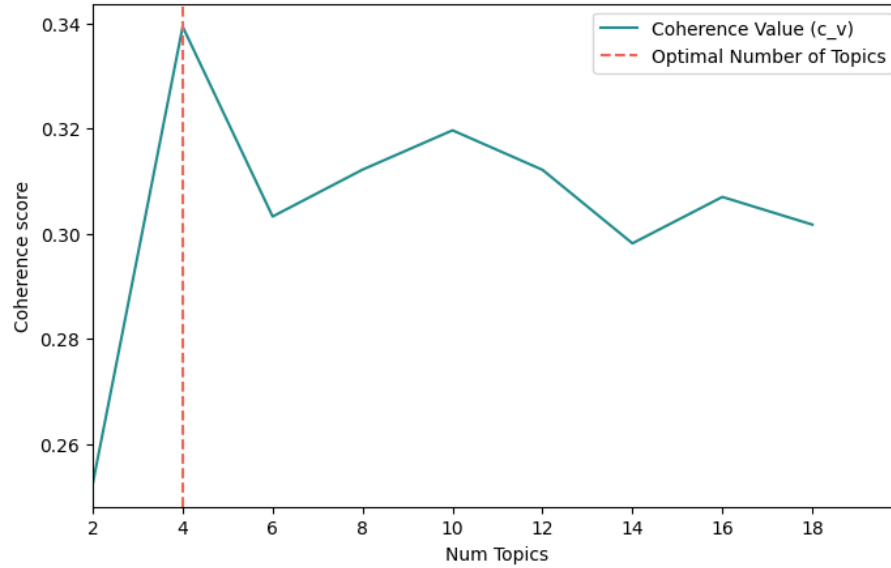


FIGURE 3.20: Federal University of Minas Gerais best number of topics for LDA based on coherence score.

Figure 3.21 illustrates the results obtained for the optimal number of topics using the Python pyLDAvis package. The left side of the figure displays the intertopic distance map, utilizing two principal components. Notably, there is no overlap or intersection observed among the circles, indicating a lack of similarity in the identified topics. In other words, these topics do not share common words. It is worth mentioning that this procedure was also conducted for all HEIs in the study.

Table 3.4 and 3.5 present an overview of the topics identified through the application of the LDA model for each Higher Education Institution (HEI). These topics serve as a valuable framework for understanding the main themes and subject addressed in each institution's tweets, clarifying the predominant discussion areas and content domains of interest.

The results of our topic modeling analysis were succinctly labeled with short identifiers, and these identified topics are provided in Table 3.4 and 3.5. The table provides a comprehensive representation of each identified topic for each Higher Education Institution (HEI). In addition, it includes valuable metrics such as coherence scores for the number of topics identified [33] and Jaccard distances between topics [34].

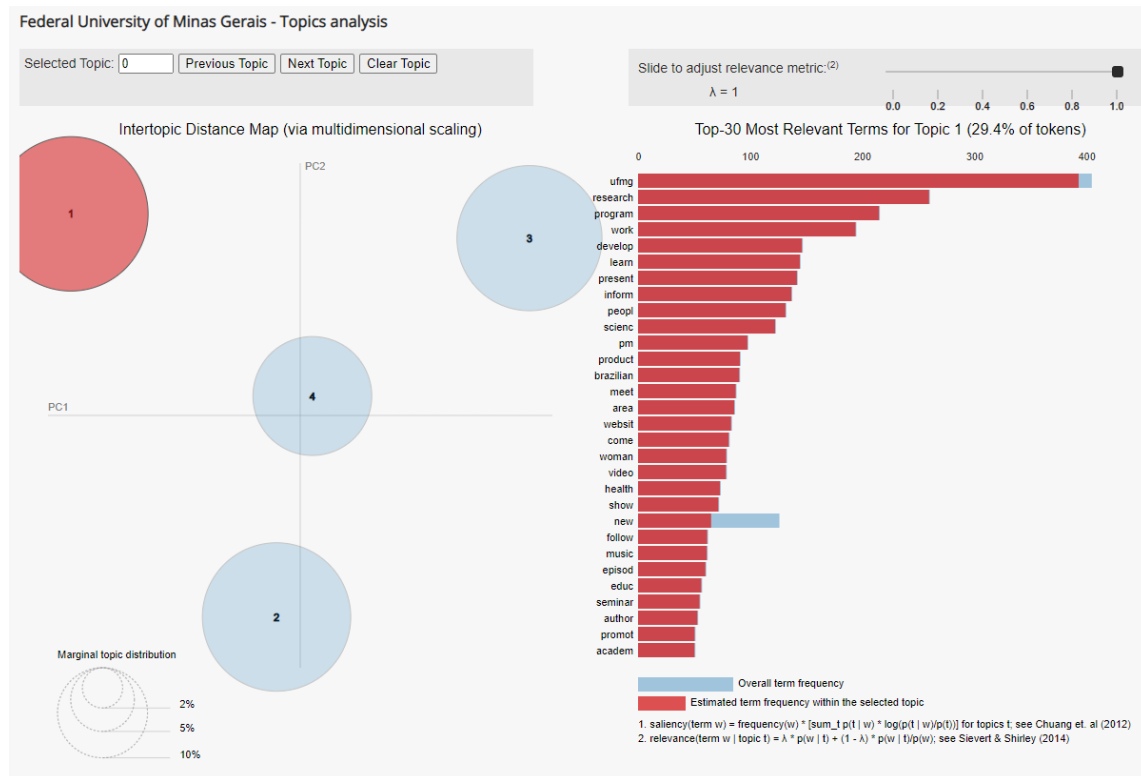


FIGURE 3.21: LDA topic model visualization for Federal University of Minas Gerais.

The coherence score serves as an indicator of topic quality, assessing the interpretability of topics from a human perspective. It gauges how well the topics align with meaningful and coherent themes that can be easily understood.

In Appendix B, you will find a comprehensive compilation of the topic 10 keywords categorized under each topic found for each Higher Education Institution (HEI). It is important to highlight that as we do not have an automated method to label the topics generated by the LDA model, the categorization process was conducted empirically by the author, guided by ChatGPT's assistance.

On the other hand, the Jaccard distance offers insights into the relative proximity or dissimilarity between topics. This metric quantifies the distance between topics, providing a continuous from 0 (indicating closely related topics) to 1 (indicating distant or dissimilar topics). These metrics collectively contribute to a comprehensive assessment of the effectiveness and distinctiveness of the topics discovered through our topic modeling analysis.

TABLE 3.4: Topics found by the Latent Dirichlet Allocation (LDA) model for each HEI.
(A)

Harvard	MIT	Stanford	Cambridge	Oxford	Princeton	Chicago
0.34	0.32	0.37	0.37	0.43	0.35	0.31
0.83	0.69	0.15	0.81	0.64	0.76	0.81
BAY	BAY	BAY			BAY	
R&D	R&D	R&D	R&D	R&D		R&D
			UE	UE		UE
E					E	E
HCM				HCM		HCM
			C			
			T			
					PD	
						DI
			CA			

TABLE 3.5: Topics found by the Latent Dirichlet Allocation (LDA) model for each HEI.
(B)

Columbia	Pennsylvania	Caltech	Boston	Porto	Oklahoma	UFMG
0.26	0.30	0.25	0.34	0.31	0.34	0.33
0.82	0.78	0.15	0.72	0.66	0.74	0.82
	BAY	BAY	BAY	BAY	BAY	BAY
	R&D	R&D				R&D
UE			UE		UE	UE
E	E			E		
C					C	
		T				
AP						AP
	S					

Chapter 4

Discussions and Results

This chapter will delve deeper into the key findings derived from the topic modeling analysis, which was implemented on the tweet data for each Higher Education Institution (HEI). It will also discuss the unsupervised clustering model based on relevant features, created using statistical metrics applied to the historical tweet data, with the objective of evaluating whether the position in the rankings has any influence on the HEIs' publication strategies. Moreover, it will present the results and discoveries yielded by both analyses.

4.1 Topic Modeling

When performing a topic modeling analysis applied to the tweet data retrieved from the higher education institutions that have been studied in this work, this methodological approach revealed a total of twelve distinct topics. For ease of reference and interpretation, we have assigned concise labels to each of these identified topics. The following is a more comprehensive delineation of these topics:

1. Research and Development (R&D)
2. Beginning of the Academic Year (BAY)
3. University Event (UE)
4. Education (E)
5. Health Care and Medicine (HCM)
6. Community (C)

7. Technology (T)
8. Academic Programs (AP)
9. Sports (S)
10. Professional Development (PD)
11. Diversity and Inclusion (DI)
12. Collaboration and Accessibility (CA)

These topics cover a wide range of subjects and recurrent themes that are in the tweets of the Higher Education Institutions (HEIs) under analysis. Notably, the number of topics discovered varied across the institutions, with a maximum of five topics identified for Chicago and Cambridge, while the minimum number of topics identified was two for MIT, Stanford, Boston, and Porto. This diversity in topic distribution highlights the distinct emphases and focuses of content in communications on the social network of each HEI.

After an in-depth analysis of Table 3.4 and 3.5, a noteworthy observation emerged: Higher Education Institutions (HEIs) typically adhere to a comparable publication strategy, characterized by the prevalence of four main topics that are recurrent in a substantial number of institutions. More specifically:

1. **Beginning of the Academic Year** was identified as a recurring topic in 10 of the 16 HEIs.
2. **Research and Development** was highlighted in the tweets of 9 HEIs.
3. **University Event** was a common theme in the social media content of 7 HEIs.
4. **Education** emerged as a substantial topic in the tweets of 5 HEIs.

This consistent thematic alignment highlights the existence of communication priorities and content themes shared between HEIs. The recurrence of these central topics implies that they occupy a significant position in the discourse and strategic communication of these institutions, reflecting their commitment to themes related to the academic calendar, research, university events, and educational initiatives.

In conclusion, the analysis reveals that top-ranked universities tend to adopt specific strategies, particularly in terms of emotions and topics. Conversely, lower-ranked universities employ less prominent strategies that do not align with their ranking positions. The

research sheds light on the role of social media in Higher Education Institutions (HEIs) marketing strategies and its potential influence on their global standings.

4.2 Grouping the Strategies

To conduct a comprehensive and quantitative analysis aimed at comparing the publishing strategies of all Higher Education Institutions (HEIs), we aim to employ a data-driven approach. This approach will allow us to simultaneously evaluate and compare all HEIs using numerical metrics. The ultimate goal is to perform an unsupervised classification, grouping HEIs based on the metrics we derive from our analysis.

