Analysis of Top-Ranked HEI Publications' Strategy on Twitter

Tiago Coelho
Faculty of Sciences of University of Porto
Rua do Campo Alegre, s/n
4169-006 Porto
Portugal
tsilvacoelho@gmail.com

Álvaro Figueira
CRACS / INESCTEC and University of Porto
Rua do Campo Alegre, s/n
4169-007 Porto
Portugal
arfiguei@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 strengthen their brand recognition. It is important for organizations to discover the true audience-aggregating themes resulting from their communication strategies, as it provides institutions with the ability to monitor their organizational positioning and identify opportunities and threats. In this work we create an automatic system capable of identifying HEI Twitter communication strategies. 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). Results show that there are different strategies, and most of HEI had to adapt them to the covid situation. The analysis also shows the prediction of topics and retweets for a HEI cannot just be based on recent historical data.**

Keywords—Higher Education Institutions, Twitter, Social Media Strategy, Machine Learning

I. INTRODUCTION

Annually, the Center for World University Rankings (CWUR) publishes an academic ranking of universities around the globe through a score made of four parameters. These parameters assess the quality of education, the alumni employment, the quality of the faculty, and research performance. Every parameter is assessed without relying on surveys nor on university data submissions. The ranking is expressed on a scale from 0 to 100 and is extremely competitive with the first places being disputed by differences of 1% or even tenths of a percentage point. This situation stems from the natural competitiveness between top institutions that dispute the best candidates to be their students, as well as to attract investment and funds from collaborations and projects, which in turn results in an overall improvement in the institution's own image.

Currently, in a world connected by social networks, these institutions cannot fail to invest in the communication they do on these networks. More than ever, their external communication strategies are relevant to their image in the community and this image ends up affecting the entire perception of excellence that they covey and identify. An initial study on the subject [1] was conducted in Facebook trying to understand how HEI really leverage social networks for communication engagement. In this context we present an investigation in which we collected all the publications of the 12 top-ranked world-wide higher education institutions (HEI), on Twitter, and analyzed them over the period of an academic year September 2019 – October 2020 (which also coincided with the outbreak pandemic) and to identify and compare the strategies that each HEI developed during this period.

II. METHODOLOGY

We used the CWUR annual ranking to identify the top 12 higher education institutions (HEI) during 2019/2020. These HEI are Caltech, Cambridge University, ETH Zurich, Harvard University, Imperial College, Johns Hopkins University, Massachusetts Institute of Technology (MIT), Stanford University, The University of Chicago, UC Berkeley, UCL and University of Oxford. Data was retrieved from Twitter using the Twitter's API collecting all of the posts from these HEI during the period September 1st of 2019 until October 31st of 2020.

We framed the experiment in this time window because: i) not all HEI start and end their academic years at the same time, ii) we believe this time frame is sufficiently adequate to accommodate an academic year and still be able to observe the effect of the pandemic situation in the communication strategy adopted by the institutions. Taking into account the date when the World Health Organization (WHO) declared Covid-19 as Pandemic, March 11th 2020, the selected period also allows us to obtain a somewhat balanced division of the dataset into pre/post Covid-19 periods, with a time window close to 6 months for each one. This makes possible to obtain behavioral changes from HEI according to the pandemic advance, characterize different editorial patterns in periods before and during the pandemic and compare the impact of same in the communication strategies of the selected HEIs.

In [2] the authors took a similar approach, but instead of relying on a pre-defined editorial model as they did, our approach is based on the emerging topics from the posts. In other words, our methodology consists of the following four steps: it starts with the characterization of HEI's communication strategies through factors external to the editorial areas present in the content, to obtain a good representation of the institutions' publishing practices and standards in this context. We then evaluate and interpret the models resulting from different approaches to identifying topics that assign an editorial standard to HEI communication strategies according to the content of their respective tweets. Thirdly, we carried out a brief analysis of the predominant sentiment of publications, we determined its evolution over the period studied and how the distribution of sentiment differs in each editorial area. Finally, using the models obtained above, we assign an editorial area to each publication in our dataset and through supervised learning, we assess the predictive capacity when trying to obtain predictions for the next topic to be addressed by a particular institution.

III. EXPLORATORY DATA ANALYSIS

By observing the publication frequency within the specified period (Table 1), we were able to determine that Imperial College was the HEI that published the most, while Stanford University was the least published. In fact, 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, with less than 1k of publications; Imperial College, Johns Hopkins and UC Berkeley with more than 2k and, finally, the remaining institutions with intermediate values. These values result from averages between 2 and 8 publications per day, depending on the HEI. The monthly distribution is depicted in Fig 1.

