Covid-19 Impact on Higher Education Institution's Social Media Content Strategy

Tiago Coelho[1,2] and Alvaro Figueira[1,2]

[1] CRACS - INESC TEC, 4200-465 Porto, Portugal
[2] Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal
tsilvacoelho@gmail.com, arf@dcc.fc.up.pt

Abstract. In recent years we have seen a large adherence to social media by various Higher Education Institutions (HEI) with the intent of reaching their target audiences and improve their public image. These institutional publications are guided by a specific editorial strategy, designed to help them better accomplish and fulfill their mission. The current Covid-19 pandemic has had major consequences in many different fields (political, economic, social, educational) beyond the spread of the disease itself. In this paper, we attempt to determine the impact of the pandemic on the HEI content strategies by gauging if these social-economical, cultural and psychological changes that occurred during this global catastrophe are actively reflected in their publications. Furthermore, we identified the topics that emerge from the pandemic situation checking the trend changes and the concept drift that many topics had. We gathered and analyzed more than 18k Twitter publications from 12 of the top HEI according to the 2019 Center for World University Rankings (CWUR). Utilizing machine learning techniques, and topic modeling, we determined the emergent content topics for each institution before, and during, the Covid-19 pandemic to uncover any significant differences in the strategies.

Keywords: Text Mining · Covid-19 · Higher Education Institutions · Twitter · Topic Modeling.

1 Introduction

As competitiveness and the expectations of current and future students increase, and as the government support and new enrollments decrease, Higher Education Institutions (HEI) need to ensure their subsistence [1]. As such, it becomes vital for a HEI to obtain additional competitive advantages in the form of differentiation, which is achieved through self-promotion and with increased interaction with wider audiences. Most of the time, HEI use social media a a tool to do so. As it has been shown in the literature [2], advantageous social media campaigns are one of the most significant drivers of loyalty to an institution.

As the Covid-19 pandemic has been unfolding we are seeing several changes in the social, economic, and political landscape at a global scale. While various populations have had to adapt their behaviors to this new context, organizations have also had to adapt their strategies to meet the new needs of the populations, the current policies and regulations.

Therefore, in this project we seek to determine the impact that the Covid-19 pandemic had on Higher Education Institutions' content strategy by: i) creating an automatic system capable of discovering the communication strategy of a given HEI through the analysis of its communication pattern and the use of machine learning techniques, and ii) compare the discovered pre/post Covid-19 editorial strategies to determine if there are significant differences. Due to the recency of the pandemic, there is still a lack of literature that explores it's impact on the publication strategy of different HEI. However, regarding content strategy identification, there are already efforts made to deepen the knowledge about the different publication strategies employed by the institutions.

Our approach differs in several aspects from previous studies. With this project we aim to study HEI communication strategies on Twitter, whereas the studies of [3][4] and [5] use Facebook posts as the social network to analyze. The communication strategy of a HEI may vary according to the social network it is employed in, and we expect to find some differences from the previously identified strategies. Unlike the methodology used in the articles [3][4] our approach uses unsupervised machine learning techniques to obtain the topics on which each HEI is focusing their communication, and therefore, does not impose a predefined model. We believe our approach creates an editorial model that is more faithful to reality.

The Center for World University Rankings, CWUR, annually publishes a Global University Ranking which measures the quality of education and training of students as well as the prestige of the faculty members and the quality of their research without relying on surveys and university data submissions. By selecting HEI based on the 2019/2020 ranking by CWUR we expect to obtain better defined communication strategies than in [4]. The selected HEI are: Caltech, Cambridge University, ETH Zurich, Harvard University, Imperial College, Johns Hopkins University, MIT, Stanford University, University of Chicago, UC Berkeley, UCL and University of Oxford.