It is important to clarify that our analysis will focus exclusively on the publishing patterns within the social media content from these institutions. We will not take into account external factors such as employment statistics, student success rates, or research funding levels. Our main objective is to utilize metrics derived from the analysis of retrieved tweets to cluster the HEIs, facilitating a comparative assessment of their publishing strategies. The resultant groupings will then be evaluated against existing institutional rankings, allowing us to explore potential correlations or distinctions between the two data sets.

4.2.1 The Feature Space Vector Model

To comprehensively summarize the HEIs' publishing behavior and reflect previous analyses, we carefully selected 10 characteristics. These resources serve as a robust representation of each Higher Education Institution's (HEI) communication strategies and social media engagement patterns. The chosen features are as follows:

1. **Mean Daily Posting Frequency:** this metric means the average number of tweets posted daily by each HEI, providing insights into their regularity of communication.
2. **Maximum Daily Posting Frequency:** represents the largest number of tweets posted by the HEI in a single day, highlighting cases of intensive communication.
3. **Weekend Posting Ratio:** this ratio illustrates the proportion of tweets posted during weekends (Saturday and Sunday), providing insights into weekend engagement practices.

4. **Night Posting Ratio:** quantifies the proportion of tweets posted at night (from 9:00 PM to 7:00 AM), indicating HEI's engagement outside business hours.
5. **Mean Positive Sentiment:** this feature represents the average positivity of sentiment expressed in HEI's tweets, reflecting the emotional tone of its communications.
6. **Mean Neutral Sentiment:** reflects the average neutrality of sentiment in HEI's tweets, meaning the balance between positive and negative sentiments present in a specific tweet.
7. **Mean Negative Sentiment:** this metric denotes the average negativity of sentiment in HEI's tweets, offering insights into the presence of negative sentiment.
8. **Mean Tweet Length (Text):** quantifies the average length of HEI's tweets in terms of text characters, highlighting typical tweet length.
9. **Total Tweet Length (Text):** this feature measures the cumulative length of all HEI's tweets in terms of text characters, indicating the overall text volume.
10. **Total Number of Links Used in Text:** represents the total count of hyperlinks included in the HEI's tweets, illustrating the use of references and external resources.

These 10 features collectively form a comprehensive signature of each HEI's posting behavior, covering various aspects such as posting frequency, sentiment, engagement patterns, and content characteristics. This signature allows for a quantitative and nuanced comparison of HEIs' communication strategies and social media engagement.

4.2.2 Clustering the HEIs

Each Higher Education Institution (HEI) is represented as a vector within a 10-dimensional vector space model. The purpose of this representation is to calculate the inter-cluster distances, thus facilitating the identification of HEIs with close spatial proximity. This investigation employs the established standard k-means algorithm to cluster HEIs based on the metric characteristics mentioned above.

In the course of our experimentation, we employ the generic k-means approach [35], and we conducted our analyses using the Floyd algorithm [36] and the Hartigan-Wong algorithm [37]. It is worth noting that the results derived from these different algorithmic methodologies exhibited a notable degree of similarity, highlighting the consistency and robustness of the clustering results.

In our quest of minimizing inter-cluster distances, we performed an exploratory analysis involving multiple cluster counts, using the “elbow method” as a basis of comparison that is represented in Figure 4.1. Ultimately, it was decided to adopt a configuration of three clusters for the purpose of categorizing the Higher Education Institutions (HEIs).

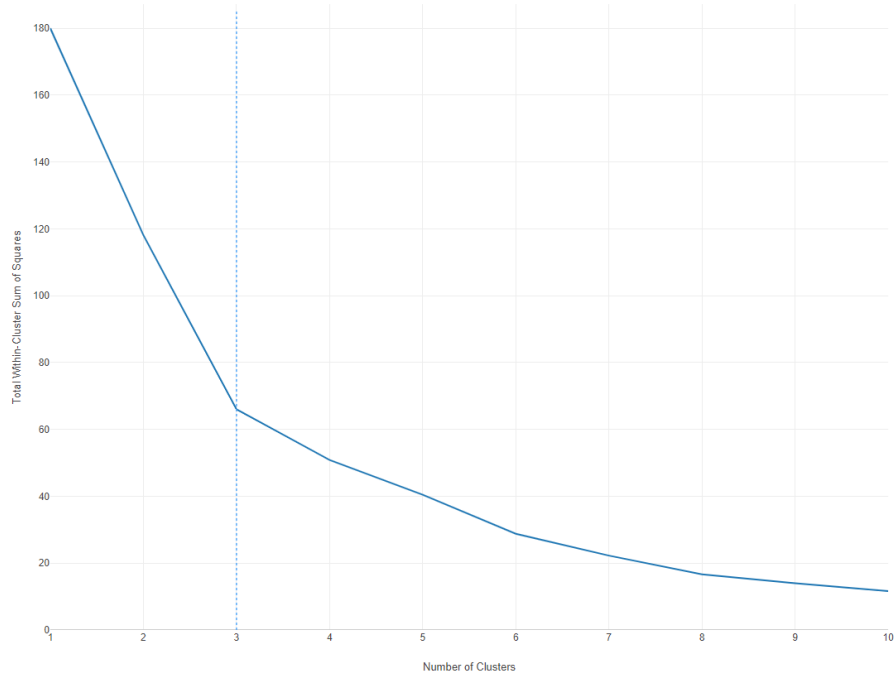


FIGURE 4.1: Elbow method results.

To provide a visual representation of the resulting clustering results, we present a mapping of each HEI, with colors indicating their respective cluster assignments, as illustrated in Figure 4.2. For this visualization, we apply a Principal Component Analysis (PCA) transformation [1] to reduce the dimensionality of the original 10-dimensional data points to two dimensions. It is important to note that the information presented in this visualization refers to data collected for the HEIs’ Twitter publications during the period from July 2021 to May 2023.

The logic underlying the adopted clustering scheme is substantiated by the conspicuous delineation of three distinct groupings: HEIs represented in blue, located mostly in the second quadrant and allocated to cluster 1; HEIs represented in orange, mainly occupying the first quadrant and assigned to cluster 2; and HEIs marked in green, predominantly clustered around the demarcation between the third and fourth quadrants, constituting cluster 3.

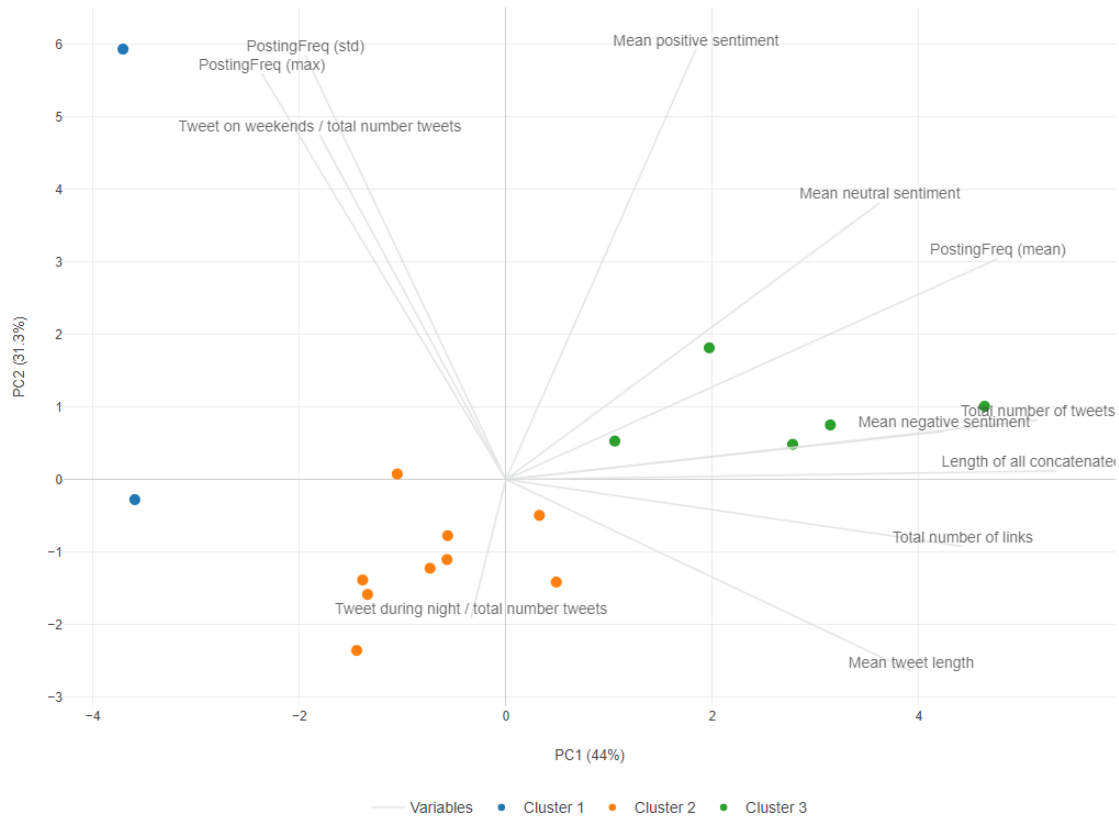


FIGURE 4.2: Positioning and clustering for all the HEIs in a 2D projection of the feature space.

To complement our analytical effort, we conducted an examination of the distribution patterns of the normalized values of the ten features within each cluster. This investigation is depicted in Figure 12, located below for reference, where we use box-plots as a means to conveying the relevant statistical information.

In Figure 4.3, the comprehensive box-plot presentation covers the results for all three clusters. Within this visual representation, our observations suggest the plausible presence of a cohesive strategic alignment among the Higher Education Institutions (HEIs) encapsulated within these respective clusters.

