Identificação de Discurso de Ódio em Redes Sociais: Um Estudo de Classificação e Aprendizado de Máquina Supervisionado

Fernanda Fornari, Leonardo da Silva Moreno

Análise e Desenvolvimento de Sistemas
Universidade Tecnológica Federal do Paraná (UTFPR) – Pato Branco, PR – Brasil

ffornari@alunos.utfpr.edu.br, leonardomoreno@alunos.utfpr.edu.br

Abstract. *With the growing impact of social media on information dissemination, it is essential to address the presence of prejudiced content, which can negatively affect individuals and groups. Furthermore, the presence of such content can compromise the quality of analyses performed in systems fueled by texts and information from social media. This situation can lead to erroneous algorithmic decisions, resulting in direct negative effects on one or more individuals. To overcome this problem, this study proposes the application of classification methods and supervised machine learning to identify texts containing discrimination or prejudice present on social media.*

Resumo. *Com o crescente impacto das redes sociais na disseminação de informações, é essencial abordar a presença de conteúdo preconceituoso, que pode afetar negativamente indivíduos e grupos. Além disso, a presença desse tipo de conteúdo pode comprometer a qualidade das análises realizadas em sistemas alimentados por textos e informações provenientes das redes sociais. Tal situação pode levar a decisões equivocadas dos algoritmos, resultando em efeitos negativos diretos em um ou mais indivíduos. Para contornar esse problema, este estudo propõe a aplicação de métodos de classificação e aprendizado de máquina supervisionado para identificar textos contendo discriminação ou preconceito presentes nas redes sociais.*

1. Introdução

Algoritmos de IA se alimentam de dados agrupados em um ou diversos conjuntos de dados que representam a fonte do conhecimento utilizado para treinar modelos inteligentes. Contudo, nem sempre esses dados são filtrados, gerando informações não confiáveis e que podem reproduzir algum tipo de preconceito, tornando o software enviesado, não confiável, e, algumas vezes, deixando de cumprir seu papel social.

Buscar a neutralidade de um algoritmo de Inteligência Artificial, ou seja, evitar o viés algorítmico, é um objetivo que muitos tentam atingir. Mittelstadt, Allo e seus parceiros do Oxford Internet Institute e Alan Turing Institute fizeram suas contribuições no artigo "The ethics of algorithms: mapping the debate" (MITTELSTADT et al., 2016), principalmente por meio de um mapa evidenciando a ética dos algoritmos, que não é uma solução para o problema, mas sim, uma sustentação para auxiliar discussões sobre o assunto. Cumprindo seu propósito, ROSSETTI e ANGELUCI (2021) fizeram, em seu artigo, seu próprio quadro (Quadro 1) apresentando sete questões éticas que são derivadas das seis preocupações citadas no trabalho de MITTELSTADT et al. (2016).

Quadro 1 – Questões éticas derivadas de preocupações citadas por MITTELSTADT *et al.* (2016)

Preocupações Éticas Trazidas por Algoritmos	**Questões Éticas Tratadas por Rossetti (2021)**
Evidências inconclusivas	Falibilidade
Evidências inextricáveis	Opacidade
Evidências mal direcionadas	Viés
Resultados injustos	Discriminação
Efeitos transformativos	Autonomia
Efeitos transformativos	Privacidade de Informações
Rastreabilidade	Responsabilidade

Fonte: Adaptado de ROSSETTI e ANGELUCI (2021), p. 7.

Algoritmos podem ser preconceituosos dependendo dos dados que são fornecidos a eles, podendo impactar negativamente na vida de diversas pessoas.

Considerando que algoritmos de IA podem expressar viés não intencional, e tomando como base os dados apresentados no Quadro 1, este trabalho terá como foco a quarta preocupação, ou seja, a discriminação causada por resultados injustos gerados por algoritmos de IA, focando-se em algortitmos que utilizam texto como dados de entrada. Assim, o objeto de estudo deste trabalho são algoritmos de análise de textos que possam ser úteis na identificação de preconceitos, evitando assim uma possível inserção desses dados em outros algoritmos de IA, melhorando a qualidade dos dados utilizados como entrada para treinamento de sistemas inteligentes, e consequentemente, a tomada de decisão.

