U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

ANÁLISE DE PUBLICAÇÕES NO TWITTER NAS INSTITUIÇÕES DO ENSINO SUPERIOR DO TOPO DE RANKING MUNDIAL

Tiago da Silva Coelho

Orientador: Professor Doutor Álvaro Figueira

[dcc]

DEPARTAMENTO DE CIÊNCIA DE COMPUTADORES
FACULDADE DE CIÊNCIAS DA UNIVERSIDADE DO PORTO

Contextualização, Motivação e Questões de Investigação

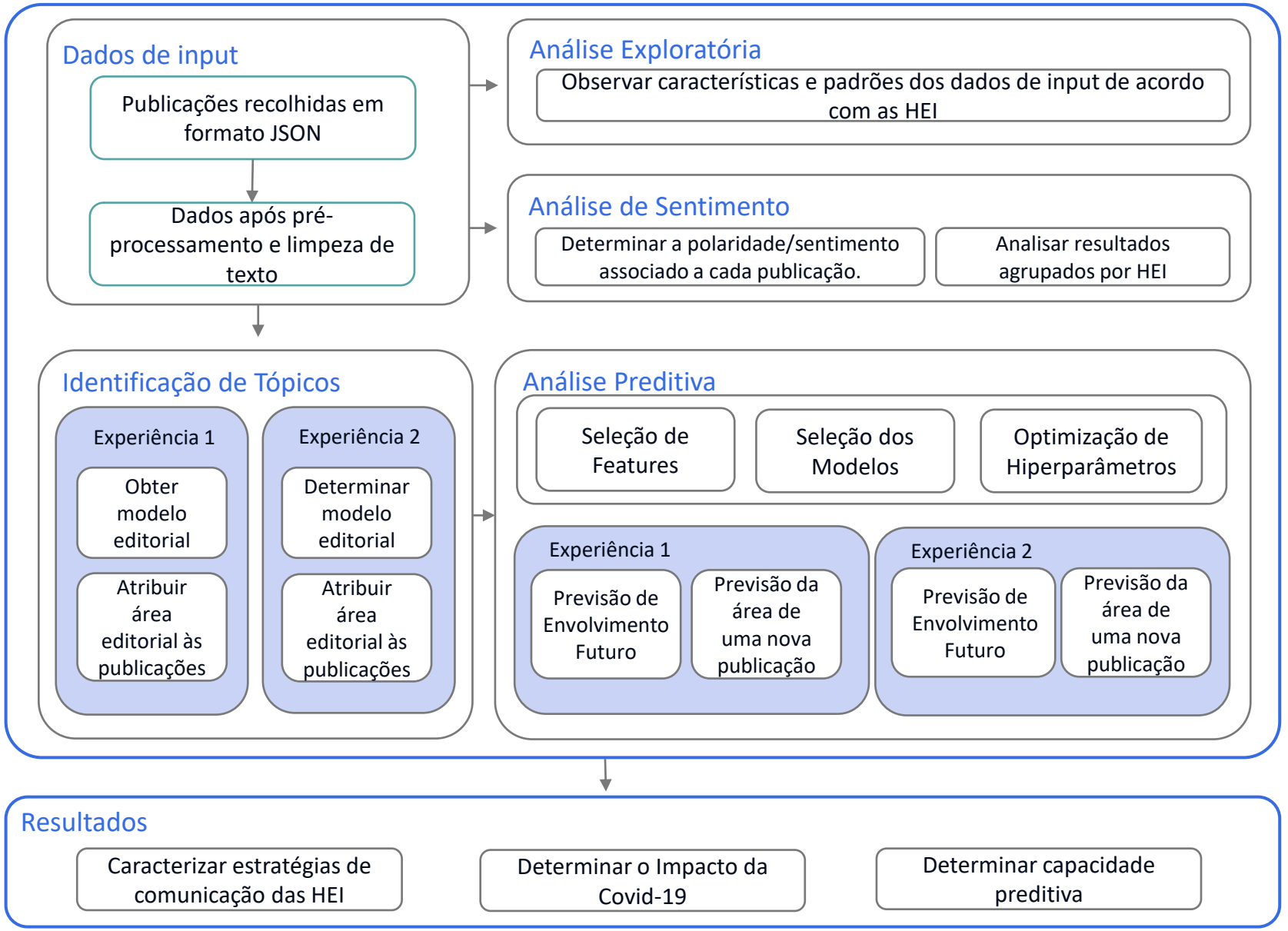
?

- Quais as estratégias empregues pelas HEI na comunicação através da rede social Twitter?
- Quais as estratégias com melhores resultados de interação/envolvimento?
- Quais as áreas editoriais de maior incidência? Quais as semelhanças estratégicas entre as instituições de acordo com essas áreas editoriais?
- Qual a influência do Covid-19 na comunicação das HEI?
- Será comum uma distribuição uniforme das publicações por área editorial ou existe uma baixa diversidade estratégica?
- Será possível prever a popularidade da publicação de uma HEI com base no seu conteúdo?
- Será possível prever o tópico da próxima publicação de uma HEI com base nas suas publicações anteriores?

Objetivos e Inovações

- Criar um sistema automático para a obtenção de estratégias de comunicação de uma HEI.
  - Modelo editorial estabelecido com base em conhecimento do domínio.
  - Modelo editorial obtido através de métodos de machine learning não supervisionados.
- Verificar o impacto da Covid-19 nas estratégias de comunicação obtidas.
- Determinar a possibilidade de previsão da interação futura de publicações.
- Determinar a possibilidade de previsão da área editorial de publicações.

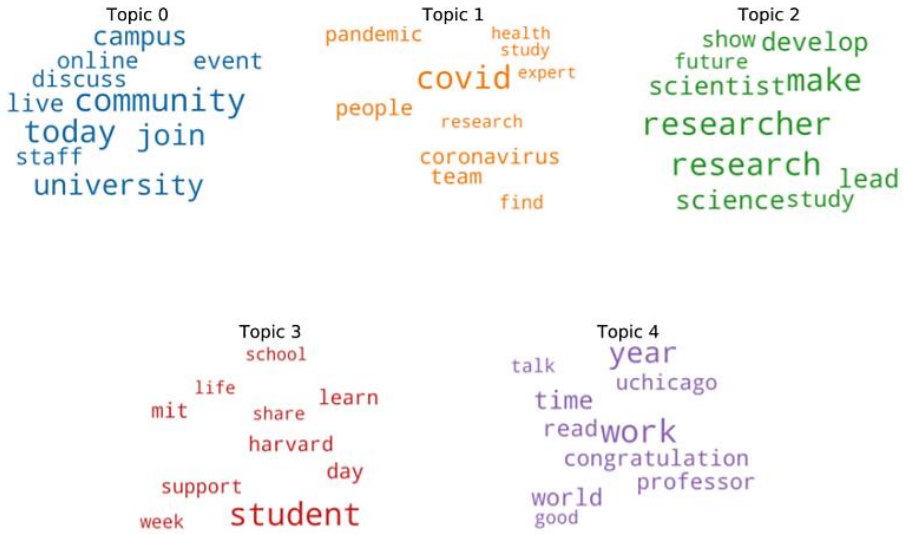
Metodologia



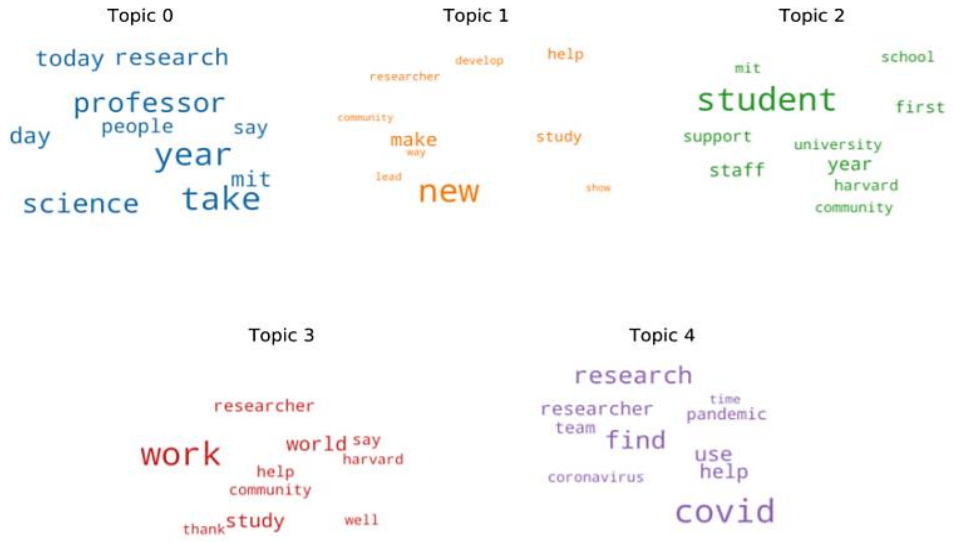
Metodologia – Experiência 1

- Consideramos dois algoritmos: Latent Dirichlet Allocation (**LDA**) e Non-Negative Matrix Factorization (**NMF**)