Notably, the data reveals that clusters 2 and 3 show a pronounced concentration of variables, highlighting a significant convergence of attributes. On the other hand, the variables in cluster 3 display a more dispersed distribution. Specifically, cluster 3 predominantly demonstrates positive values, whereas cluster 2 predominantly depicts negative values. Cluster 1 presents a mixed composition of both positive and negative attributes. This observation highlights the effectiveness of the clusterization model, which effectively

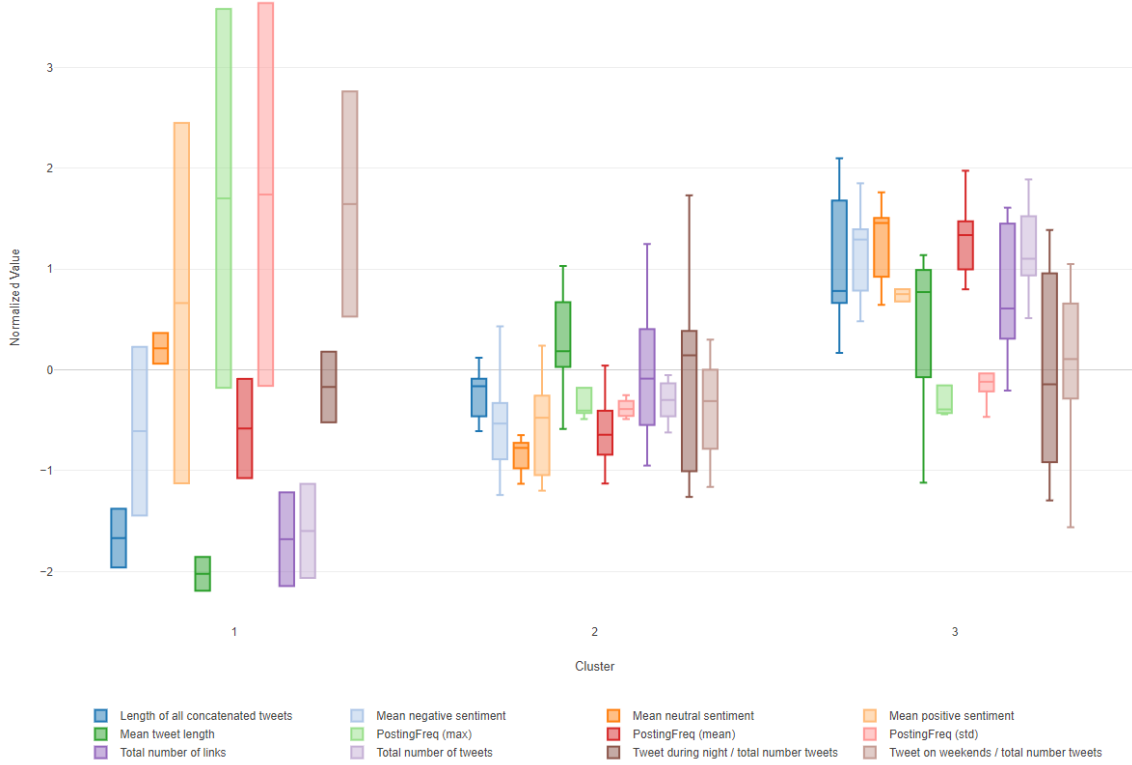


FIGURE 4.3: Distribution for all the features separated by cluster.

segregated institutions into distinct clusters based on the features employed by the unsupervised model.

It is imperative to emphasize that before carrying out this analysis, a rigorous standardization procedure was carefully applied to all data. This standardization process ensured data uniformity and facilitated meaningful comparisons throughout the analysis.

4.2.3 Analysis of the Clustering Results

To facilitate a comparative analysis of these results with the established rankings, we have structured Table 4 accordingly. In this table, we have incorporated the cluster assignments as the second column, in conjunction with the four ranking lists. The table has been carefully organized based on the “cluster” column, thus aggregating Higher Education Institutions (HEIs) belonging to the same cluster for ease of reference and evaluation.

The clustering results manifest distinct patterns between Higher Education Institutions (HEIs) in different clusters. Notably, cluster 3 HEIs, with the exception of the University of Minas Gerais, consistently occupy top-tier positions within the ranking lists.

TABLE 4.1: Cluster assignment together with the rankings scores of 2022.

Higher Education Institution	Cluster	CWUR	Shanghai	USNews	QS
University at Buffalo	1	300	301-400	280	338
University of Lisbon	1	200	201-300	197	356
Boston University	2	99	101-150	65	112
Caltech	2	10	9	9	6
Columbia University	2	8	8	6	19
MIT	2	2	3	2	1
Stanford University	2	3	2	3	3
University of Cambridge	2	4	4	8	3
University of Chicago	2	7	10	15	10
University of Oklahoma	2	400	501-600	425	651-700
University of Porto	2	308	201-300	255	295
UFMG	3	500	401-500	256	651-700
Harvard University	3	1	1	1	5
Princeton University	3	6	6	16	20
University of Oxford	3	5	7	5	2
University of Pennsylvania	3	9	15	13	13

Conversely, the HEIs in cluster 2 exhibit a more diversified distribution, spanning a wide spectrum of positions in the rankings.

Cluster 1 comprises the University of Lisbon and University of Buffalo, located in the intermediate position in the ranking 200th and 300th in CWUR, respectively. A comparative examination between the University of Lisbon and the University of Buffalo reveals divergent publishing strategies. Buffalo strategy appears to be evolving towards a more consolidated approach, while Lisbon approach appears to be experiencing greater volatility with less distinct objectives.

In summary, it is noticeable that Higher Education Institutions (HEIs) affiliated with clusters 2 and 3 consistently presents high position in the ranking, even though there is a presence of HEIs occupying intermediary positions within these clusters. On the other hand, HEIs affiliated with cluster 1 present a notable challenge in terms of facile characterization based on the attributes used for the clustering process, as exemplified by the radar graph represented in Figure 4.4. This observation underscores the high complexity inherent in the profiles of HEIs covered within these clusters, since all three cases present a clear demarcation of institutions based on the features used.

To facilitate a comparison of these results with the rankings, we have prepared Table 4.1. This table incorporates the cluster assignments in the second column, together with the four ranking lists. To increase clarity and ease of reference, the table has been

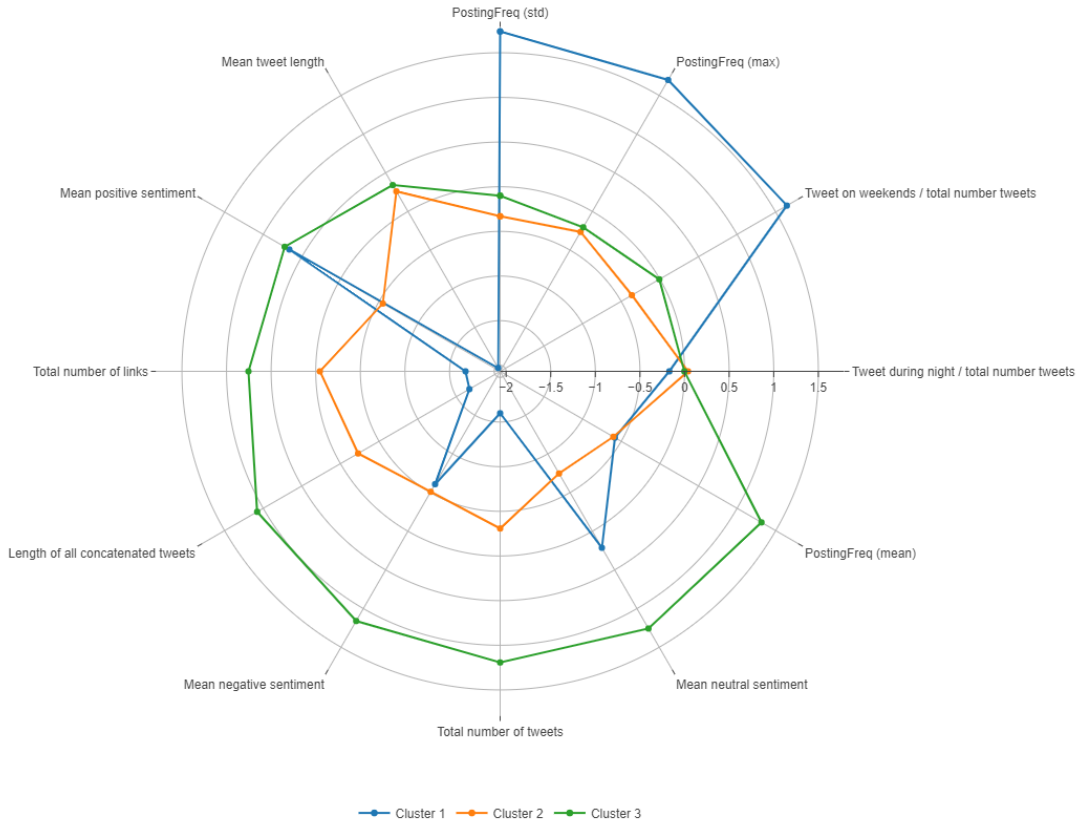


FIGURE 4.4: Distribution of each variable in each cluster highlighting the major changes.

organized based on the "cluster" column, effectively grouping HEIs that share the same cluster designation.

The results of the cluster analysis provides notable insights into the publication strategies adopted by Higher Education Institutions (HEIs). In particular, Cluster 1 encompasses both the University of Lisbon and the University at Buffalo, which share a middle-ranking position. This clustering suggests a potential similarity in their publication strategies, as inferred from the examined features.

Cluster 3 is mainly consists of HEIs that achieve top-10 positions in the ranking, with the University of Minas Gerais being a noteworthy exception. On the other hand, cluster 2 encompasses HEIs cover a wide range of ranking positions, making them difficult to categorize directly based on the analyzed features in this study.

Interestingly, institutions outside the top 10 ranking positions are distributed across all three clusters, suggesting that publication strategies may not be strictly linked to ranking positions. Therefore, the clusters appear well defined, revealing distinct patterns, as represented in the radar graph in Figure 4.4.

In summary, Cluster 1 consists mainly of HEIs occupying intermediate ranking positions, Cluster 3 tends to secure higher rankings, while Cluster 2 HEIs defy simple characterization based on the features examined in this study.

Chapter 5

Conclusions

5.1 Synthesis

In this study, we embarked on an exploration of the complex relationship between Higher Education Institutions (HEIs) and their publication strategies based on their ranking position (CWUR), with a special focus on how these strategies can be influenced in relation to the rankings of these institutions.

To achieve this, we utilized Natural Language Processing (NLP) techniques on the historical Twitter data of the HEIs. This included sentiment analysis, emotion analysis, and topic analysis. Moreover, a clustering segmentation was executed to identify institutions with similar publication strategies and to gain valuable insights from the segmentation.

5.2 Contributions

5.2.1 Objectives

The data was collected using a collector built in-house, which interfaced with the Twitter API using Python as a framework. We utilized an extensive dataset derived from [16], focusing on the common period data spanning from July 2022 to May 2023. This data was subjected to a comprehensive exploratory data analysis, encompassing aspects such as community engagement, publication frequency, temporal patterns, and content characteristics.

The outcomes of our analysis clarify several important conclusions:

- Firstly, we discern a notable correlation between HEIs that occupy top positions in the rankings and the consolidation of their publication strategies, with the exception

of the Federal University of Minas Gerais (500th position in the ranking) which is fully aligned with the publication strategy of the best classified.