2. Descrição da base de dados

label	id	key	text	text_len	context_name
1	1645735418938150000	aleijada	iala essas piranha puta vadia safada vagabunda estranha chata boba desgracada prostitu	280	
1	1645735092667530000	aleijada	se eu presenciar isso vou presa pq ela ia ficar aleijada de tanto q eu ia bater na nela	104	
0	1645724510304320000	aleijada	se so piora vai ficar logo e aleijada	53	
0	1645655628650630000	aleijada	guria eu tambem tenho certeza que sou toda errada mas e o medo de sair aleijada de la	99	
0	1645615543670510000	aleijada	limpe vc n e aleijada	38	
0	1645614973664590000	aleijada	prende ele e da um nome nao e bom ele ficar saindo ja vi quatro gatos morrerem atrope	171	A Gata
0	1645614170820200000	aleijada	eu n vou conseguir fzr pq eu fiquei aleijada hj	56	
1	1645604369952540000	aleijada	aleijada ainda	62	
1	1645602449305090000	aleijada	pq q eu quero saber de jogo de paraolimpiada eu nao sou aleijada polyana	81	
0	1645567789938400000	aleijada	quando q a aleijada do cca voltara ein	47	

Figura 1. Recorte da Base de Dados

A fonte de dados utilizada neste trabalho foi extraída por meio de API da rede social Twitter. Foram extraídos 2797 tweets, utilizando um filtro por palavras-chaves específicas.

A base de dados conta com 18 variáveis e 1 classe, que foram classificadas e descritas na seção 4 deste artigo.

Dentre as informações extraídas, as informações de maior destaque são a label, que classifica os textos como preconceituoso ou não preconceitos, que é o foco principal deste estudo, a variável texto, que contém o conteúdo textual que será utilizado para classificação, e a key, que é a palavra chave utilizada como filtro para obter a base de dados.

3. Distribuição de frequência

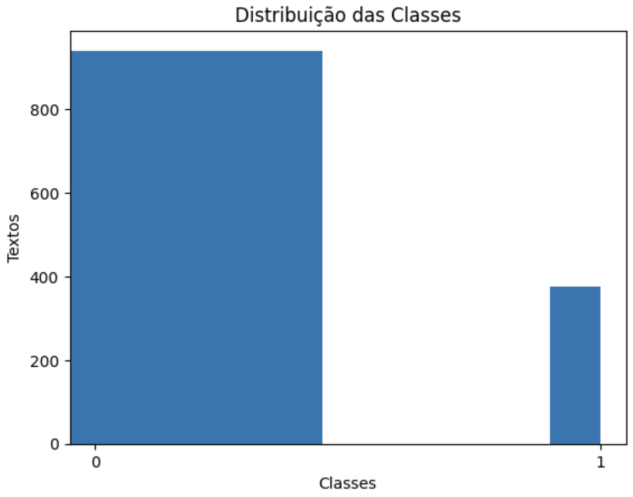

Figura 2. Distribuição das classes por quantidade

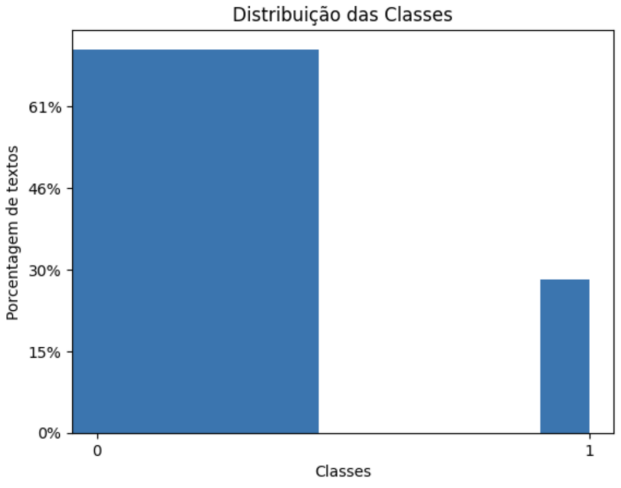

Figura 3. Percentual de Distribuição das Classes

Aproximadamente 30% dos textos analisados apresentam preconceito (classificação 1).