- Iterativamente, fizemos a otimização dos seus hiperparâmetros e obtivemos as seguintes áreas editoriais



Top 10 palavras mais relevantes para cada área editorial (LDA)



Top 10 palavras mais relevantes para cada área editorial (NMF)

Cálculo da relevância para um par (termo, tópico): $relevância(w,k) = log(\phi_{kw})$

Metodologia - Experiência 2

- Utilizando os critérios da CWUR, determinamos o modelo editorial a utilizar
- Através de valores de TF-IDF e cálculos de semelhança entre palavras representadas por embeddings, determinamos o método a utilizar para atribuição de uma área editorial a uma publicação

1. Cálculo da similaridade entre 2 embeddings

$$similaridade(A, B) = \frac{A \cdot B}{||A|| \times ||B||}$$

2. Cálculo dos valores TF-IDF

$$tf(t, d) = f_{t,d}$$

$$idf(t, D) = log\left(\frac{(1 + |D|)}{(1 + df(t))}\right) + 1$$

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D)$$

3. Método de atribuição de tópico dominante

$$Score(t, d, D, k) = tfidf(t, d, D) * similaridade(t, centroide_k)$$

$$Score(k, d) = \sum_{i=1}^{i=|d|} Score(t_i, d, D, k)$$

$$TopicoDominante(d) = \max(Score(k_i, d)), k \in [0, 5]$$

Metodologia - Previsão

Modelos de Previsão Comparados

- Linear Regression / Logistic Regression
- Decision Tree (DT)
- DT Bagging
- DT Boosting
- Random Forest (RF)
- K-Nearest Neighbours (KNN)
- Support Vector Machines (SVM)
- Multilayer Perceptron (MLP)

Medidas de Erro Utilizadas

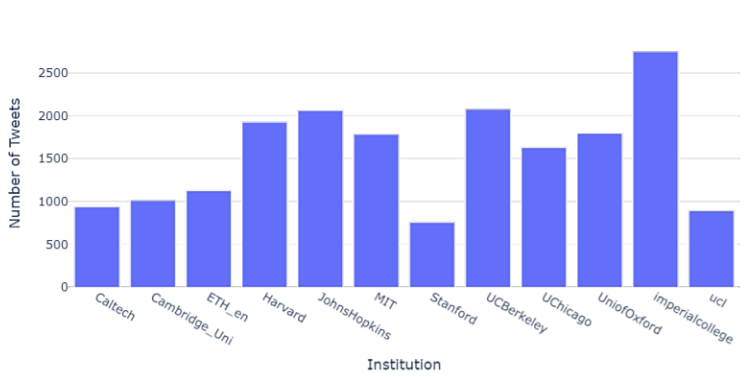
Regressão

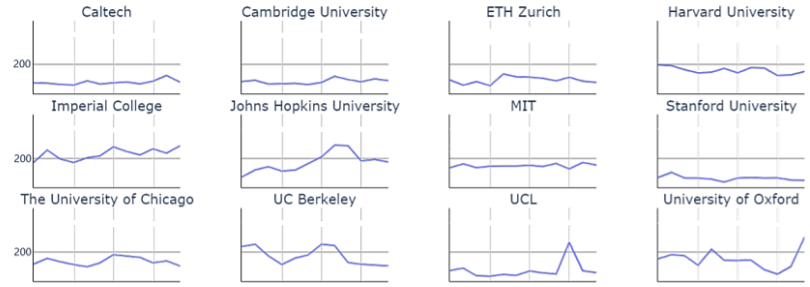
- Mean Squared Error (MSE)
- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- R-squared (R2)

Classificação

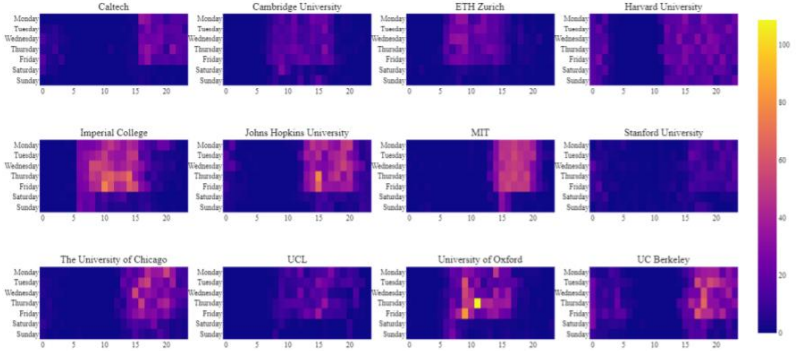
- Accuracy / Exatidão
- Precision / Precisão
- Recall / Sensibilidade
- F1-Score