- Additionally, our observations revealed that higher-ranked HEIs tend to exhibit more pronounced sentiments, engage in longer tweets, and maintain higher posting frequencies.
- Besides that, we performed sentiment, emotion, and topic analysis of the tweet content for each HEI. Our comprehensive analysis revealed that the top-ranked HEIs share surprisingly similar publication strategies, characterized by the prevalence of four primary topics, Beginning of the Academic Year, Research and Development, University Event and Education, that cover the majority of these analyzed institutions.
- In conclusion, the clustering analysis revealed a noteworthy observation that institutions positioned beyond the top 10 ranking positions are evenly dispersed among all three clusters. This observation suggests that there may not be a strict correlation between publication strategies and ranking positions. As a result, the clusters exhibit well-defined boundaries, exposing discernible and distinct patterns.

Furthermore, all the research objectives outlined in this study were successfully achieved, leading to the following conclusions:

1. The CWUR ranking was selected as the main reference point for our study, as it demonstrated that there are no significant differences between the most used existing rankings, as proven by the support of the results from the Kendall τ and RBO metrics.
2. Through exploratory data analysis, coupled with sentiment and emotion analysis, a wealth of insights pertaining to the publication characteristics of HEIs was gleaned.
3. The use of topic analysis in conjunction with the clustering approach played a key role in defining categories and delineating clear boundaries, which in turn served as a guiding framework for classifying of HEIs based on their strategies.
4. The discussion of the results was based on the topic and clustering analyses, providing solid bases for our conclusions regarding the categories of HEIs and their relationship with the positions ranking. Furthermore, these analyses, when compared, offer compelling evidence in support of our findings.

5.2.2 Preliminary Publications

During the course of this dissertation's development, one paper and one book chapter were published. These publications utilized the ongoing content generated from the progressive work conducted in this applied research. This is a testament to the significant contributions and advancements made in this field through this research.

Initially, a paper was submitted to the WEBIST-2022 conference in the area of social network analytics, as cited by Figueira and Nascimento [15]. This article included the initial findings and discoveries from the exploratory data analysis and the first version of the clustering approach, which used data collected until July 2022. The article was accepted and presented at this conference in October 2022. It generated our first publication:

- Figueira, A., & Nascimento, L. V. (2022). Do Top Higher Education Institutions' Social Media Communication Differ Depending on Their Rank?. *Technology*, 2(3), 2.

The subsequent publication, as cited by Figueira and Nascimento [16], served as an expansion of the preceding work. This extension was prompted by an invitation to contribute to a book in the Lecture Notes in Business Information Processing (LNBIP) Series, published by Springer. In order to augment our original work, we utilized Twitter data collected until February 2023. We then conducted an analysis of sentiment, emotion, and topic within this data set:

- Figueira, A., & Nascimento, L. (2022, October). An NLP Approach to Understand the Top Ranked Higher Education Institutions' Social Media Communication Strategy. In *International Conference on Web Information Systems and Technologies* (pp. 166-185). Cham: Springer Nature Switzerland.

5.3 Future Work

For future research directions, it would be beneficial to construct a predictive model that utilizes a Neural Network to be able to capture more patterns in the dataset. This model could integrate the gathered retrieval information to anticipate emotions and topics in upcoming posts from each Higher Education Institution (HEI). Furthermore, incorporating the capability to analyze images and videos contained within these posts would significantly enhance the comprehensiveness of the analysis. This approach would provide

a more holistic understanding of the data, thereby contributing valuable insights to the field.

Appendix A

Sentiment Evolution

In this section, we will illustrate the monthly evolution of sentiment posts across all Higher Education Institutions. Negative sentiment will be depicted with a red line, neutral sentiment with a blue line, and positive sentiment with a green line.

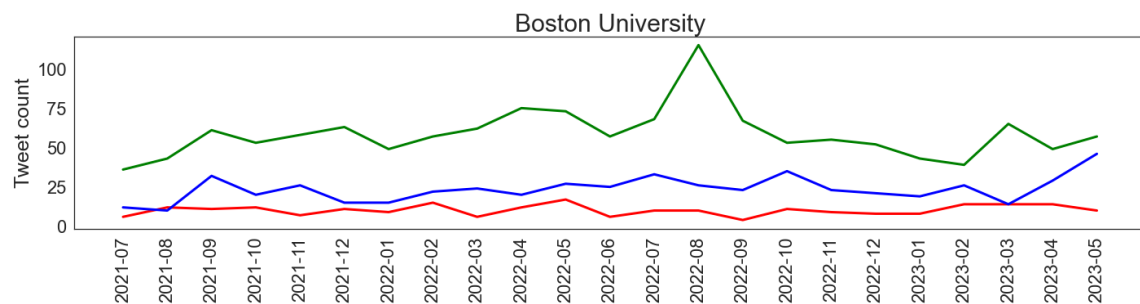


FIGURE A.1: Monthly evolution of sentiment for Boston University posts.

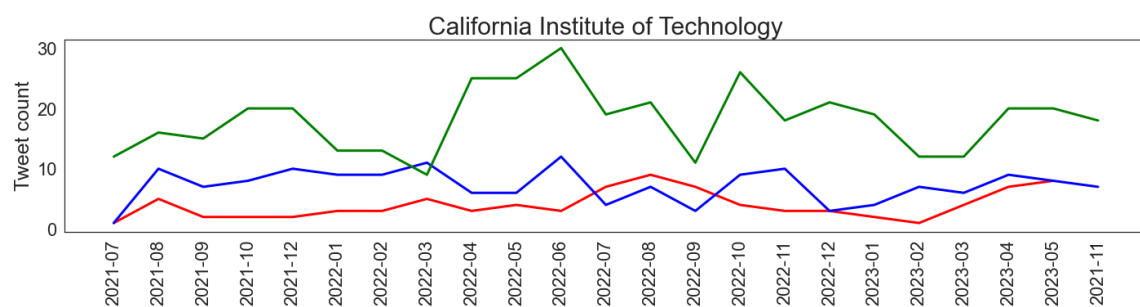


FIGURE A.2: Monthly evolution of sentiment for California Institute of Technology posts.

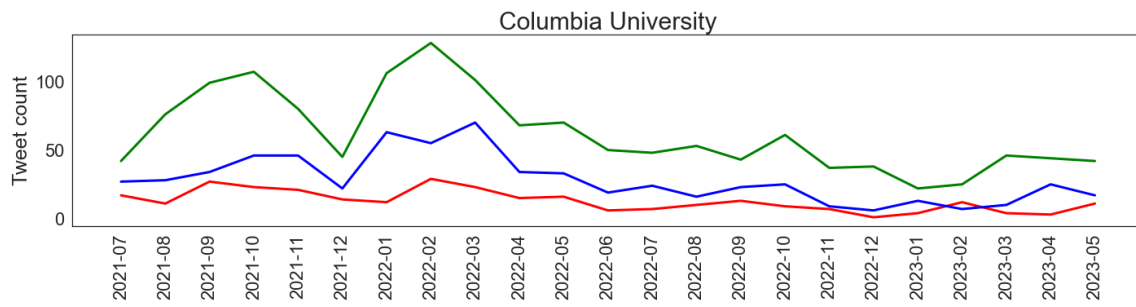


FIGURE A.3: Monthly evolution of sentiment for Columbia University posts.

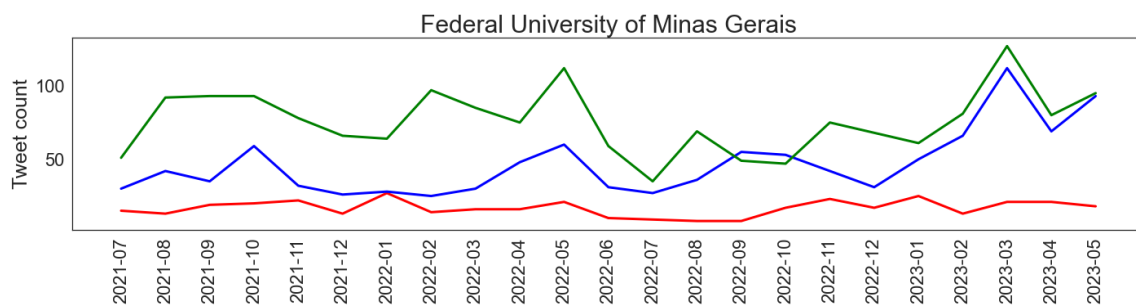


FIGURE A.4: Monthly evolution of sentiment for Federal University of Minas Gerais posts.

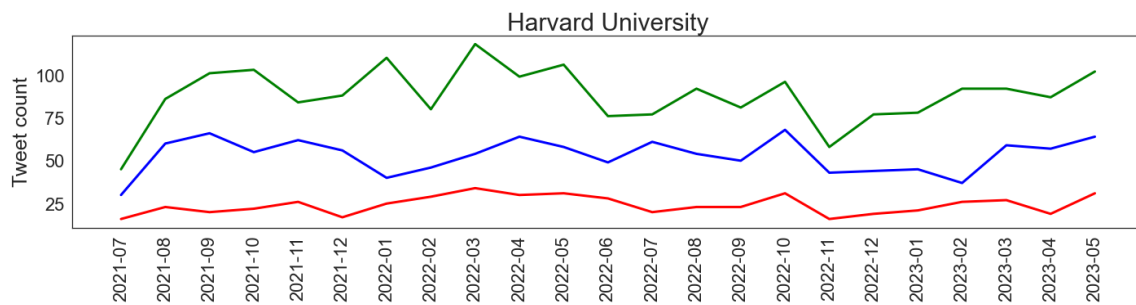


FIGURE A.5: Monthly evolution of sentiment for Harvard University posts.

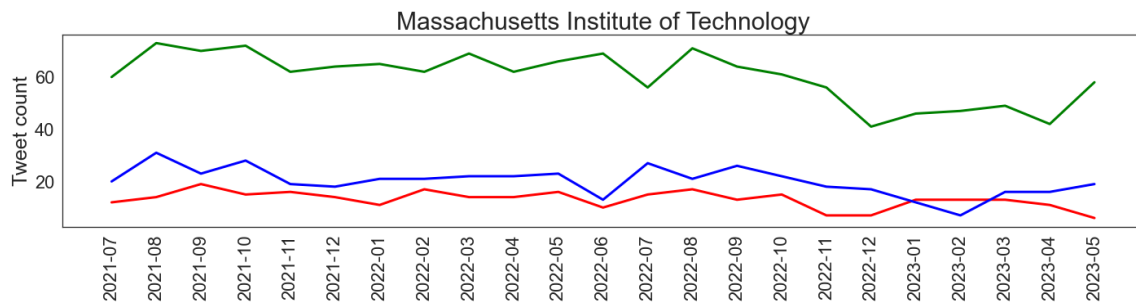


FIGURE A.6: Monthly evolution of sentiment for Massachusetts Institute of Technology posts.