2 Data Set Characterization

Using the Twitter API we gathered all the posts from each of the 12 selected HEI for the period between the 1st of September 2019 to the 31th of October 2020. Even though not all HEI start and end their academic year at these exact dates, we believe that this period is the most fitting in representing an academic year for the majority of them. Taking into account the date on which the World Health Organization (WHO) declared Covid-19 as a Pandemic, March 11 of 2020, the selected period allows us to obtain a somewhat balanced division of the dataset into the pre/post Covid-19 periods, with a time window close to 6 months each.

When observing the posting frequency within the specified period we were able to determine that Imperial College was the HEI that published the most, while Stanford University was the one which published the least. Actually, the numbers are quite different for each organization. If we divide them by publication frequency we have: Stanford, UCL and Caltech in the same group, under 1k publications; Imperial College, Johns Hopkins and UC Berkeley with over 2k and, lastly, the remaining institutions with values in between. These values range between 2 and 8 posts a day, depending on the HEI. The average number of retweets and favorites is also significantly different. In what concerns retweets, it ranges from 10 (ETH Zurich) to 119 (Stanford) and the average number of favorites per tweet ranges from 13 (UCL) to 157 (Harvard). The standard deviations for both these metrics in each HEI are extremely high: this is the result of outlier trending tweets with high retweet and favorite counts that inflate the distribution.

With regards to posting patterns we were able to determine some common preferences by most organizations. All organizations have a more or less restrict set of hours that represent their activity period which results in all of them having a period of the day where little to no publications are made. The same can be said about the weekdays, with every organizations showing less activity during the weekend. To better understand posting frequency of each HEI throughout the year we aggregated their respective posts by month, and no evident similarities were found between the various HEI that could indicate the same trend/event. The same was made for the average number of favorites and retweets (Figure 1). To facilitate a comparison, we used the same scale for all graphs. The red dashed line represents the date Covid-19 was declared a pandemic.

Analyzing the depicted time series we can find numerous differences between the 12 HEI. However, there are some similarities in regard to seasonality and trend that we cannot overlook. The first evident aspect to point out is that the average number of favorites and retweets usually presents the same behavior due to the high correlation between the two, i.e, it is expected to see that tweets with a high/low number of favorites also have have a high/low number of retweets (and vice versa). Although, in Caltech, Cambridge, Stanford and Berkeley that is not always the case. Another noticeable aspect is that in most HEI the peak average number of retweets and favorites occur after the Covid-19 pandemic start, whereas Cambridge and Chicago figure as exceptions. The final interesting similarity is noticed in Harvard, Stanford, UC Berkeley and Oxford, in which taking into account the number of retweets we may say that these HEI did experience very similar events characterized by two major peaks and two major valleys, although not with the same intensity and not always at the same exact periods.

3 Emerging Topics Identification

For the topic modeling task we used the Non-Negative Matrix Factorization (NMF) and Latent Dirichlet Allocation (LDA) techniques because of their bet-