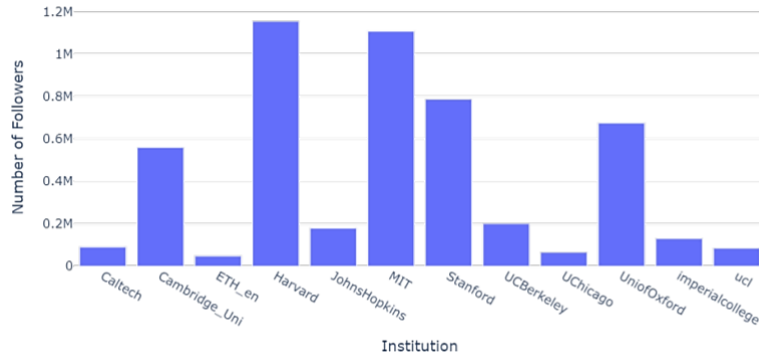
Análise Exploratória

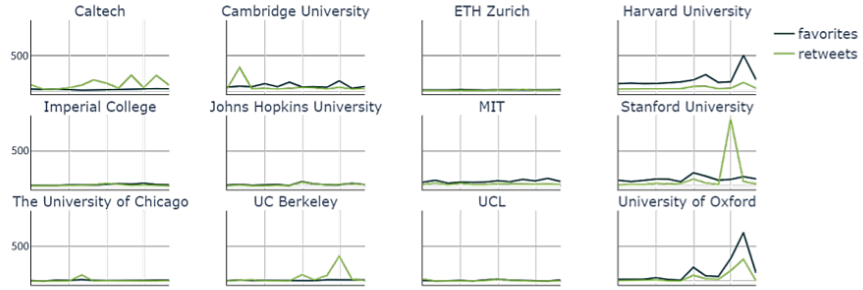




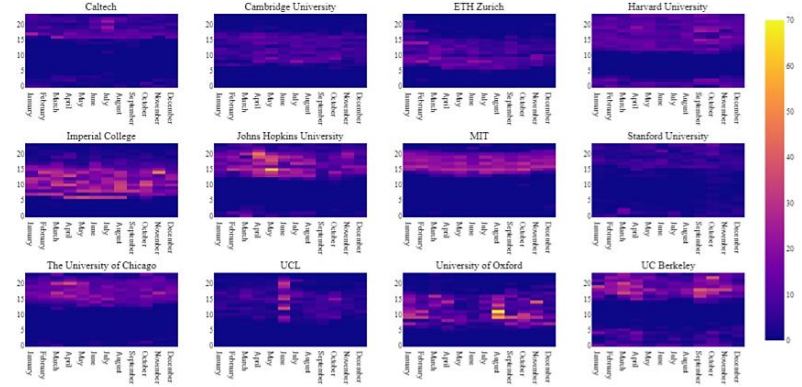
Publicações

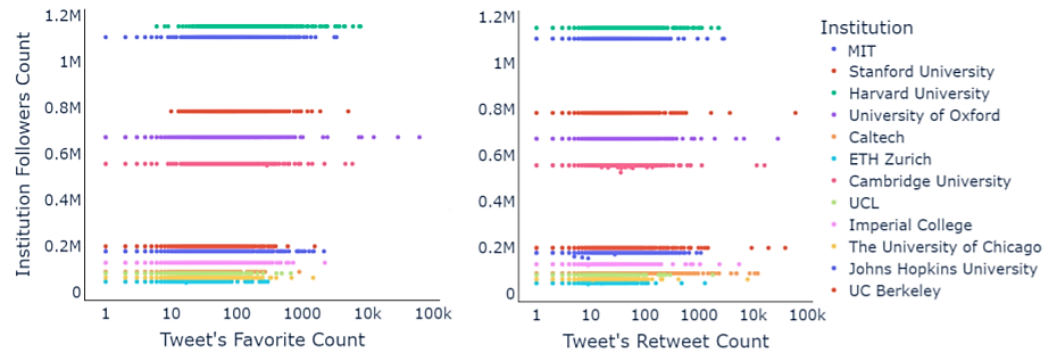


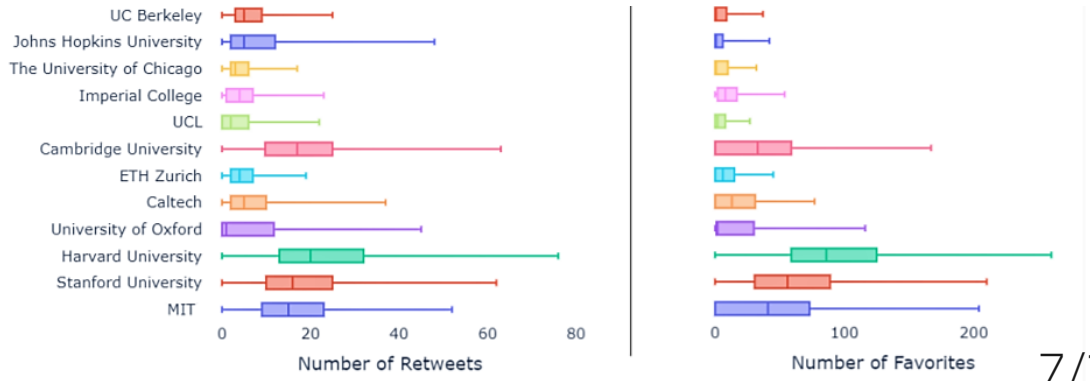




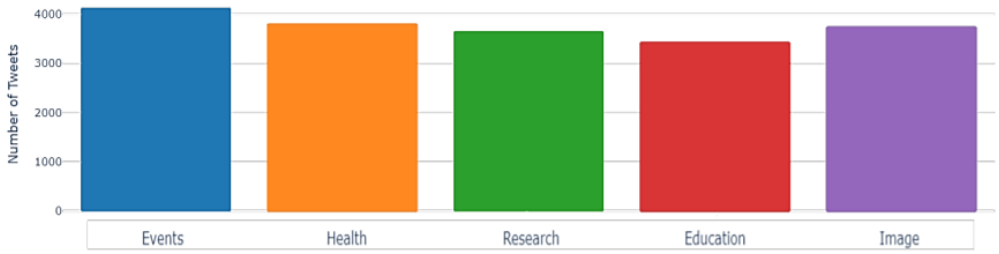
Envolvimento



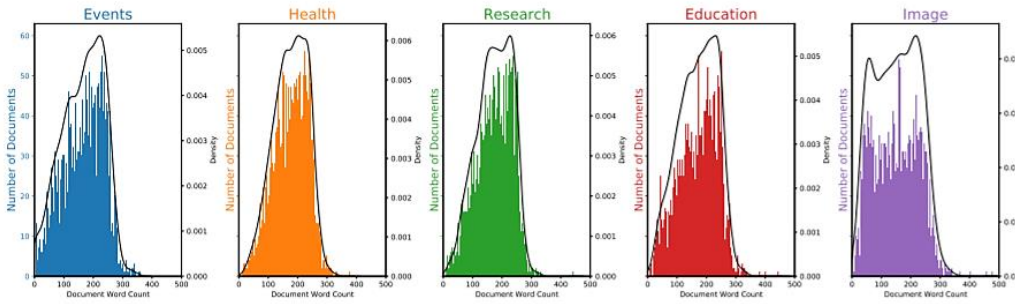




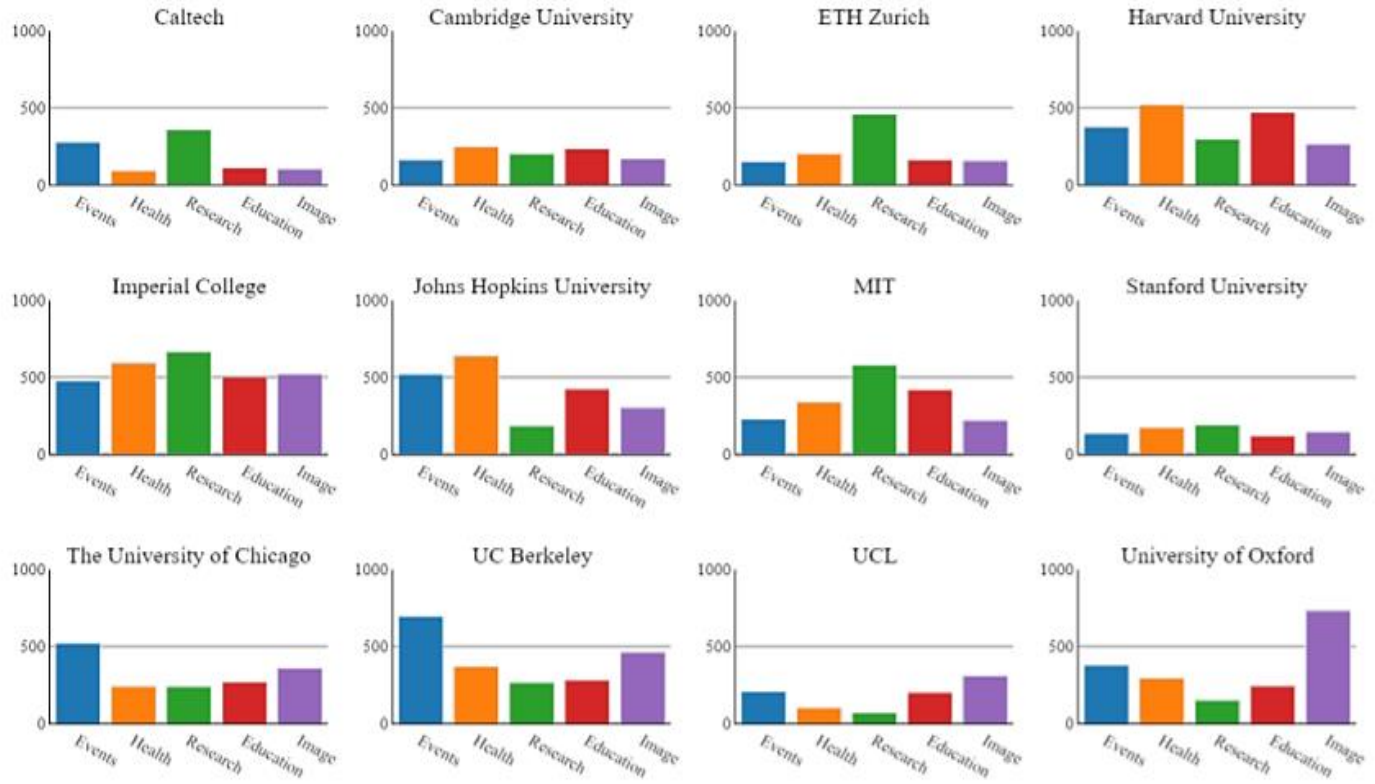