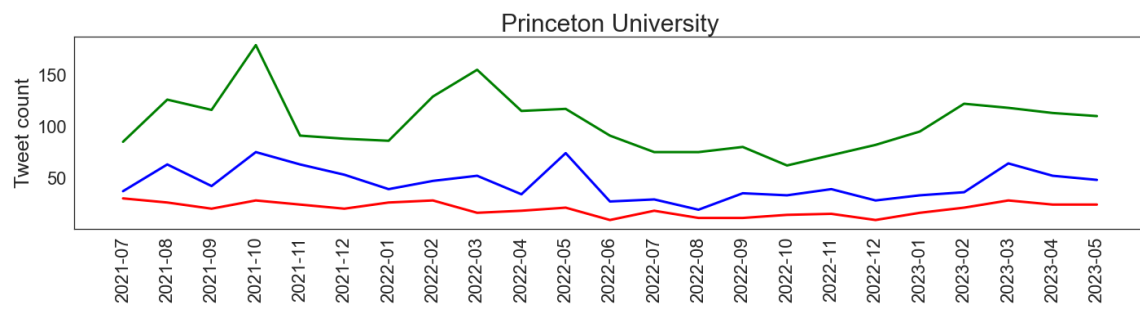


FIGURE A.7: Monthly evolution of sentiment for Princeton University posts.

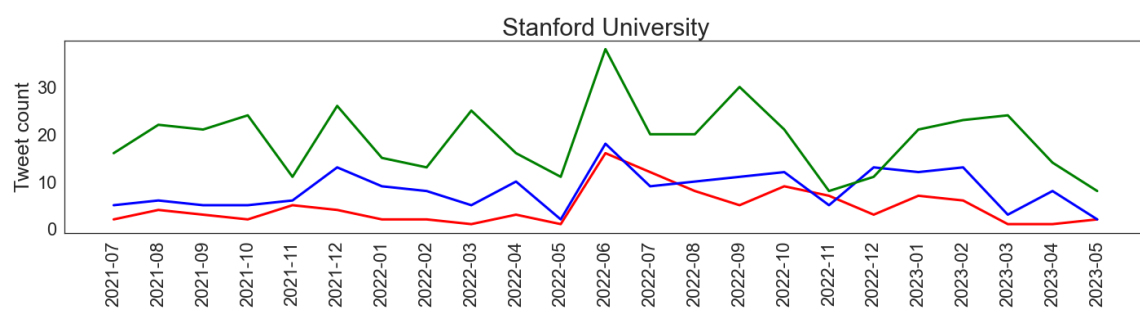


FIGURE A.8: Monthly evolution of sentiment for Stanford University posts.

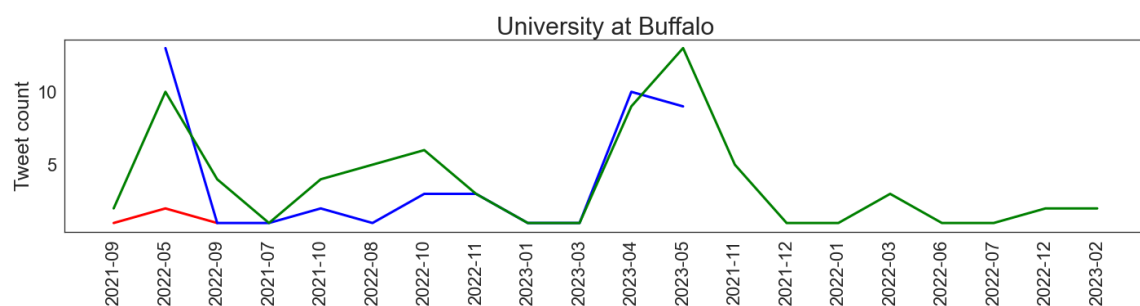


FIGURE A.9: Monthly evolution of sentiment for University at Buffalo posts.

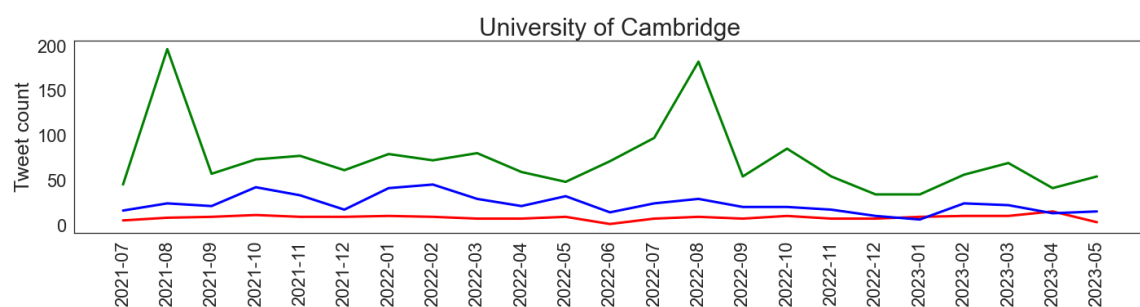


FIGURE A.10: Monthly evolution of sentiment for University of Cambridge posts.

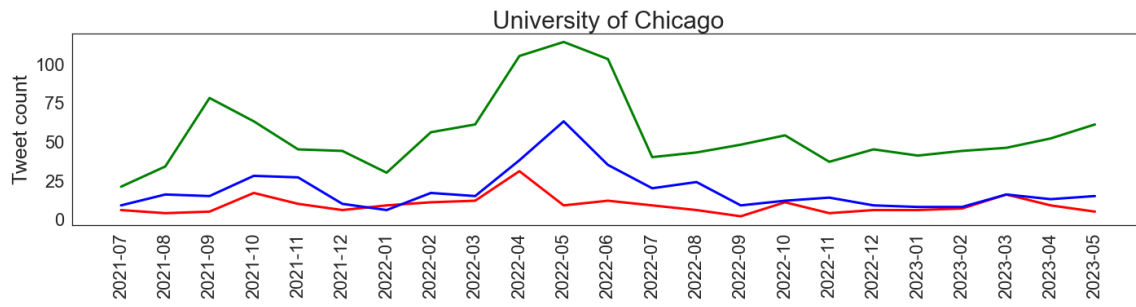


FIGURE A.11: Monthly evolution of sentiment for University of Chicago posts.

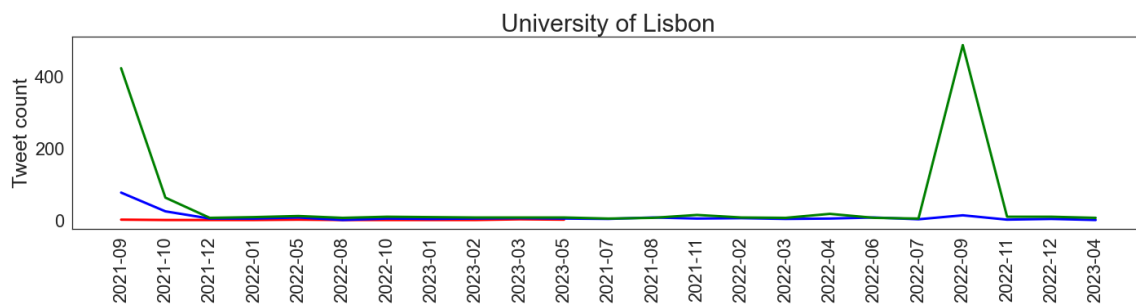


FIGURE A.12: Monthly evolution of sentiment for University of Lisbon posts.

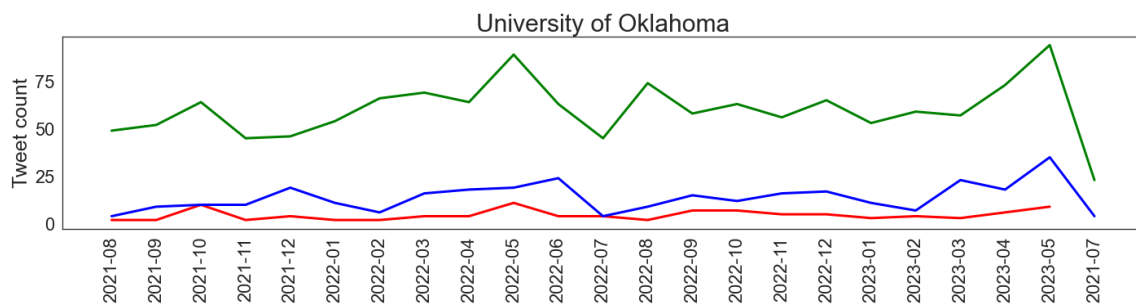


FIGURE A.13: Monthly evolution of sentiment for University of Oklahoma posts.

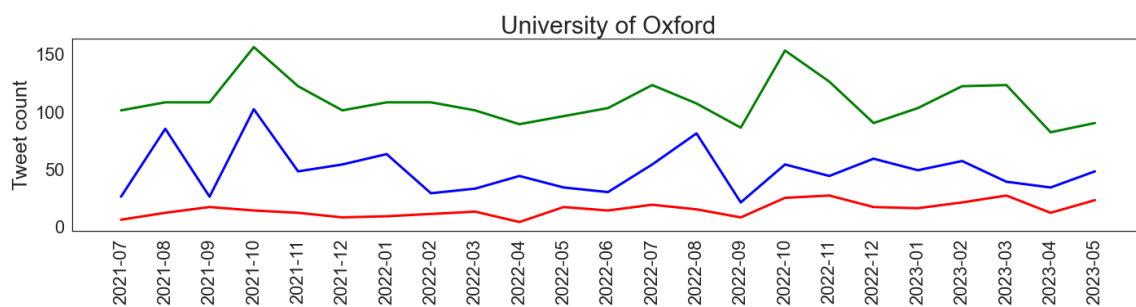


FIGURE A.14: Monthly evolution of sentiment for University of Oxford posts.