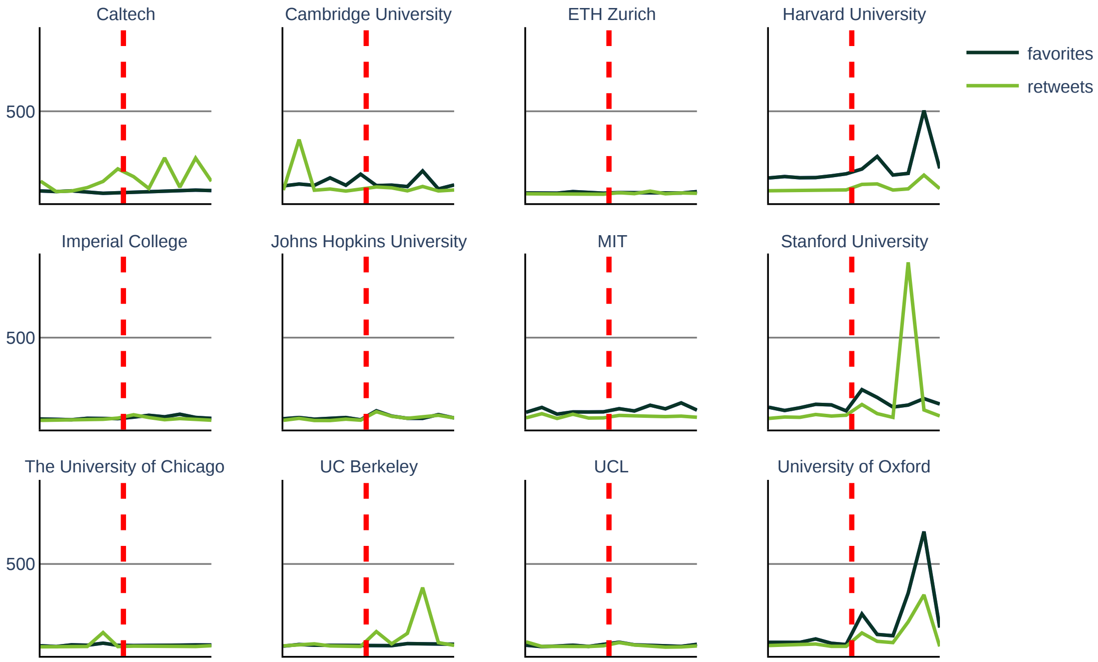

Fig. 1. Monthly average number of favorites and retweets for each HEI.

ter performance in the comparative analysis [6] on datasets which are similar to ours (short text, sourced from microblogs). NMF is an unsupervised matrix factorization (linear algebraic) method that is able to obtain topics from a collection of documents by performing both dimension reduction and clustering simultaneously. LDA topic modeling discovers topics that are latent in a collection of text documents by using a generative probabilistic model and Dirichlet distributions to infer possible topics based on the words present in the said documents. When choosing the LDA implementation, we opted to use the LDA Mallet wrapper for Genisim because it uses an optimized Gibbs Sampling[7] as the inference algorithm, which is more precise than the faster Variational Bayes algorithm[8] used by Gensim's LdaModel. After data cleaning steps, stop-word removal and tokenization we transformed the text to include bigrams and trigrams with hyperparameters determined trough iterative testing: total collected count above 50 and a score above 100 in Gensim's default bigram scoring function, based on the original [9]. Then, we lemmatized the text, performed POS tagging, keeping only nouns, verbs, adjectives and adverbs and created the two main inputs needed for topic modeling, the dictionary and the corpus. In order to determine the ideal number of topics for LDA and NMF models we decided to create models with an increasing number of topics and plot their measured performance. We evaluate the models in terms of topic coherence: A metric that scores a topic (between 0 and 1) by measuring the degree of semantic similarity between the high scoring words in it. Since the coherence of the models seem to increase with the number of topics (which can lead to an overfitting), we decided to select the models with the highest coherence values before the growth starts flattening out, or it reflects a major decrease. Based on the previously mentioned criteria, the number of topics was selected for both models (k=5), with the LDA model having a coherence score of 0.37 and the NMF model a coherence score of 0.28. The LDA model also revealed better human interpretability of the derived topics. Observing the inter-topic semantic similarities with the help of the pyLDAvis library, the 5-topic LDA model also proved superior, with greater inter-topic distance, making it easier to differentiate the topics. Therefore, we decided to advance with the LDA model with k=5 to characterize the publication strategies of the HEI. The topics discovered are represented in the following figure (Figure 2) using 10 of their most frequent words in the form of word clouds.