Análise Editorial: Experiência 1



Número de publicações por tópico

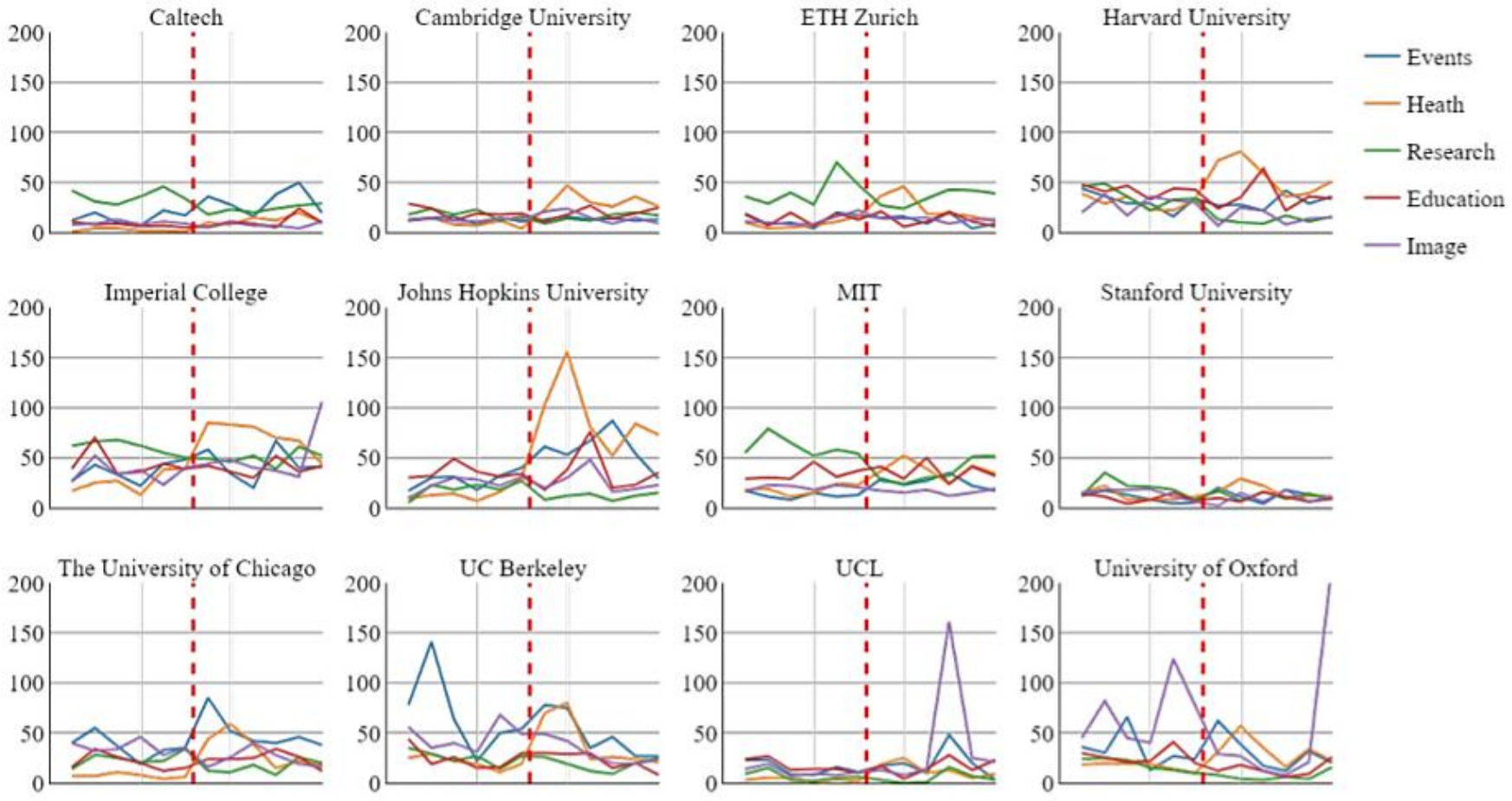


Número de palavras por tópico



Distribuição de publicações por tópico para cada HEI

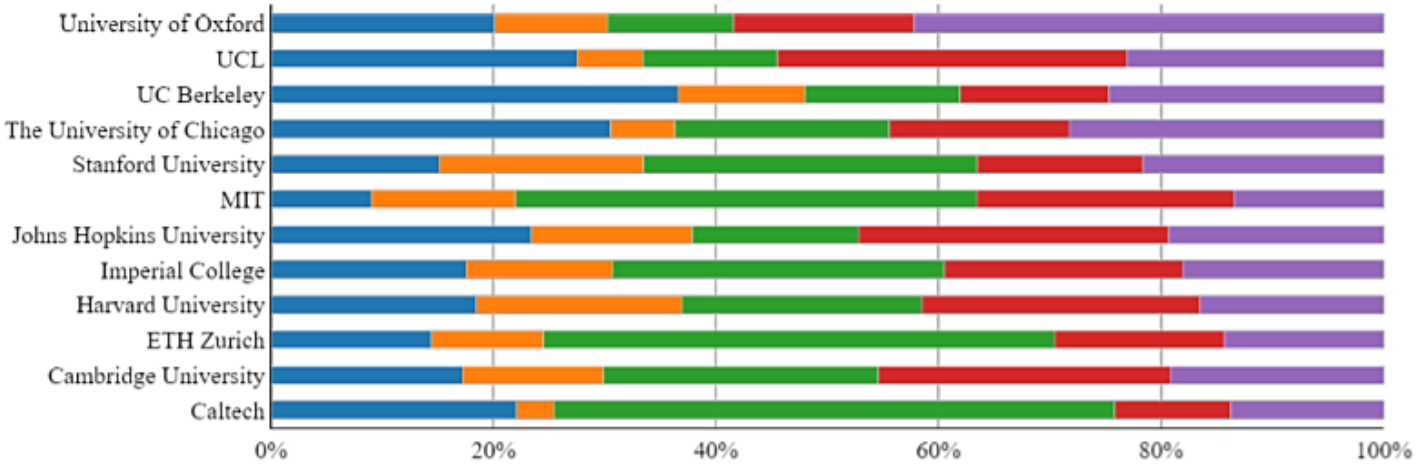
Análise Editorial: Experiência 1



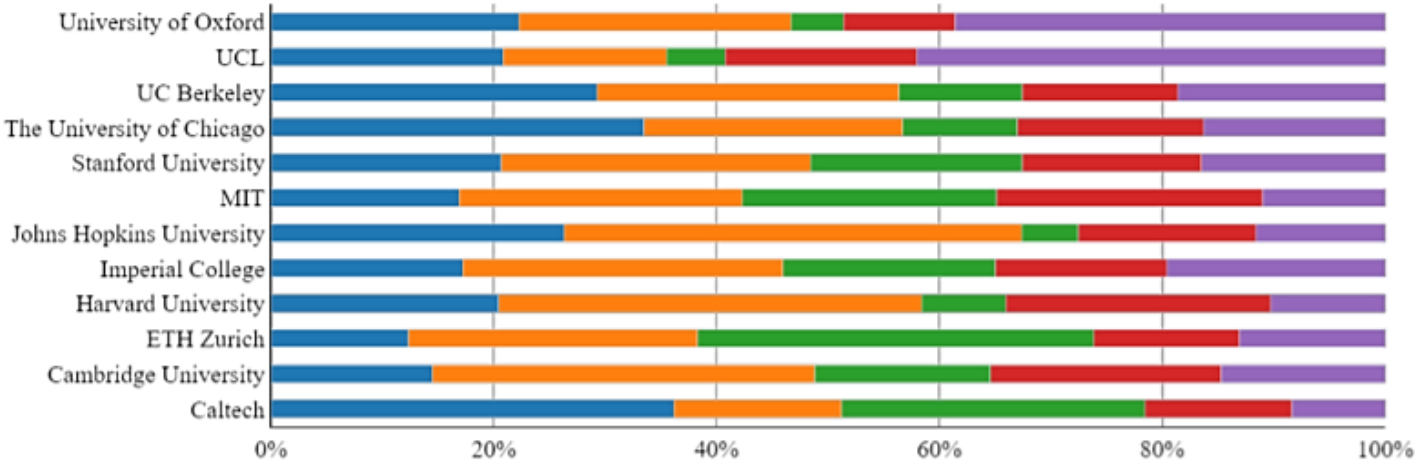
Evolução temporal do número de publicações por tópico