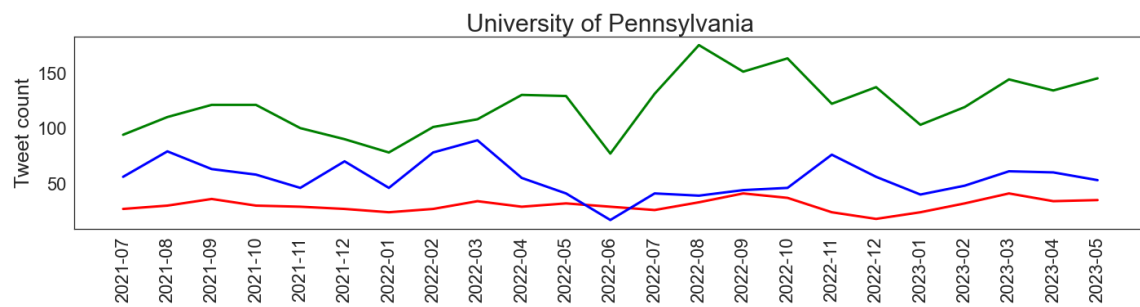


FIGURE A.15: Monthly evolution of sentiment for University of Pennsylvania posts.

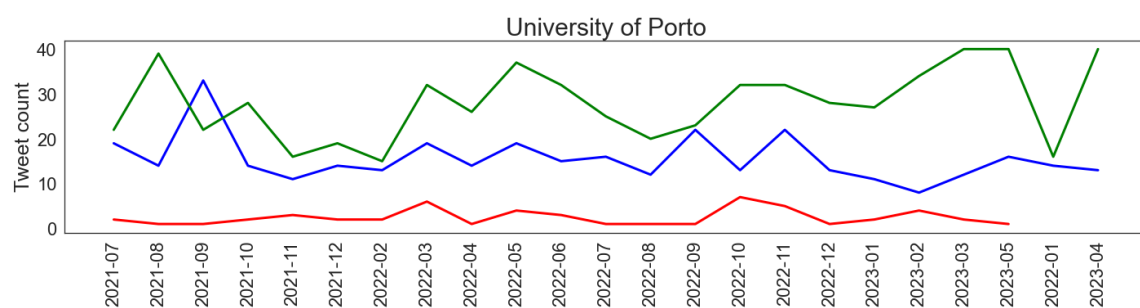


FIGURE A.16: Monthly evolution of sentiment for University of Porto posts.

Appendix B

HEIs Topics Keywords

Discrimination

Within this section, you will visualize the 10 keywords associated with each topic across all Higher Education Institutions, as determined by the LDA model. It is essential to keep in mind that the number of topics was optimized as elucidated in Chapter 3. As there is no automated method for topic labeling, the process was carried out empirically, with guidance from ChatGPT. You can view the label for each topic in the respective figure legend for each Higher Education Institution (HEI).

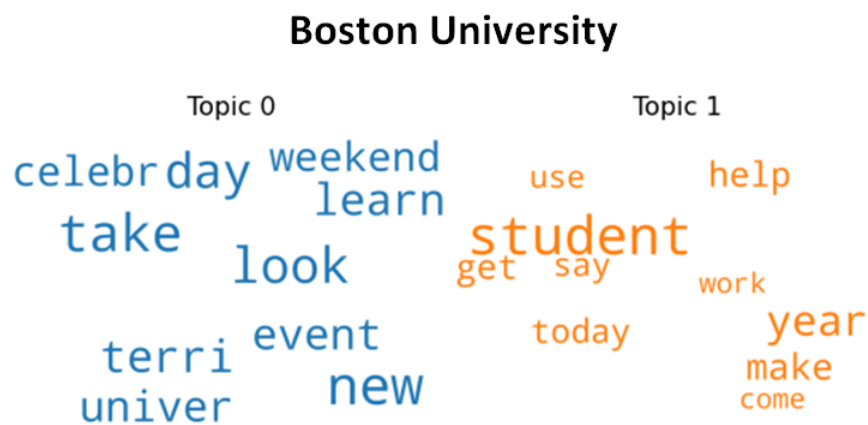


FIGURE B.1: Boston University topics keywords. Topic 0: University Event and Topic 1: Beginning of the Academic Year.

California Institute of Technology

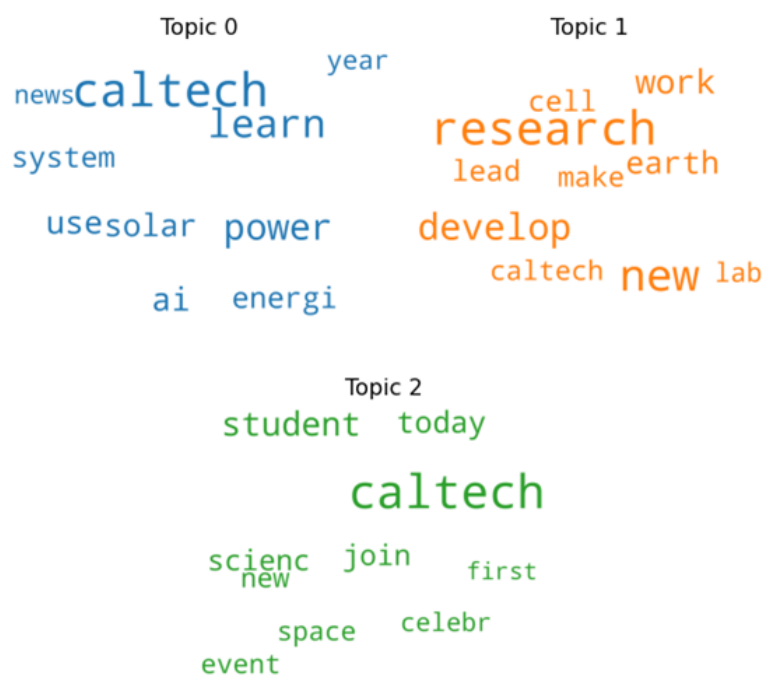


FIGURE B.2: California Institute of Technology topics keywords. Topic 0: Technology, Topic 1: Research and Development and Topic 2: Beginning of the Academic Year.

Columbia University



FIGURE B.3: Columbia University topics keywords. Topic 0: Academic Programs, Topic 1: Education, Topic 2: Community and Topic 3: University Event.

Federal University of Minas Gerais



FIGURE B.4: Federal University of Minas Gerais topics keywords. Topic 0: University Event, Topic 1: Research and Development, Topic 2: Beginning of the Academic Year and Topic 3: Academic Programs.

Harvard University



FIGURE B.5: Harvard University topics keywords. Topic 0: Health Care and Medicine, Topic 1: Education, Topic 2: Beginning of the Academic Year and Topic 3: Research and Development.

Massachusetts Institute of Technology

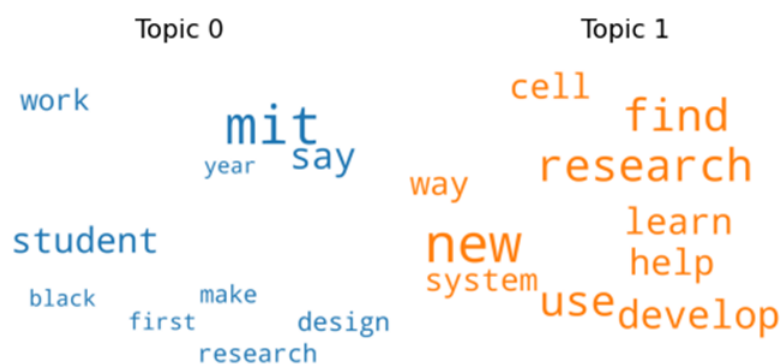


FIGURE B.6: Massachusetts Institute of Technology topics keywords. Topic 0: Beginning of the Academic Year and Topic 1: Research and Development.

Princeton University

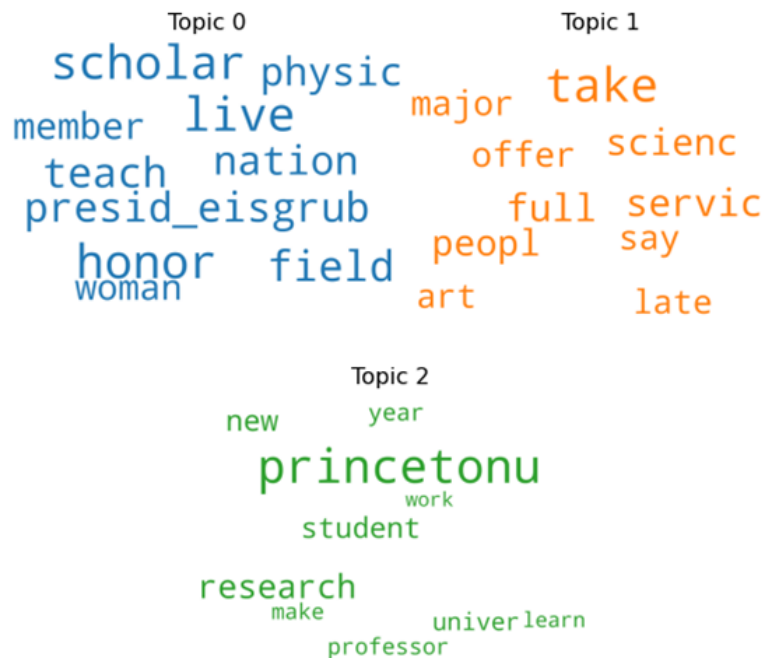


FIGURE B.7: Princeton University topics keywords. Topic 0: Education, Topic 1: Professional Development and Topic 2: Beginning of the Academic Year.

Stanford University



FIGURE B.8: Stanford University topics keywords. Topic 0: Beginning of the Academic Year and Topic 1: Research and Development.