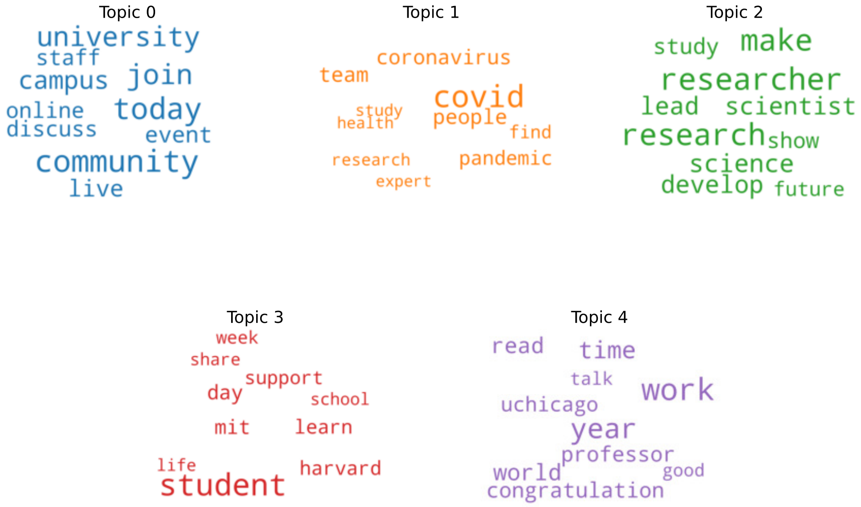

Fig. 2. Word clouds for each LDA generated topic.

Analyzing the word clouds we try to interpret the topics according the context they are inserted in. Topic 0 seems to be about college events: words like ”university”, ”join”, ”online”, ”live”, ”today”, ”discuss” and ”event” emphasize that idea. Topics 1 and Topic 2 are clearly related to public health/Covid-19 pandemic and research/investigation, respectively. Topic 3 seems to relate to education, evidenced by words: ”student”, ”learn”, ”school”, ”support”. Finally, Topic 4, seems to be the hardest to categorize, but pulling not only from the top 10 most salient words but from the top 30 we can get a clearer picture. We believe this topic regards public image due to words like ”congratulation”, ”work”, ”good”, ”great”, ”woman”, ”celebrate”, ”hope”, ”world” and ”history”. Having the topics 0 to 4 properly identified, we will refer to them hereafter as Events, Health, Research, Education and Image, respectively. From this basis, to better understand the publication frequency per topic, we assign the dominant topic to each tweet based on the tweet content.

One of the most glaring aspects that is easily observable in every HEI is the sudden and massive increase in the number of Health publications after the Covid-19 WHO pandemic declaration. Possibly correlated, we can also observe a shared decreasing trend of the Research topic over the academic year, with it’s biggest valley coinciding, in many cases, with the Health topic increase. This could be due to the semantic similarities between the Health and Research topics, since their inter-topic distance is the lowest out of any other possible pair, causing them to share some salient terms like ”research” or ”study”. We may say that Imperial College, UCL and University of Oxford possibly shared the same event, since all exhibit similar peaks in the number of Image publications between June 2020 and August 2020. In the University of Oxford’s case, this peak coincides with their absolute maximum in average number of retweets and favorites. In the same way, Stanford and MIT also have similar peaks in Research publications in October 2019, as well as Caltech and ETH Zurich, later that academic year, in January 2020.

4 Covid19 Impact Analysis

To determine each HEI strategy before and after the pandemic we displayed their posting behavior, color coded by topic, as horizontal stacked bar charts (Figure 5).

Before the pandemic, We can see that organizations like Stanford (30%), MIT (41%), Imperial College (30%), ETH Zurich (46%) and Caltech (50%) had Research as their primary topic of publications. Only UC Berkeley (37%) and The University of Chicago (31%) had Events as their primary topics, although the latter also had an almost equally matching preference for Image (28%). Focusing on Education oriented tweets we have UCL (31%), Johns Hopkins University (28%), Harvard University (25%) and Cambridge University (26%). However, this topic’s degree of dominance is not as high as the previous ones. All the HEI with this primary topic have a much more balanced distributions and a prominent second topic with almost as much focus as the first one. Both UCL and Johns Hopkins University prioritize the Events topic in their editorial model while Harvard University and Cambridge University prioritize Research. Lastly, the University of Oxford has 42% of tweets falling into the Image topic, the only one with this topic preference. Curiously, there are no HEI with a strategy that favors publications regarding the topic Health.