Análise Editorial: Experiência 1

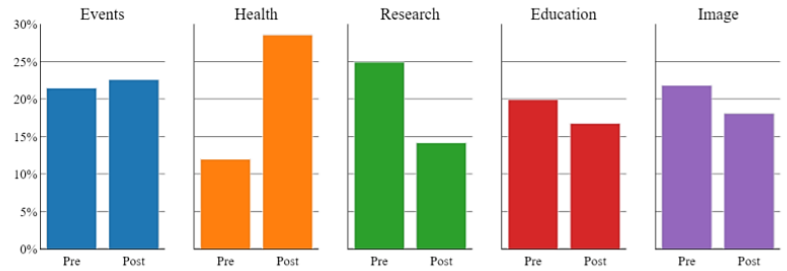
Antes da Covid-19



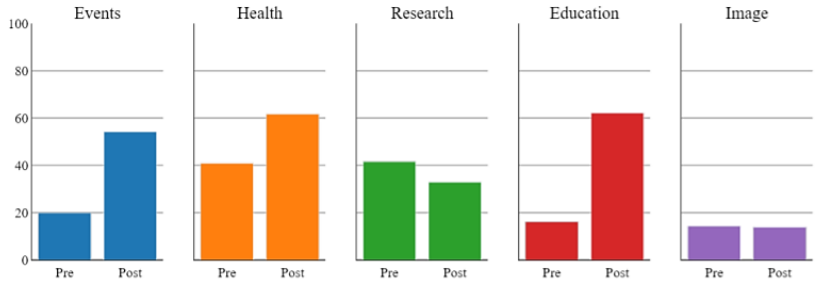
Depois da Covid-19



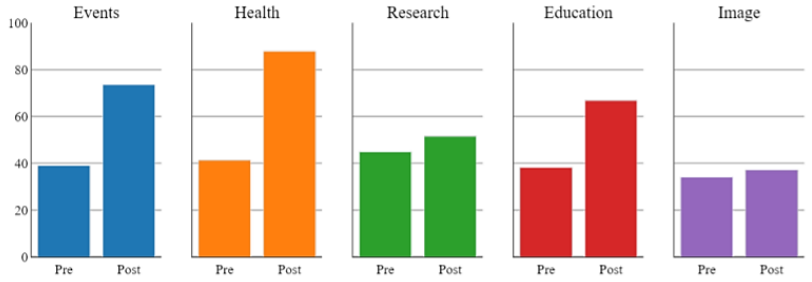
Publicações



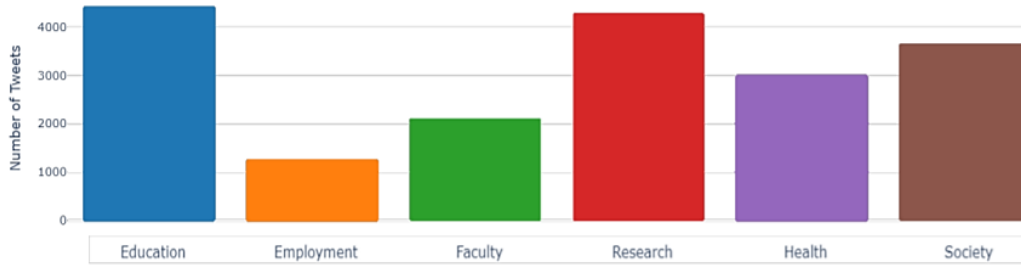
Retweets



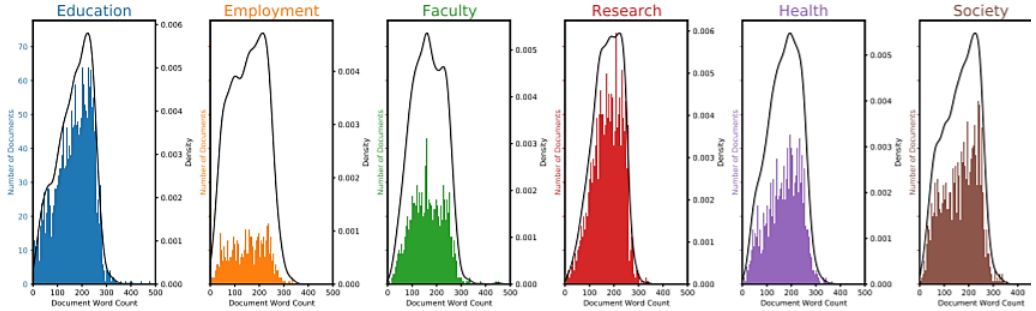
Favoritos



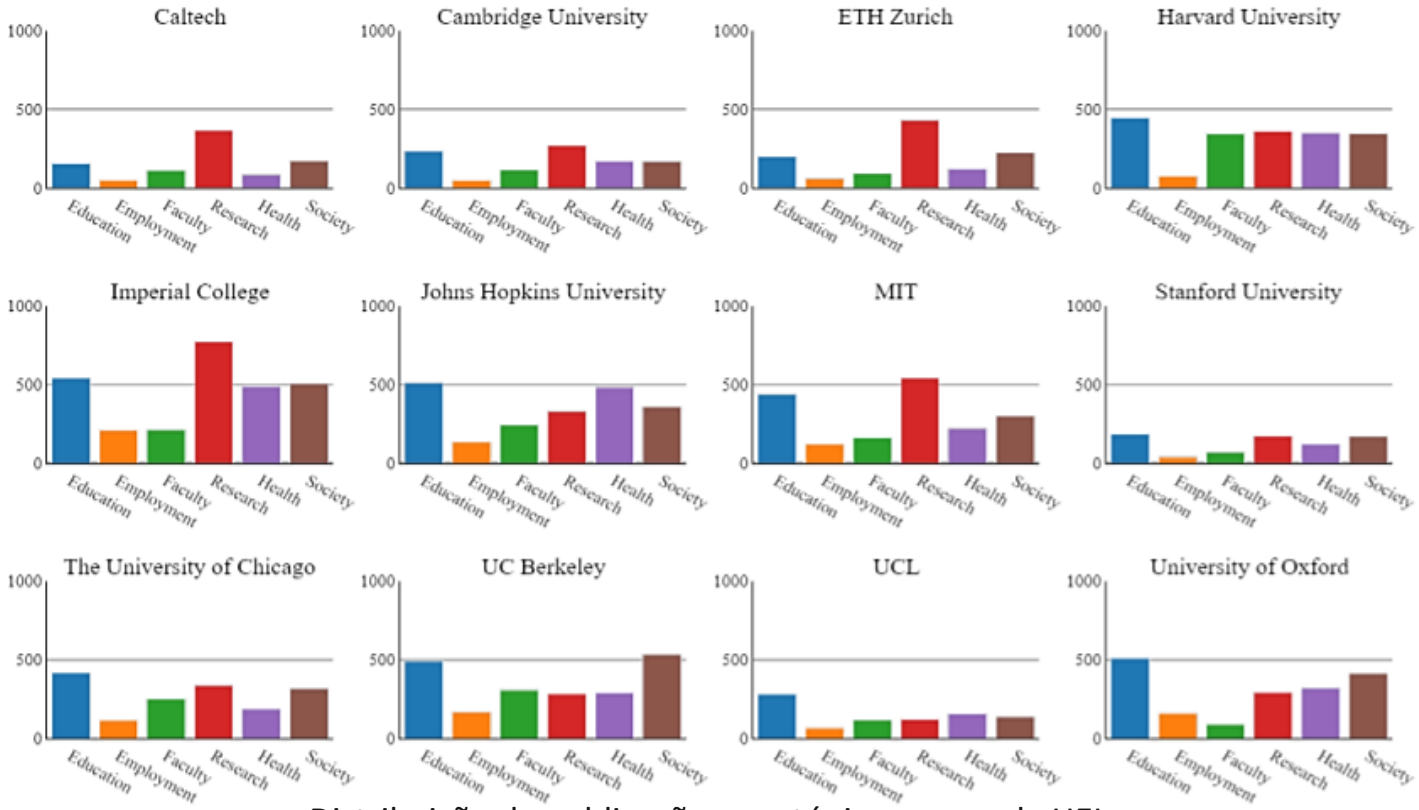