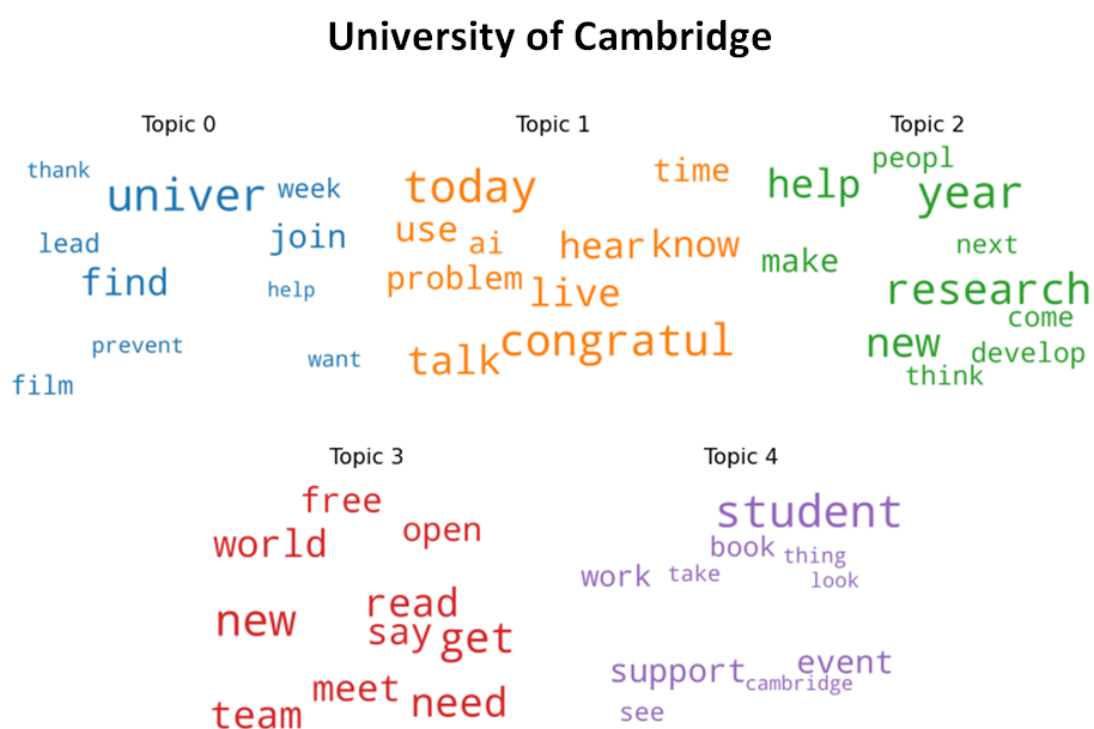


FIGURE B.9: University of Cambridge topics keywords. Topic 0: Community, Topic 1: Technology, Topic 2: Research and Development, Topic 3: Collaboration and Accessibility and Topic 4: University Event.

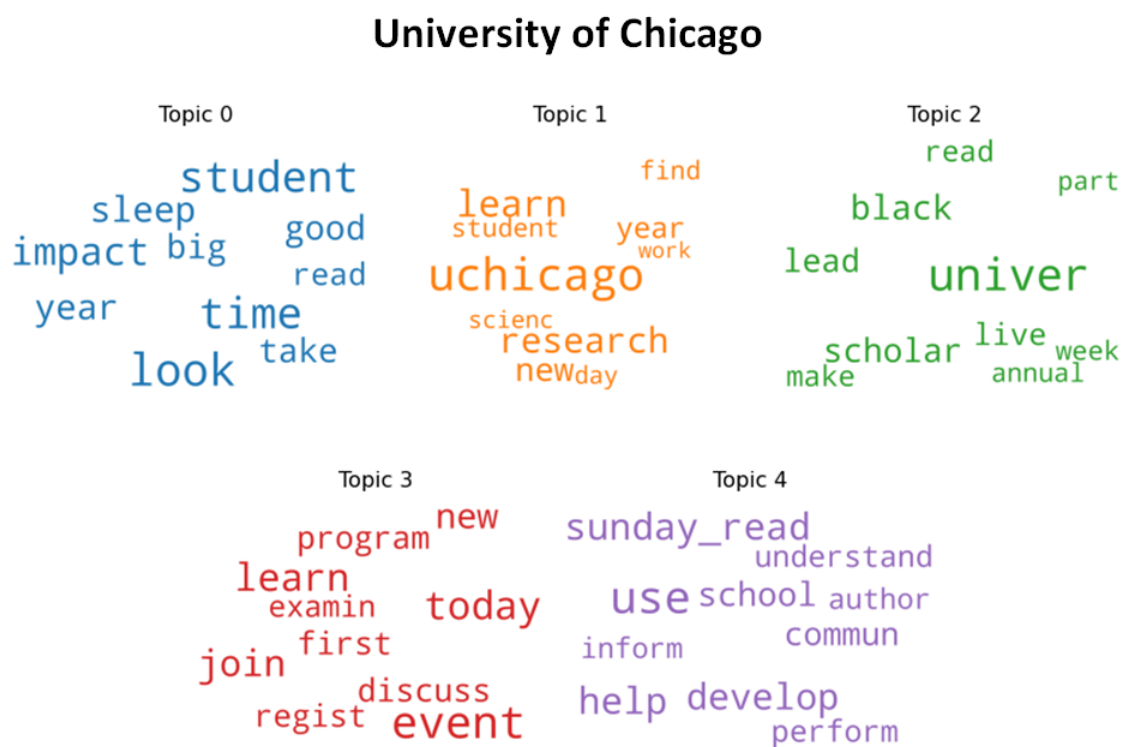


FIGURE B.10: University of Chicago topics keywords. Topic 0: Health Care and Medicine, Topic 1: Education, Topic 2: Diversity and Inclusion, Topic 3: University Event and Topic 4: Research and Development.

University of Oklahoma

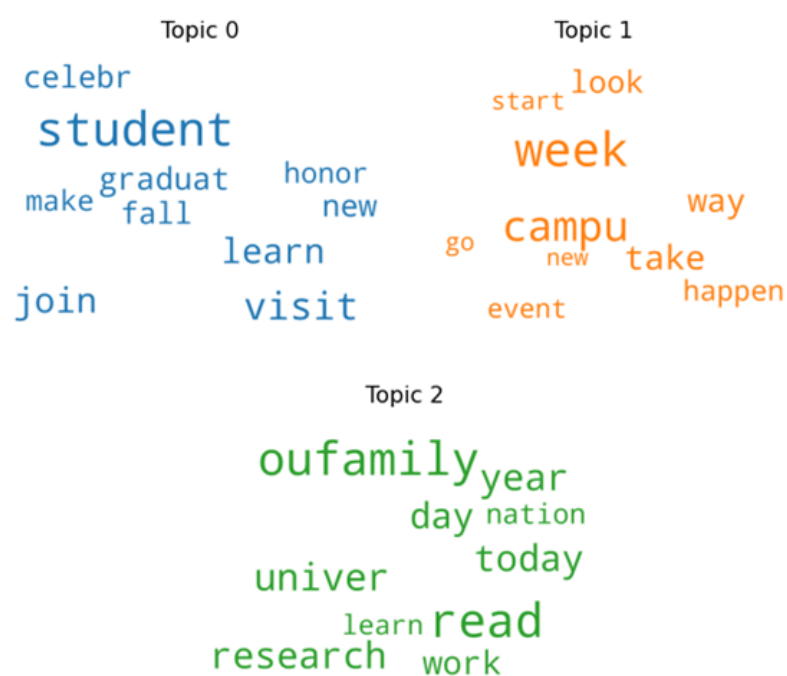


FIGURE B.11: University of Oklahoma topics keywords. Topic 0: Beginning of the Academic Year, Topic 1: University Event and Topic 2: Community.

University of Oxford



FIGURE B.12: University of Oxford topics keywords. Topic 0: Health Care and Medicine, Topic 1: Research and Development and Topic 2: University Event.

University of Pennsylvania

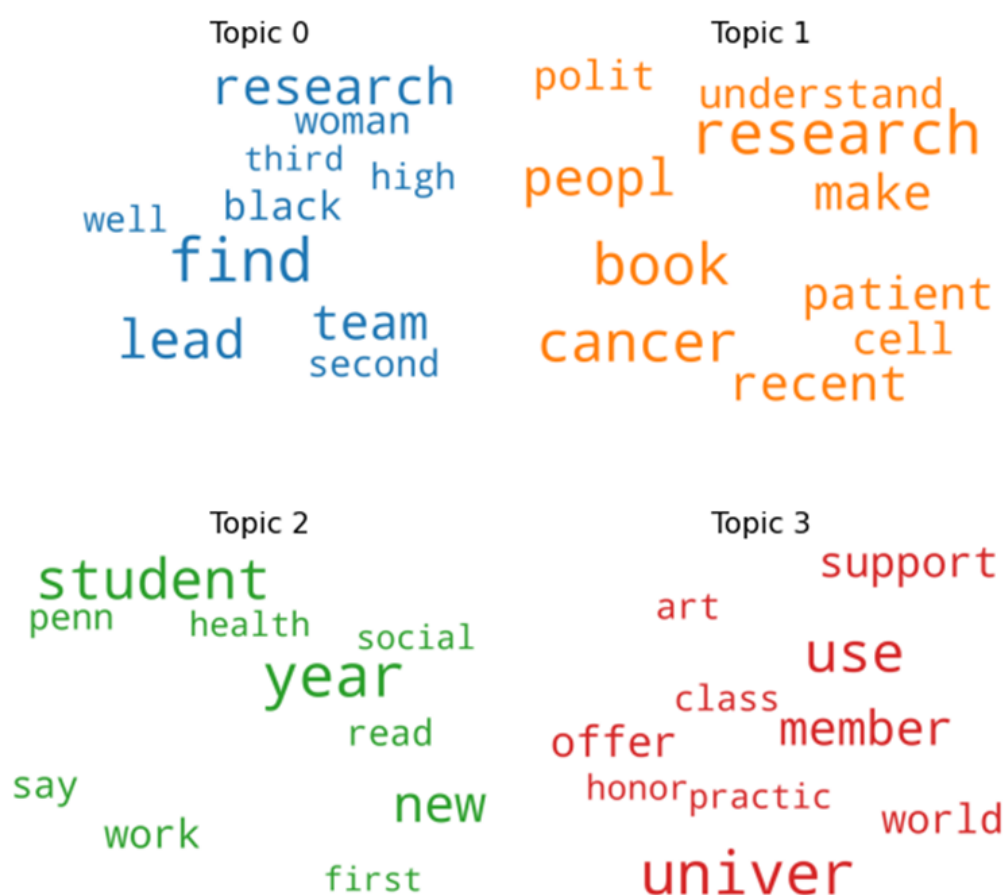


FIGURE B.13: University of Pennsylvania topics keywords. Topic 0: Sports, Topic 1: Research and Development, Topic 2: Beginning of the Academic Year and Topic 3: Education.

University of Porto



FIGURE B.14: University of Porto topics keywords. Topic 0: Education and Topic 1: Beginning of the Academic Year.