TABLE I. NUMBER OF POSTS AND FOLLOWERS BY HEI

HEI	Posts	Followers
Caltech	936	89 373
Cambridge University	1 010	557 387
ETH Zurich	1 126	46 978
Harvard University	1 923	1 152 771
Imperial College	2 746	129 082
Johns Hopkins University	2 060	178 596
MIT	1 782	1 106 098
Stanford University	756	784629
University of Chicago	1 629	63 834
UC Berkeley	2 074	199 923
UCL	890	83 130
University of Oxford	1 795	672 782

As there are significant differences in size between the institutions, it is important to take into account these differences in size when analyzing the involvement of institutions. We also found it relevant to observe the distribution of attributes representing involvement, the 'retweet_count' and 'favorite_count', in an initial attempt to identify the institutions with the greatest success/reach in their publications.

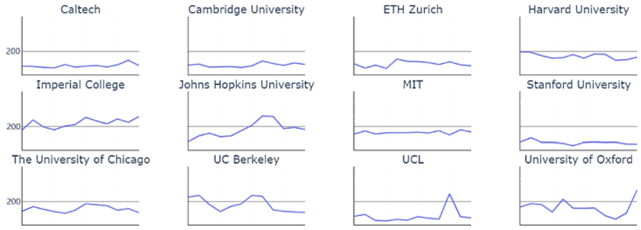


Fig. 1. Monthly posts by HEI.

Initially, with the help of the SpaCy library, the text fields of the collected tweets are subjected to pre-processing procedures. Then, we performed the optimization of hyperparameters of the respective algorithms, namely the number of topics to be used in the identification process. This is achieved through an iterative process, where for each value of number of topics k a model is trained with that number of topics and its coherence is evaluated according to the pipeline defined in [3]. In this phase, we use the Gensim library, one of most popular the libraries specialized in learning topic models.

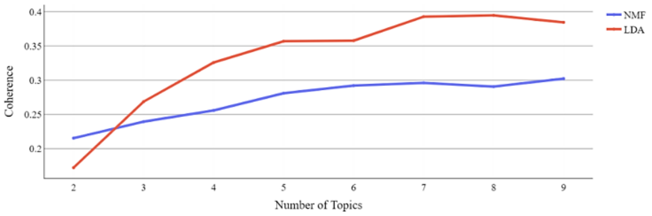


Fig. 2. Coherence according to the number of chosen topics.

As the coherence value tends to increase the greater the number of topics used (which can lead to overfitting), we limit the number of topics to 10 and consider, for the ideal number of topics, the highest value before the growth starts to flatten, or reflect a significant decrease. According to these criteria, we can conclude that the 5-topic models are the most suitable for our analysis, both for the NMF algorithm and for the LDA, having coherence values of 0.28 and 0.36 respectively. This approach is largely confirmed in [3].

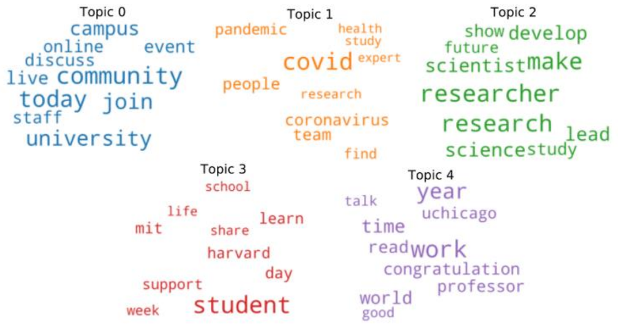


Fig. 3. World cloud of topics obtained through LDA.

Topic 0 seems to focus on university events: words like ’university’, ’join’, ’online’, ’live’, ’today’, ’discuss’ and event’ emphasize this idea. Topics 1 and 2 are clearly related to public health/Covid-19 and research, respectively. Topic 3 appears to be related to education, evidenced by words such as: 'student', 'learn', 'school', 'support'. Finally, Topic 4 seems to be the hardest to categorize, but not only looking at the 10 most relevant words, when looking at the 30 most relevant words, we can get a clearer picture. We believe this topic concerns the public image due to words like ’congratulation’, ’work’, ’good’, ’great’, ’woman’, ’celebrate’, ’hope’, ’world’ and ’history’. After obtaining the editorial model to be used, we proceeded to classify each publication according to its dominant topic. For sentiment analysis, we first obtain the polarity using the TextBlob library. Then, based on these polarity values, and using threshold values, we categorize these values into 3 different categories, 'Positive', 'Neutral ' and 'Negative', corresponding to the positive, neutral and negative sentiment, each covering the values of]0.33,1], [0.33, -0.33] and]-0.33, -1], respectively.