Análise Editorial: Experiência 2



Número de publicações por tópico

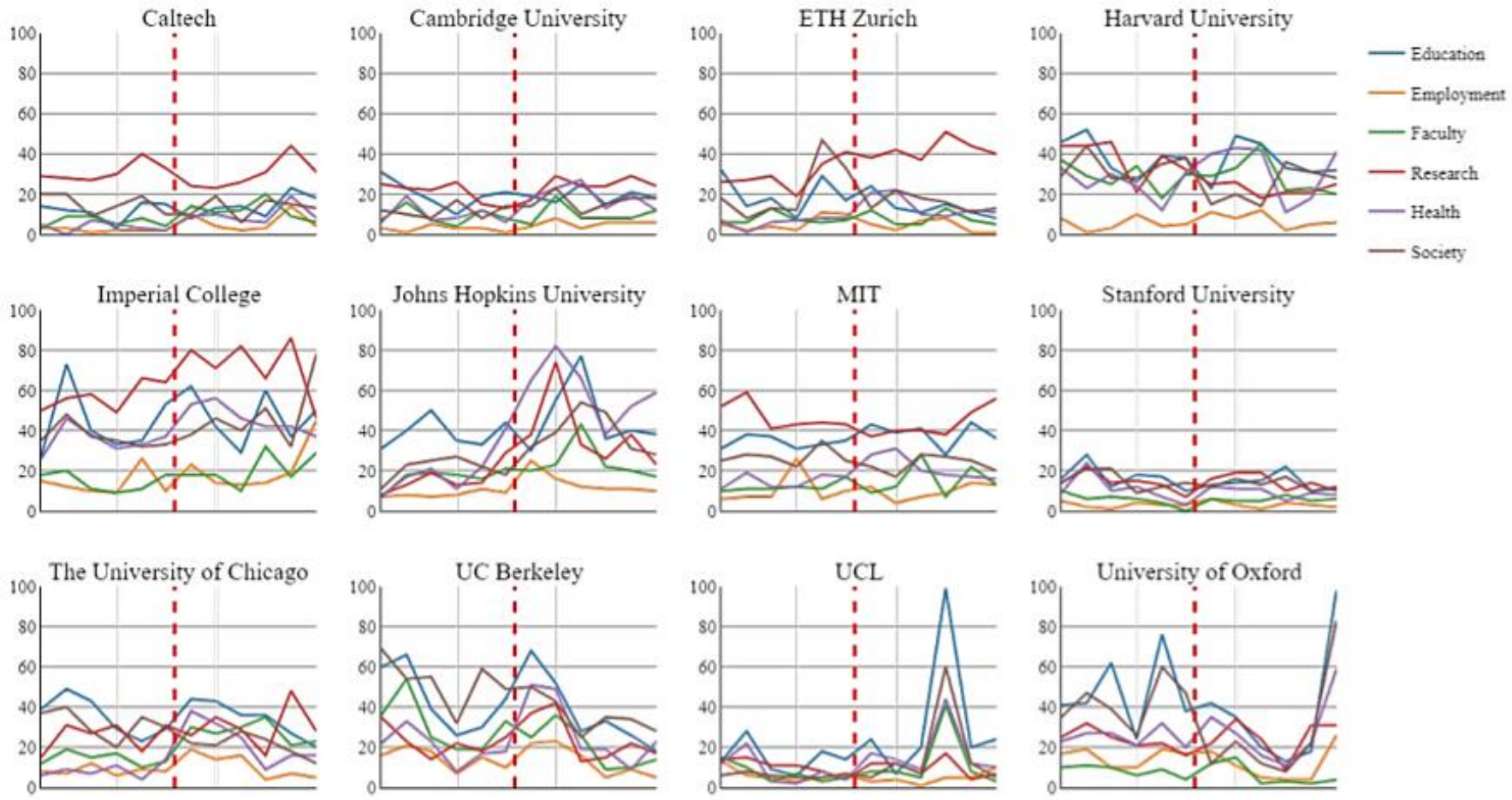


Número de palavras por tópico



Distribuição de publicações por tópico para cada HEI

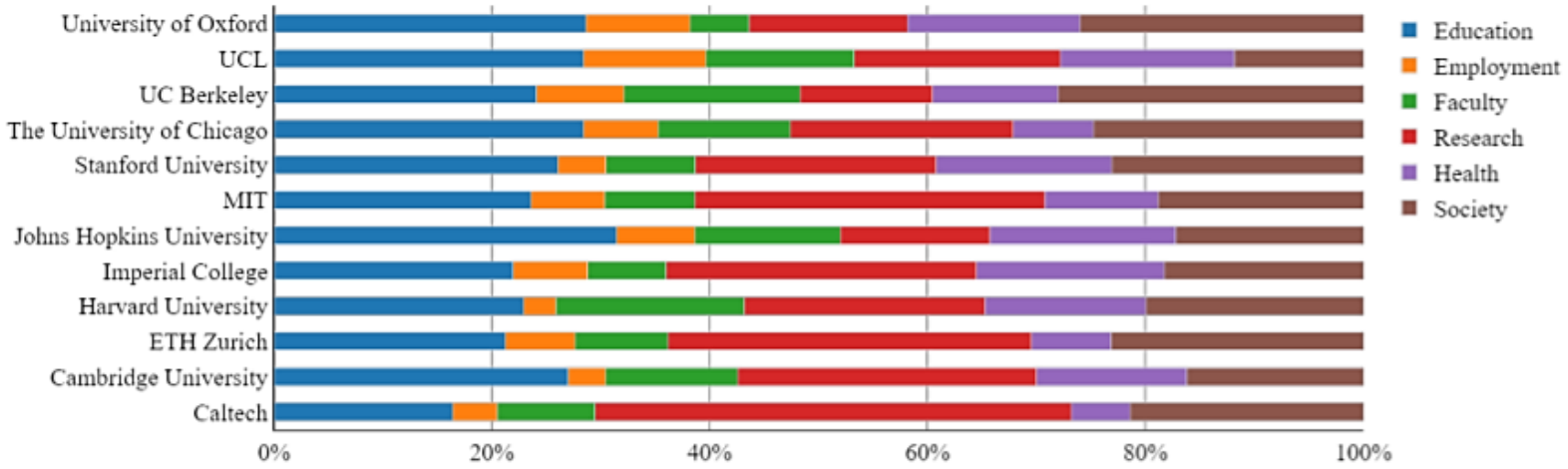
Análise Editorial: Experiência 2



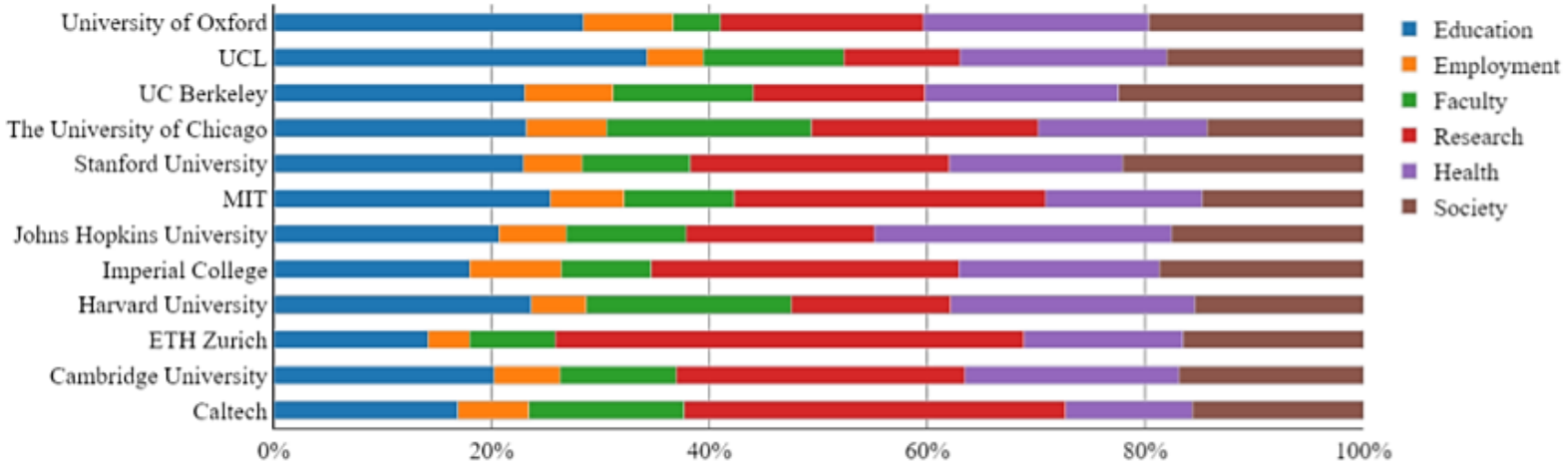
Evolução temporal do número de publicações por tópico