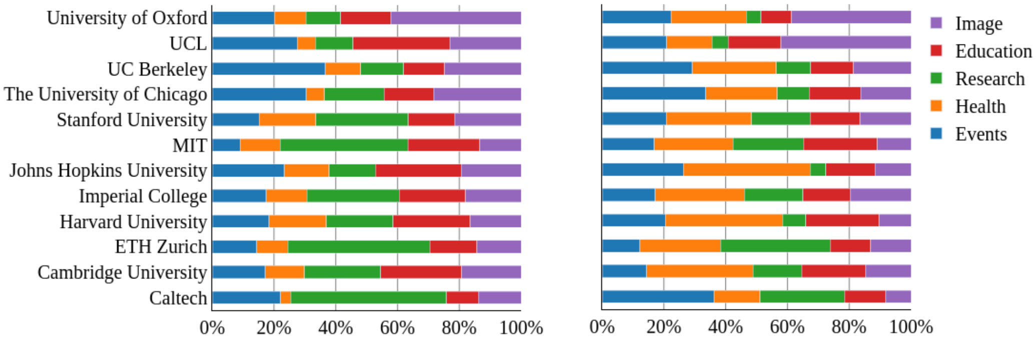

Fig. 3. Pre and Post Covid-19 topic frequencies, respectively

After the Covid-19 Pandemic start, we can clearly see significant changes in every institution's strategy. As expected, the Health topic, who previously wasn't the first topic of choice in any HEI, suddenly becomes the primary focus of UC Berkeley (27%), Stanford University (28%), MIT (25%), Johns Hopkins University (41%), Imperial College (29%), Harvard University (38%) and Cambridge University (34%). UC Berkeley, Imperial College, Harvard University and Cambridge University maintain their previous dominant topic as the second most prevalent, while the remaining do not. The University of Oxford, University of Chicago, and ETH Zurich maintained a similar strategy by keeping the same primary topic of publications. However, its dominance has decreased slightly due to the rise of the Health topic present in every single one of the post Covid-19 strategies. UCL and Caltech have had the most drastic changes, being the only ones with a complete switch of dominant topic, but not a switch to Health dominance. UCL joins Oxford University with 42% Image related content and Caltech switches from Research to Events with 36%.

To compare the topics themselves in the 2nd time window, we created the bar charts displayed in Figure 5, representing the share and favorites/retweets values per topic.

We can see that before Covid, Research was the most common topic of publication, represented by 25% of Tweets. After the pandemic, it became the least common, suffering a decrease of about 11%. Education and Image also suffered a 3% decrease each, making Events and Heath the only topics whose publication share increased after the pandemic. Health went from the least common to most common, with an increase of 17%.

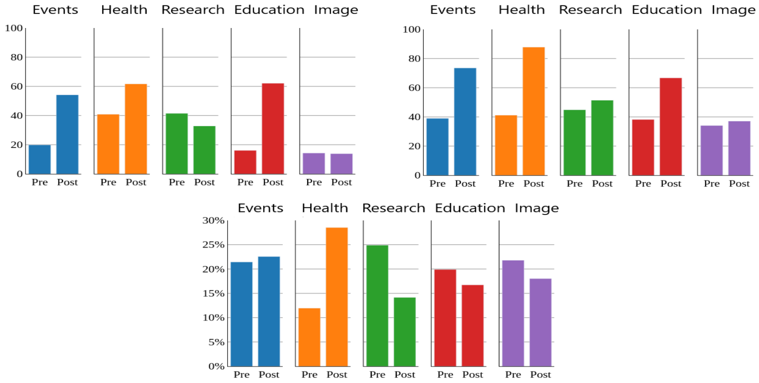

Fig. 4. Average number of retweets (Top-left), Average number of favorites (Top-Right) and Topic share percentage (Bottom), before and after the Covid-19 Pandemic.

In regard to the number of retweets and favorites, Events, Health and Education had similar results, each one increasing their number of interactions during the pandemic. Research and Image had a slight decrease in the average number of retweets but still had an increase in the average favorite count, marking an overall increase in average number of favorite across all topics during the pandemic.