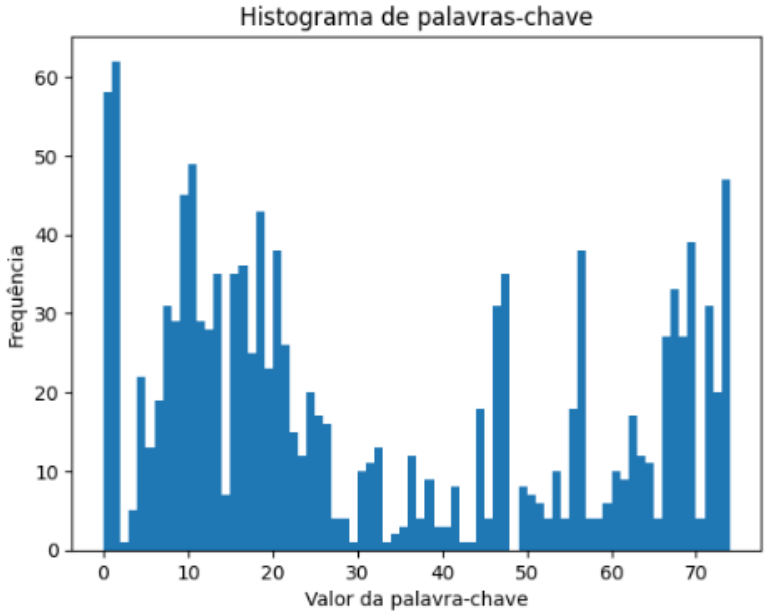

Figura 4. Distribuição das palavras-chaves

4. Tipos de dados

Os tipos de cada variável presente na base de dados, bem como suas descrições são apresentadas abaixo.

- **label (bool):** classificação (0/1):
  1 para texto preconceituoso;
  0 para texto não preconceituoso.
- **id (id):** identificador único do tweet;
- **key (text):** palavra chave utilizada na busca;
- **text (text):** texto do tweet;
- **text_len (numeric):** tamanho do texto do tweet, quantidade de caracteres;
- **context_name (text):** contexto que o tweet se encaixa;
- **context_description (text):** descrição do contexto que o tweet se encaixa;
- **context_annotations (text):** demais anotações do contexto que o tweet se encaixa;
- **impression_count (numeric):** número de visualizações(impressões) que o tweet tinha no momento da coleta;
- **like_count (numeric):** número de curtidas que o tweet tinha no momento da coleta;
- **retweet_count (numeric):** número de compartilhamentos que o tweet tinha no momento da coleta;
- **reply_count (numeric):** número de respostas que o tweet tinha no momento da coleta;
- **quote_count (numeric):** número de menções que o tweet tinha no momento da coleta;

- **has_link (bool):** Possui link no texto ou não. 1 para texto com link. 0 para texto sem link;
- **has_emoji (bool):** Possui emoji no texto ou não. 1 para texto com emoji. 0 para texto sem emoji;
- **lang (text):** Idioma do tweet. Todos serão 'pt' pois já foi filtrado;
- **created_at (date):** Data que o tweet foi postado;
- **edit_history_tweet_ids (list):** Lista de ids de edição dos tweets;
- **referenced_tweets (list):** Lista de referências do tweet.

5. Análise estatística

5.1. Cálculos estatísticos

	label	id	key	text	text_len	context_name	context_description	context_annotations	impression_count	like_count	retweet_count	reply_count	quote_count	has_link	has_emoji	lang	created_at	edit_history_tweet_ids	referenced_tweets
count	1317	1317	1317	1317	1317	332	301	5	1317	1317	1317	1317	1317	1317	1317	1317	1317	1317	1304
unique	NaN	NaN	74	1294	NaN	104	85	5	NaN	NaN	NaN	NaN	NaN	1	2	1	1292	1311	1229
top	NaN	NaN			NaN				NaN	NaN	NaN	NaN	NaN	False	False	pt			
freq	NaN	NaN	62	11	NaN	45	45	1	NaN	NaN	NaN	NaN	NaN	1317	1187	1317	3	3	16
mean	0.293090	1.646241	NaN	NaN	88.635535	NaN	NaN	NaN	127.891420	1.670463	1.084282	0.387244	0.016705	NaN	NaN	NaN	NaN	NaN	NaN
std	0.455352	1.072663	NaN	NaN	65.854767	NaN	NaN	NaN	864.812804	11.245769	26.687537	0.985652	0.159864	NaN	NaN	NaN	NaN	NaN	NaN
min	0.000000	1.642210	NaN	NaN	5.000000	NaN	NaN	NaN	0.000000	0.000000	0.000000	0.000000	0.000000	NaN	NaN	NaN	NaN	NaN	NaN
25%	0.000000	1.645743	NaN	NaN	43.000000	NaN	NaN	NaN	8.000000	0.000000	0.000000	0.000000	0.000000	NaN	NaN	NaN	NaN	NaN	NaN
50%	0.000000	1.646363	NaN	NaN	66.000000	NaN	NaN	NaN	23.000000	0.000000	0.000000	0.000000	0.000000	NaN	NaN	NaN	NaN	NaN	NaN
75%	1.000000	1.647279	NaN	NaN	109.000000	NaN	NaN	NaN	60.000000	1.000000	0.000000	1.000000	0.000000	NaN	NaN	NaN	NaN	NaN	NaN
max	1.000000	1.647571	NaN	NaN	322.000000	NaN	NaN	NaN	27323.000000	237.000000	812.000000	24.000000	4.000000	NaN	NaN	NaN	NaN	NaN	NaN