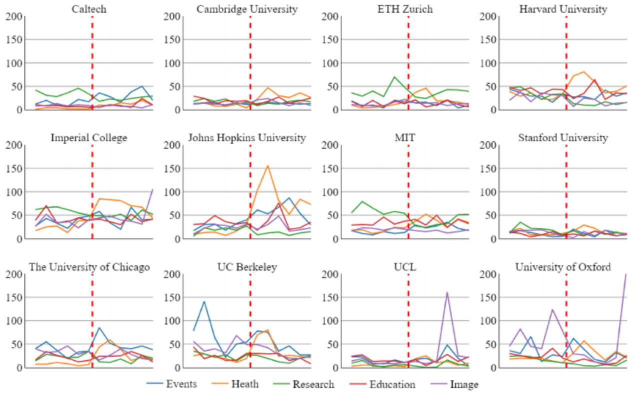


Fig. 4. Monthly distribution of topics in posts, per HEI.

In the graphs in Fig. 4 the time axis is shared for all observed HEIs. The red vertical line represents the date of Covid-19's declaration as a Pandemic, March 11, 2020. One of the most evident aspects, which is easily observable in all HEIs, is the sudden and massive increase in the number of Health publications after the declaration of Covid-19 as a pandemic by the WHO. Possibly correlated, we can also observe a shared decreasing trend of the Research theme throughout the academic year, with its greatest value coinciding, in many cases, with the increase in the Health topic.

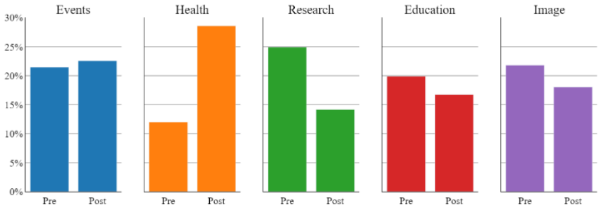


Fig. 5. Percentage of posts per topic in the pre- and pos-pandemic.

The polarity distributions, and hence the sentiment distributions of the different HEIs, are practically identical. They can all be generally characterized by a high density of neutral publications, with a decreasing number of publications as we approach the extremes, almost similar to a normal distribution. Unlike a normal distribution, this distribution does not have bilateral symmetry, which is the main aspect that differentiates them. The number of negative publications is significantly lower than the number of positive publications.

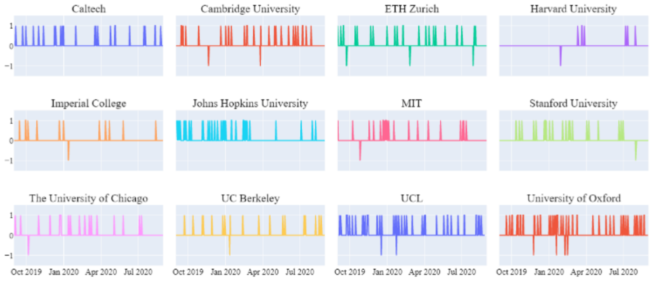


Fig. 6. Evolution of sentiment polarity per HEI..

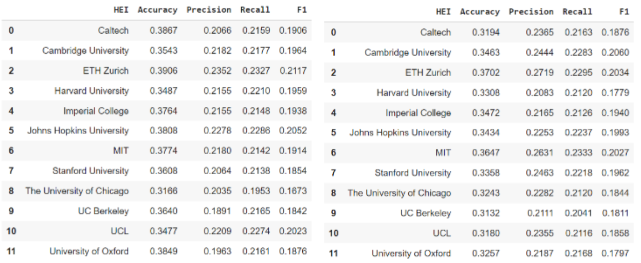
IV. Predictive Analysis

We also tested the predictive capacity of the models obtained in a classification trying to predict the main topic of the next post for each HEI. For this, we used our constructed model of discovered topics, where only the main editorial model used for the attribution of the dominant topic. The classification task also consists of supervised training of a model, capable of predicting the dominant topic of a publication by an HEI, based on topics from previous publications, in a time window of 7, and of 3 days.