5 Conclusions and Future Work

In this work we used the CWUR ranking to identify 12 of the highest ranking HEI, from which we collected 18727 tweets, for the period of 01/09/2019 to 31/08/2020. We then performed LDA topic modeling to identify topic areas present in each HEI content strategy and compare the resulting strategies before and after the Covid-19 WHO pandemic declaration.

Based on the results presented, we found that the editorial strategies of HEI Twitter can be characterized as revolving around five key content topics (events, health, research, education, and image) or editorial domains, and that there are significant strategic changes in topic distribution in each HEI after the pandemic declaration of Covid-19 WHO, indicating a change in content strategy in social media due to the current pandemic situation. In addition, we can also observe that the Covid-19 pandemic had an overall positive impact on public interaction with HEI social media publications. This could possibly be due to an increased interest in the topics discussed by the HEI according to its new content strategy, or even just the increased online presence of the public resulting from the pandemic restrictions leading to remote work and education.

Throughout the studied period, we can observe opposing trends for the Health and Research topics, with Health taking over the strategy focus that was previously placed on Research. This can be explained by the semantic similarity between the topics, as previously stated, but also could be due to the fact that most of the research efforts during that period were actively being directed towards the health industry. The Events topic was the only one besides

Health that managed to increase in all 3 of the studied metrics reflecting a prevalent interest/need for social interaction derived from the isolation caused by the pandemic.

Lastly, it's noteworthy to point out that, despite having the lowest average number of publications per day, Stanford has managed to achieve some of the best interaction numbers, with the highest average of retweets per tweet at 120 and the 3rd highest average of favorites per tweet. On the contrary, despite having the highest average number of publications per day, Imperial College was in 9th place in average retweets and the 7th place in average favorites, making a strong case for publication frequency not being a good predictor of engagement.

Apart from the previous conclusions, our contribution in this paper is also in the form of a proposed general methodology for understanding and measure social media content strategies in Higher Education Institutions. As a future work we will explore how the sentiment (expressed in each tweet) and associated with each topic unfolds during the pandemic and how it affected the postings of Higher Education Institutions.

References

1. Mitchell, M., Leachman, M., Masterson, K.: A Lost Decade in Higher Education Funding State Cuts Have Driven up Tuition and Reduced Quality (2017).
2. Çiçek, M.,Erdogmus, I.: The Impact of Social Media Marketing on Brand Loyalty. In: Procedia - Social and Behavioral Sciences, vol. 58, pp. 1353–1360 (2012). https://doi.org/:10.1016/j.sbspro.2012.09.1119
3. Figueira, A.: Uncovering Social Media Content Strategies for Worldwide Top-Ranked Universities. In: CENTERIS/ProjMAN/HCist 2018, vol. 138, pp. 663–670 (2018). https://doi.org/:10.1016/j.procs.2018.10.088
4. Oliveira, L., Figueira, A.: Measuring Performance and Efficiency on Social Media: A Longitudinal Study. In: ECSM 2018 5th European Conference on Social Media, pp. 198–207 (2018).
5. Peruta, A., Shields, A.: Social media in higher education: understanding how colleges and universities use Facebook. In: Journal of Marketing for Higher Education, vol. 27, pp. 131–143 (2017). https://doi.org/:10.1080/08841241.2016.1212451
6. Albalawi, R., Yeap, T., Benyoucef, M.: Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. In: Frontiers in Artificial Intelligence, vol. 3 (2020). https://doi.org/:10.3389/frai.2020.00042
7. Hoffman, M., Blei, D., Bach, F.: Online Learning for Latent Dirichlet Allocation. In: Advances in Neural Information Processing Systems, vol. 23, pp. 856–864 (2010).
8. Yao, L., Mimno, D., McCallum, A.: Efficient Methods for Topic Model Inference on Streaming Document Collections. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 937-–946 (2009). https://doi.org/:10.1145/1557019.1557121
9. Mikolov, T., Sutskever, I., Chen, K,, Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, pp. 3111—3119 (2013). https://doi.org/:10.5555/2999792.2999959