Análise Editorial: Experiência 2

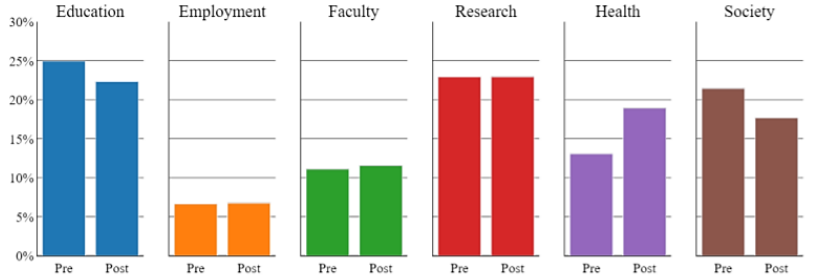
Antes da Covid-19



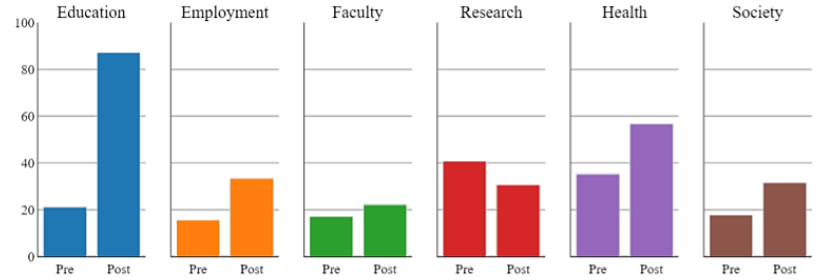
Depois da Covid-19



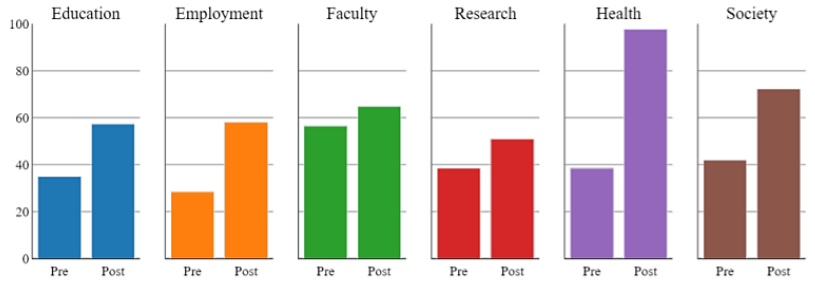
Publicações



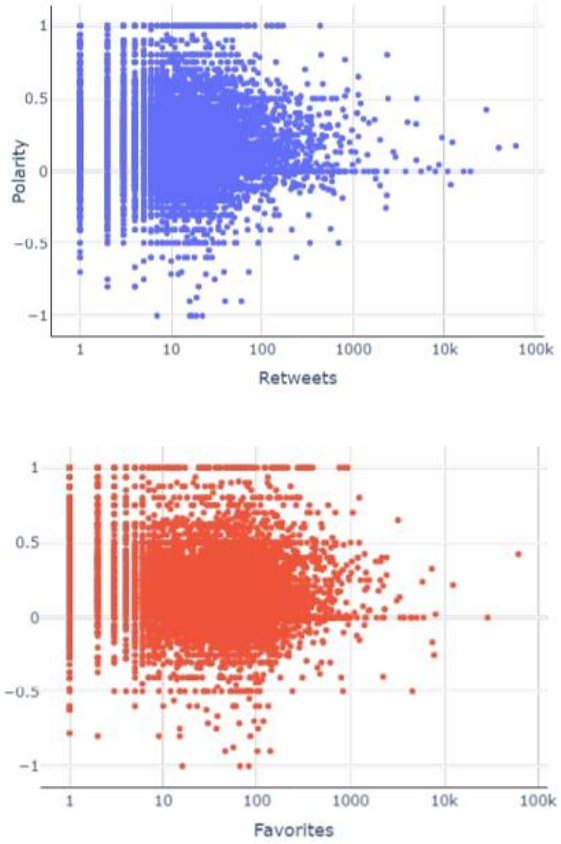
Retweets



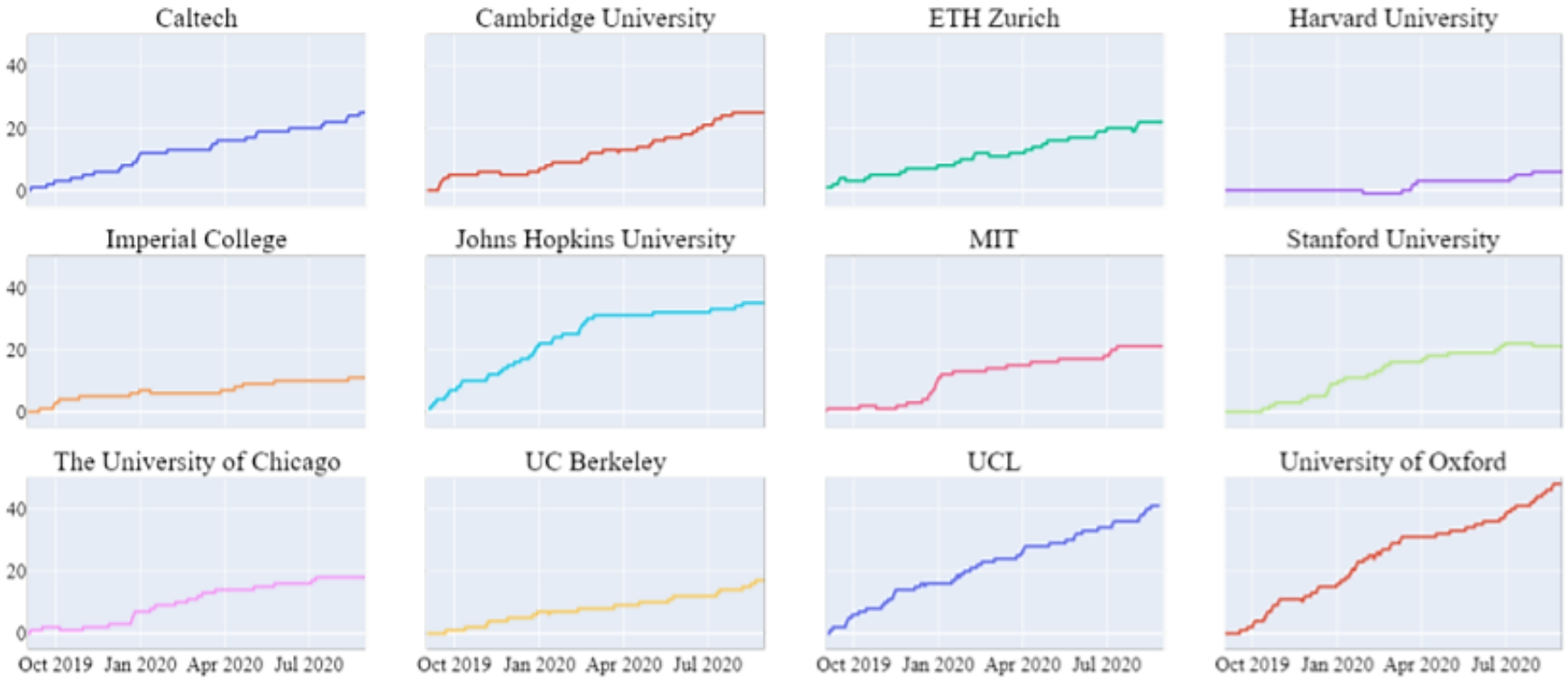
Favoritos



Análise de Sentimento



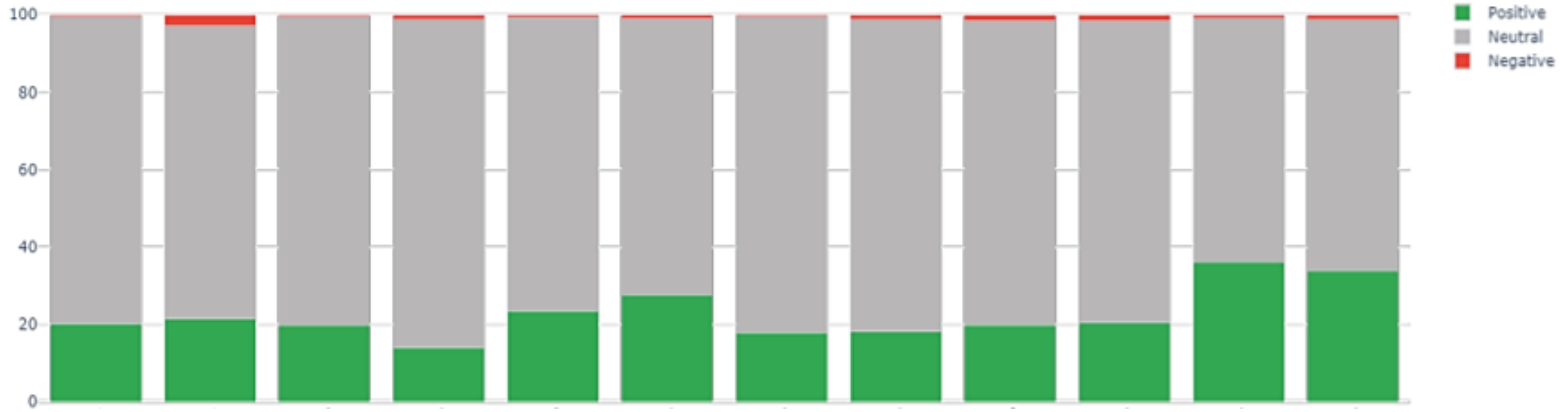
Relações polaridade/retweets e polaridade/favoritos



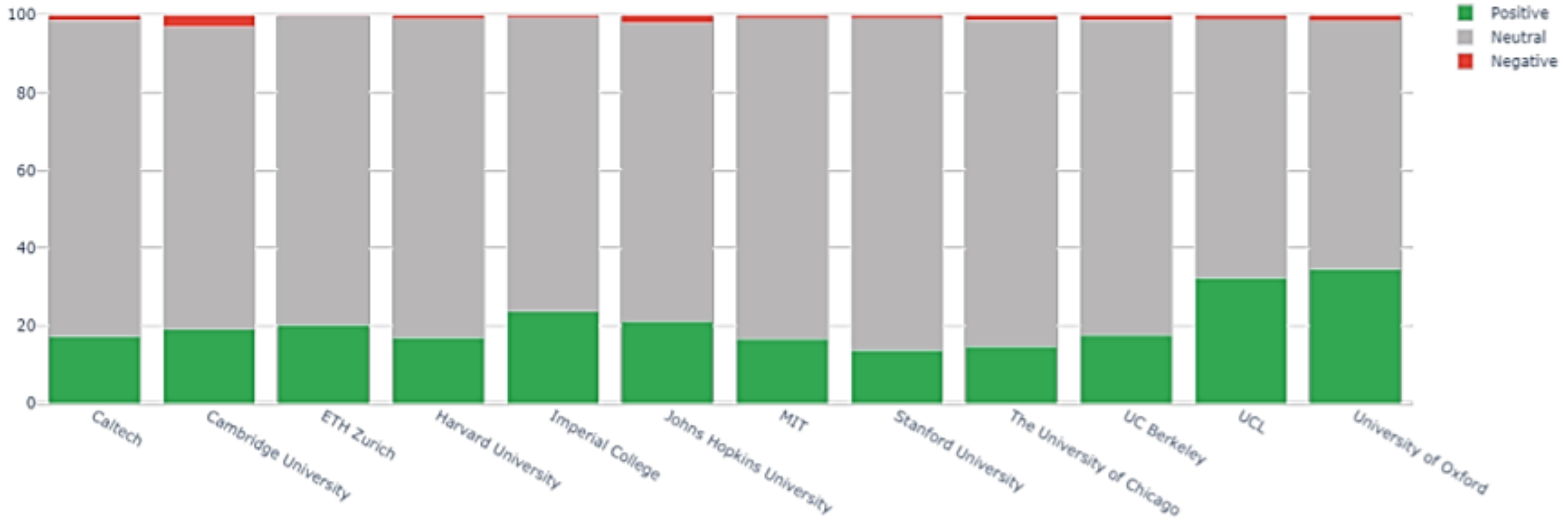
Evolução acumulada do sentimento por HEI