Bibliography

- [1] G. A. Olcay and M. Bulu, "Is measuring the knowledge creation of universities possible?: A review of university rankings," *Technological Forecasting and Social Change*, vol. 123, no. C, pp. 153–160, 2017. [Online]. Available: <https://EconPapers.repec.org/RePEc:eee:tefoso:v:123:y:2017:i:c:p:153-160> [Cited on pages 1 and 13.]
- [2] I. Aguillo, J. Bar-Ilan, M. Levene, and J. Ortega, "Comparing university rankings," *Scientometrics*, vol. 85, no. 1, pp. 243–256, 2010. [Online]. Available: <https://doi.org/10.1007/s11192-010-0190-z> [Cited on page 2.]
- [3] N. C. Liu, "The story of academic ranking of world universities," *International Higher Education*, no. 54, 2009. [Online]. Available: <https://doi.org/10.6017/ihe.2009.54.8409> [Cited on page 2.]
- [4] A. F. Van Raan, "Challenges in ranking of universities," in *Invited paper for the First International Conference on World Class Universities, Shanghai Jaio Tong University, Shanghai*. Citeseer, 2005, pp. 133–143. [Online]. Available: <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=0f62b2ea85a99549549ecadf4b01d8fe11b59e90> [Cited on page 2.]
- [5] J. Gajić, "Importance of marketing mix in higher education institutions," *The European Journal of Applied Economics*, vol. 9, no. 1, pp. 29–41, 2012. [Online]. Available: <https://journal.singidunum.ac.rs/files/2012-9-1/importance-of-marketing-mix-in-higher-education-institutions.pdf> [Cited on page 2.]
- [6] L. Oliveira and Álvaro Figueira, "Improving the benchmarking of social media content strategies using clustering and kpi," *Procedia Computer Science*, vol. 121, pp. 826–834, 2017, cENTERIS 2017 - International Conference on ENTERprise Information Systems / ProjMAN 2017 - International Conference on Project MANagement / HCist 2017 - International Conference on Health and Social Care

- Information Systems and Technologies, CENTERIS/ProjMAN/HCist 2017. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050917323098> [Cited on page 5.]
- [7] 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. [Online]. Available: <https://doi.org/10.1080/08841241.2016.1212451> [Cited on page 5.]
- [8] C. Assimakopoulos, I. Antoniadis, O. G. Kayas, and D. Dvizac, "Effective social media marketing strategy: Facebook as an opportunity for universities," *International Journal of Retail & Distribution Management*, vol. 45, no. 5, pp. 532–549, 2017. [Online]. Available: <https://doi.org/10.1108/IJRDM-11-2016-0211> [Cited on page 6.]
- [9] A. Irfan, A. Rasli, Z. Sulaiman, A. Sami, and M. I. Qureshi, "Use of social media sites by malaysian universities and its impact on university ranking," *International Journal of engineering and Technology*, vol. 7, no. 4.28, pp. 67–71, 2018. [Online]. Available: <https://core.ac.uk/download/pdf/288476961.pdf> [Cited on page 6.]
- [10] T. Coelho and Figueira, "Analysis of top-ranked hei publications' strategy on twitter," in *2021 IEEE International Conference on Big Data (Big Data)*, 2021, pp. 5875–5877. [Online]. Available: <https://ieeexplore.ieee.org/document/9671383> [Cited on page 6.]
- [11] T. Coelho and A. Figueira, "Covid-19 impact on higher education institution's social media content strategy," in *Computational Science and Its Applications – ICCSA 2021*, O. Gervasi, B. Murgante, S. Misra, C. Garau, I. Blečić, D. Taniar, B. O. Apduhan, A. M. A. Rocha, E. Tarantino, and C. M. Torre, Eds. Cham: Springer International Publishing, 2021, pp. 657–665. [Online]. Available: https://doi.org/10.1007/978-3-030-86960-1_49 [Cited on page 7.]
- [12] A. C. María del Rocío Bonilla Quijada, Eva Perea Muñoz and J.-L. Olmo-Arriaga, "Engaging students through social media. findings for the top five universities in the world," *Journal of Marketing for Higher Education*, vol. 32, no. 2, pp. 197–214, 2022. [Online]. Available: <https://doi.org/10.1080/08841241.2020.1841069> [Cited on page 7.]

- [13] I. Sörensen, S. Fürst, D. Vogler, and M. S. Schäfer, "Higher education institutions on facebook, instagram, and twitter: Comparing swiss universities' social media communication," *Media and Communication*, vol. 11, no. 1, pp. 264–277, 2023. [Online]. Available: <https://doi.org/10.17645/mac.v11i1.6069> [Cited on page 7.]
- [14] B. T. M. Yalalem Assefa and T. Kumar, "Modelling the predictive effects of social media on the relationship between students' engagement and academic achievement in higher education institutions," *Cogent Education*, vol. 10, no. 1, p. 2207801, 2023. [Online]. Available: <https://doi.org/10.1080/2331186X.2023.2207801> [Cited on page 8.]
- [15] 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. [Online]. Available: <https://www.scitepress.org/publishedPapers/2022/115925/pdf/index.html> [Cited on pages 8 and 59.]
- [16] 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. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-43088-6_9 [Cited on pages 8, 57, and 59.]
- [17] S. Symeonidis, D. Effrosynidis, and A. Arampatzis, "A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis," *Expert Systems with Applications*, vol. 110, pp. 298–310, 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0957417418303683> [Cited on page 9.]
- [18] Álvaro Figueira, "Uncovering social media content strategies for worldwide top-ranked universities," *Procedia Computer Science*, vol. 138, pp. 663–670, 2018, cENTERIS 2018 - International Conference on ENTERprise Information Systems / ProjMAN 2018 - International Conference on Project MANagement / HCist 2018 - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN/HCist 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050918317344> [Cited on page 9.]

- [19] A. 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*, ser. TEEM'18. New York, NY, USA: Association for Computing Machinery, 2018, p. 923–929. [Online]. Available: <https://doi.org/10.1145/3284179.3284342> [Cited on page 10.]
- [20] S. N. Brohi, T. R. Pillai, S. Kaur, H. Kaur, S. Sukumaran, and D. Asirvatham, "Accuracy comparison of machine learning algorithms for predictive analytics in higher education," in *Emerging Technologies in Computing*, M. H. Miraz, P. S. Excell, A. Ware, S. Soomro, and M. Ali, Eds. Cham: Springer International Publishing, 2019, pp. 254–261. [Online]. Available: https://doi.org/10.1007/978-3-030-23943-5_19 [Cited on page 10.]
- [21] J. Son, J. Lee, K. R. Larsen, and J. Woo, "Understanding the uncertainty of disaster tweets and its effect on retweeting: The perspectives of uncertainty reduction theory and information entropy," *Journal of the Association for Information Science and Technology*, vol. 71, no. 10, pp. 1145–1161, 2020. [Online]. Available: <https://doi.org/10.1002/asi.24329> [Cited on page 10.]
- [22] H. Kennedy, T. Kunkel, and D. C. Funk, "Using predictive analytics to measure effectiveness of social media engagement: A digital measurement perspective," *Sport Marketing Quarterly*, vol. 30, no. 4, pp. 265–277, 2021. [Online]. Available: https://www.academia.edu/download/66694142/2021_Kennedy_Kunkel_Funk_in-press_Predictive_analytics_SMQ-online.pdf [Cited on page 11.]
- [23] C. A. Palacios, J. A. Reyes-Suárez, L. A. Bearzotti, V. Leiva, and C. Marchant, "Knowledge discovery for higher education student retention based on data mining: Machine learning algorithms and case study in chile," *Entropy*, vol. 23, no. 4, p. 485, 2021. [Online]. Available: <https://www.mdpi.com/1099-4300/23/4/485> [Cited on page 11.]
- [24] J. C. Soares, R. Limongi, and E. D. Cohen, "Engagement in a social media: an analysis in higher education institutions," *Online Information Review*, vol. 46, no. 2, pp. 256–284, 2022. [Online]. Available: <https://www.emerald.com/insight/content/doi/10.1108/OIR-06-2020-0242/full/html> [Cited on page 11.]

- [25] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938. [Online]. Available: <http://www.jstor.org/stable/2332226> [Cited on page 13.]
- [26] A. P. Field, "Kendall's coefficient of concordance," *Encyclopedia of Statistics in Behavioral Science*, vol. 2, pp. 1010–11, 2005. [Online]. Available: https://www.discoveringstatistics.com/repository/kendall's_coefficient_of_concordance_ebs.pdf [Cited on page 13.]
- [27] W. Webber, A. Moffat, and J. Zobel, "A similarity measure for indefinite rankings," *ACM Trans. Inf. Syst.*, vol. 28, no. 4, nov 2010. [Online]. Available: <https://doi.org/10.1145/1852102.1852106> [Cited on page 14.]
- [28] K. Chauhan and A. Pillai, "Role of content strategy in social media brand communities: a case of higher education institutes in india," *Journal of Product & Brand Management*, vol. 22, no. 1, pp. 40–51, 2013. [Online]. Available: <https://www.emerald.com/insight/content/doi/10.1108/10610421311298687/full/html> [Cited on page 31.]
- [29] F. Sabate, J. Berbegal-Mirabent, A. Cañabate, and P. R. Lebherz, "Factors influencing popularity of branded content in facebook fan pages," *European management journal*, vol. 32, no. 6, pp. 1001–1011, 2014. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0263237314000607> [Cited on page 31.]
- [30] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, ser. WSDM '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 399–408. [Online]. Available: <https://doi.org/10.1145/2684822.2685324> [Cited on page 42.]
- [31] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003. [Online]. Available: <http://portal.acm.org/citation.cfm?id=944937> [Cited on page 42.]
- [32] D. Ramachandran and R. Parvathi, "Analysis of twitter specific preprocessing technique for tweets," *Procedia Computer Science*, vol. 165, pp. 245–251, 2019, 2nd International Conference on Recent Trends in Advanced Computing ICRTAC

- DISRUP - TIV INNOVATION , 2019 November 11-12, 2019. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050920300910> [Cited on page 42.]
- [33] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, Jul. 2011, pp. 262–272. [Online]. Available: <https://aclanthology.org/D11-1024> [Cited on page 43.]
- [34] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, "Using of jaccard coefficient for keywords similarity," in *Proceedings of the international multiconference of engineers and computer scientists*, vol. 1, no. 6, 2013, pp. 380–384. [Online]. Available: https://www.iaeng.org/publication/IMECS2013/IMECS2013_pp380-384.pdf [Cited on page 43.]
- [35] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297. [Online]. Available: https://digitalassets.lib.berkeley.edu/math/ucb/text/math_s5_v1_article-17.pdf [Cited on page 50.]
- [36] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, 1980. [Online]. Available: <https://doi.org/10.1109/TCOM.1980.1094577> [Cited on page 50.]
- [37] 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. [Online]. Available: <http://www.jstor.org/stable/2346830> [Cited on page 50.]