Our objective is to forecast the dominant topic of a publication, information from the same institution varying the quantity of previous knowledge in topics per HEI, from one previous week to the previous 3 days. The most important pieces of information to obtain from this temporal window are the dominant topics, with the possibility of finding some regularity in their occurrence. The selected aggregation method of topics consists of the mode of all dominant topic values existing on the same day, for the same institution. Through this iterative process, for each institution, we were able to associate the dominant topics of the 7 (or 3) days prior to it to a publication, as well as other metrics related to that publication period (average number of retweets, average number of favorites and average polarity value) and other metadata (year, month, day of the week). The conversion process resulted in a data frame with 71 dimensions, representing an increase of 57 dimensions. Therefore, we selected 20 the most important features (using mutual information score) an used them for the prediction model.

V. Results

For the task of predicting the dominant topic of a publication, we used a classification task and obtained the results presented in Fig 7. We first considered a time frame of 7-days of prior knowledge (left) and then of 3-days of prior knowledge (right).



	HEI	Accuracy	Precision	Recall	F1
0	Caltech	0.3867	0.2066	0.2159	0.1906
1	Cambridge University	0.3543	0.2182	0.2177	0.1964
2	ETH Zurich	0.3906	0.2352	0.2327	0.2117
3	Harvard University	0.3487	0.2155	0.2210	0.1959
4	Imperial College	0.3764	0.2155	0.2148	0.1938
5	Johns Hopkins University	0.3808	0.2278	0.2286	0.2052
6	MIT	0.3774	0.2180	0.2142	0.1914
7	Stanford University	0.3608	0.2064	0.2138	0.1854
8	The University of Chicago	0.3166	0.2035	0.1953	0.1673
9	UC Berkeley	0.3640	0.1891	0.2165	0.1842
10	UCL	0.3477	0.2209	0.2274	0.2023
11	University of Oxford	0.3849	0.1963	0.2161	0.1876

	HEI	Accuracy	Precision	Recall	F1
0	Caltech	0.3194	0.2365	0.2163	0.1876
1	Cambridge University	0.3463	0.2444	0.2283	0.2060
2	ETH Zurich	0.3702	0.2719	0.2295	0.2034
3	Harvard University	0.3308	0.2083	0.2120	0.1779
4	Imperial College	0.3472	0.2165	0.2126	0.1940
5	Johns Hopkins University	0.3434	0.2253	0.2237	0.1993
6	MIT	0.3647	0.2631	0.2333	0.2027
7	Stanford University	0.3358	0.2463	0.2218	0.1962
8	The University of Chicago	0.3243	0.2282	0.2120	0.1844
9	UC Berkeley	0.3132	0.2111	0.2041	0.1811
10	UCL	0.3180	0.2355	0.2116	0.1858
11	University of Oxford	0.3257	0.2187	0.2168	0.1797

Fig. 7. Predicting tge next topic metric results.

In this task, the model that presents the best results for accuracy, precision and sensitivity is Random Forest, while the MLP presents the best F1-score.

VI. Conclusions

We could also observe that the Covid-19 pandemic had an overall positive impact on public interaction with HEI publications in the studied media, and the areas that showed a greater increase in interaction were Education and Health. , this may possibly be due to a growing interest in the teaching tools/content shared by HEIs, stemming from the pandemic restrictions leading to remote learning. In the area of Health, it may possibly be due to an increase in the search for information about the pandemic arising from the uncertainty and lack of knowledge characteristic of this initial phase that the world has experienced. However, this global increase in interaction may also simply be due simply to the increased online presence of the audience, resulting from the aforementioned pandemic restrictions that lead to remote work and learning.

The editorial area identified as Image is a misleading representation of the content contained therein. This area is better represented through its division into the Faculty, Employment and Society areas. The term Image suggests a self-promotional behavior, and our results revealed that the number of publications was similar to the level of all other areas. Regarding the results of the predictive analysis, we concluded that the information collected is not enough to be able to carry out the predictive analysis. For the prediction of the dominant topic of a publication, the best models are only 44% accurate. Again, we conclude that there are more factors than knowing previous posts, and some metadata from them, that define the content of a new post.

References

[1] Adam Peruta and Alison Shields. Social media in higher education: understanding how colleges and universities use facebook. Journal of Marketing for Higher Education, 27:1–13,

[2] 07 2016. doi:10.1080/08841241.2016.1212451.

[3] Álvaro Figueira. Uncovering social media content strategies for worldwide top-ranked universities. Procedia Computer Science, 18, 10 2018. doi:10.1016/j.procs.2018.10.088.

[4] David M. W. Powers. Evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. Journal of Machine Learning Technologies, 2:37–73, 2011.

[5] David M. Blei. Probabilistic topic models. Commun. ACM, 55(4):77–84, April 2012. ISSN: 0001-0782. doi:10.1145/2133806.2133826.