Análise de Sentimento

Antes da Covid-19

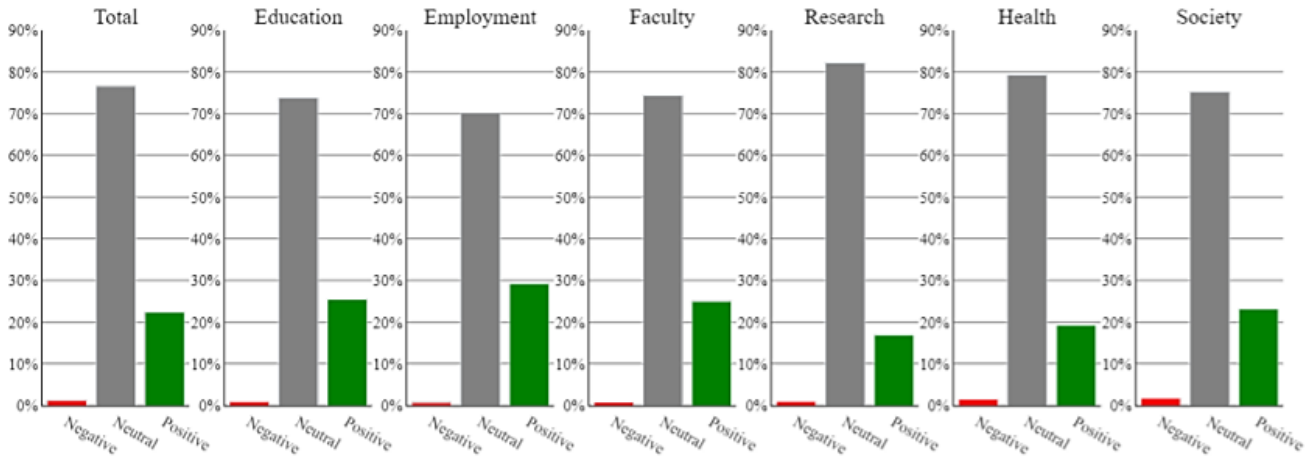


Depois da Covid-19

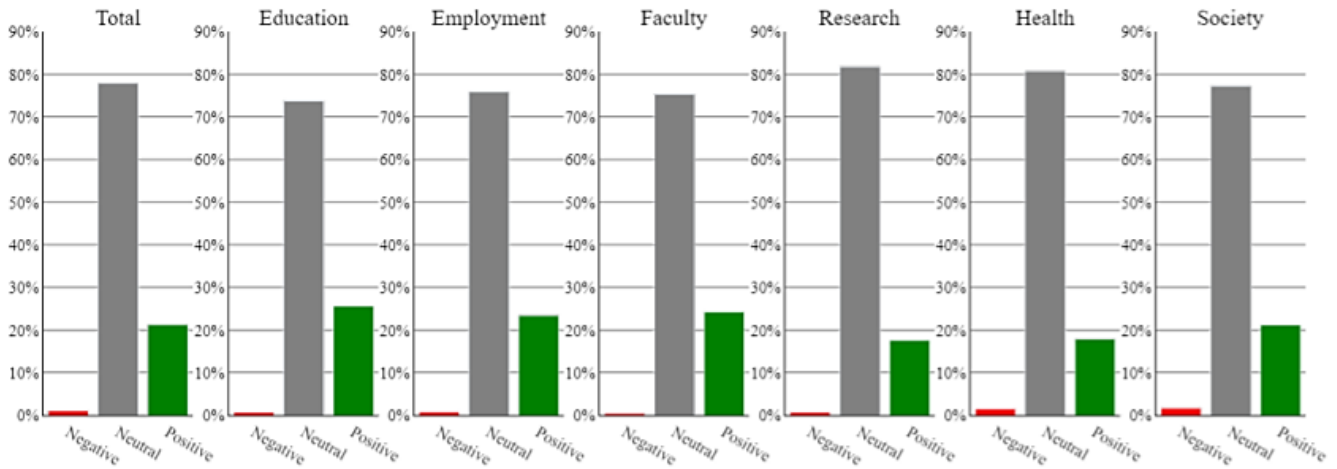


Análise de Sentimento: Experiência 2

Antes da Covid-19



Depois da Covid-19



Análise Preditiva: Experiência 1

Modelo	MSE	RMSE	MAE	R2
Regressão Linear	91.1129	9.5453	6.8918	0.2998
Regressão de Ridge	91.1036	9.5448	6.8932	0.2999
Regressão de Lasso	91.1118	9.5452	6.8943	0.2998
Árvore de Decisão	81.3809	9.0211	6.3894	0.3746
Bagging	131.3328	11.4601	8.5720	-0.0093
Boosting	87.6635	9.3629	6.9248	0.3263
Random Forest	77.7842	8.8195	6.2419	0.4022
K-NN	93.6225	9.6759	6.7605	0.2805
SVM	93.7397	9.6819	6.3183	0.2796
MLP	83.9911	9.1647	6.5326	0.3546

Tarefa de Regressão – Previsão de Retweets

Modelo	Exatidão	Precisão	Sensibilidade	F1
Regressão Logística	0.4082	0.3681	0.3791	0.3675
Árvore de Decisão	0.3691	0.3423	0.3390	0.3221
Bagging	0.4043	0.3616	0.3636	0.3266
Boosting	0.4102	0.3868	0.3753	0.3456
Random Forest	0.4336	0.4235	0.3981	0.3784
K-NN	0.4043	0.3928	0.3752	0.3508
SVM	0.4160	0.3854	0.3820	0.3583
MLP	0.4238	0.3823	0.3933	0.3791

Tarefa de Classificação – Previsão de Área Editorial

Análise Preditiva: Experiência 2

Modelo	MSE	RMSE	MAE	R2
Regressão Linear	96.9085	9.8442	7.0110	0.2814
Regressão de Ridge	96.8996	9.8438	7.0112	0.2815
Regressão de Lasso	96.9085	9.8442	7.0110	0.2814
Árvore de Decisão	87.4559	9.3518	6.6444	0.3515
Bagging	83.7966	9.1540	6.6214	0.3786
Boosting	90.4698	9.5116	7.0274	0.3292
Random Forest	82.6370	9.0905	6.4214	0.3872
K-NN	84.7157	9.2041	6.4311	0.3718
SVM	94.1320	9.7022	6.2479	0.3020
MLP	84.0360	9.1671	6.4982	0.3769

Tarefa de Regressão – Previsão de Retweets

Modelo	Exatidão	Precisão	Sensibilidade	F1
Regressão Logística	0.2788	0.2173	0.2122	0.2086
Árvore de Decisão	0.3091	0.1642	0.1805	0.1487
Bagging	0.3414	0.1966	0.2042	0.1789
Boosting	0.3414	0.1903	0.2033	0.1762
Random Forest	0.3515	0.2099	0.2117	0.1893
K-NN	0.3414	0.1886	0.2033	0.1772
SVM	0.3455	0.2057	0.2140	0.1937
MLP	0.3475	0.2022	0.2153	0.1988

Tarefa de Classificação – Previsão de Área Editorial

Conclusões

- Comparação e interpretação dos resultados obtidos pelos modelos editoriais de ambas as experiências
  - Uniformidade
  - Estabilidade
  - Credibilidade
- Impacto da Covid-19 na estratégia de comunicação das HEI
- Possíveis explicações para as tendências observadas
- Utilidade dos modelos de previsão obtidos

Conclusões

Questões de investigação

- Quais as estratégias empregues pelas HEI na comunicação através da rede social Twitter?
- Qual o impacto do Covid-19 na comunicação das HEI?

Trabalho Futuro

- Análise das respostas às publicações das HEI
- Análise do tipo de publicação efetuada (video, imagem, url, etc)

- Parte do trabalho desenvolvido nesta dissertação foi tema de um artigo cientifico aceite pela 21ª Conferência Internacional de Ciências de Computação e as Suas Aplicações (ICCSA) e consequentemente incluído na Springer Lecture Notes in Computer Science (DOI: 10.1007/978-3-030-86960-1_49) assim como indexado por Scopus, EI Engineering Index, Thomson Reuters Conference Proceedings Citation Index (incluído no ISI Web of Science), e vários outros serviços de indexação.
- Coelho, T., & Figueira, A. (2021, September). Covid-19 Impact on Higher Education Institution's Social Media Content Strategy. In *International Conference on Computational Science and Its Applications* (pp. 657-665). Springer, Cham.
- Pendente novo artigo a ser publicado na IEEE BigData 2021 conference.