Figura 5. Cálculos estatísticos

5.2. Valores nulos

	NOT NULL	NULL	TOTAL
label	1317	1480	2797
id	2797	0	2797
key	2797	0	2797
text	2797	0	2797
text_len	2797	0	2797
context_name	818	1979	2797
context_description	743	2054	2797
context_annotations	10	2787	2797
impression_count	2797	0	2797
like_count	2797	0	2797
retweet_count	2797	0	2797
quote_count	2797	0	2797
has_link	2797	0	2797
has_emoji	2797	0	2797
lang	2797	0	2797
created_at	2797	0	2797
edit_history_tweet_ids	2797	0	2797
referenced_tweets	2755	42	2797

Figura 6. Apresentação de valores nulos presentes na base de dados

A base de dados possui 4 variáveis que possuem registros nulos. 4 variáveis que possuem grandes quantidades de registros nulos:

- label
- context_name
- context_description
- context_count

E 1 variável que possui apenas 42 registros nulos:

- referenced_tweets

Para o propósito do trabalho, preencher essas variáveis não vale a pena. Essas variáveis contém o contexto no qual o tweet está inserido, por exemplo: Política, Música, Televisão e Religião, e o que coletamos é suficiente para nos dar uma ideia de possíveis contextos onde tweets preconceituosos estão inseridos. Enriquecer essa variável exigiria modelos de aprendizado de máquina, para analisar o texto e identificar qual o assunto e o contexto que ele se encaixa, que não é o objetivo foco deste trabalho.

6. Visualização de dados

As palavras chaves com maiores porcentagem de rótulos 1 (com preconceito):

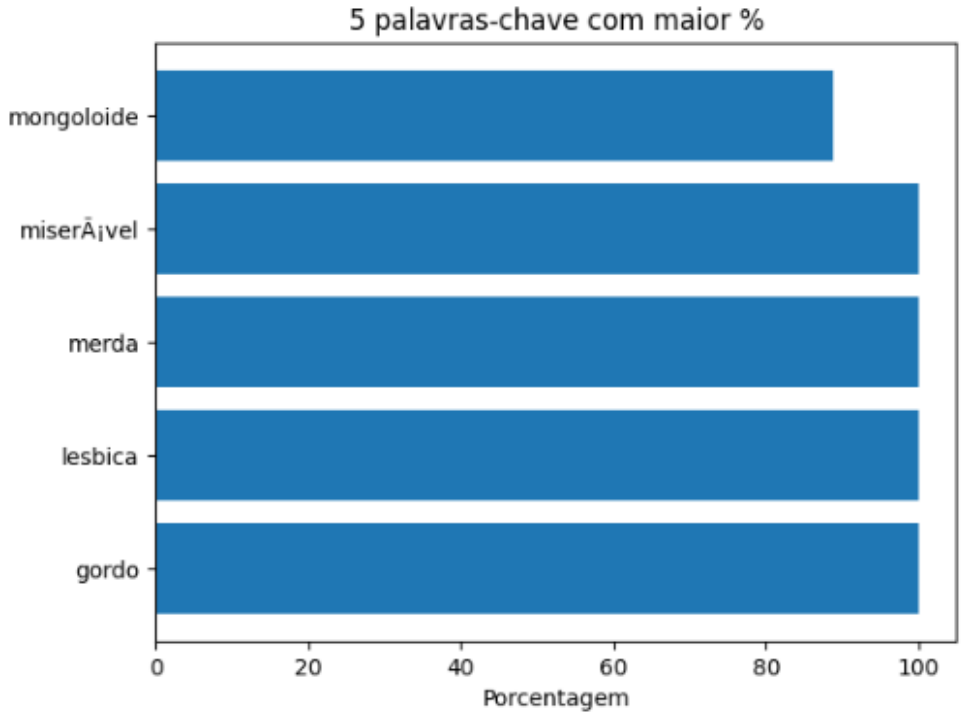

Figura 7. Palavras-chave com maior porcentagem de rótulos 1 (com preconceito)

O total de textos rotulados com cada uma dessas palavras chaves:

- gordo: 100.00% (1.0)
- lesbica: 100.00% (1.0)
- merda: 100.00% (1.0)
- miserável: 100.00% (1.0)
- mongoloide: 88.89% (18.0)

Com essas informações, podemos analisar que 4 das 5 palavras possuem apenas um único texto rotulado em cada uma delas, mostrando a escassez da base e a necessidade de mais textos rotulados com essas palavras-chaves.

A palavra mongoloide, por ser uma palavra que define um agrupamento de povos ou raça, é comumente usada em manifestações preconceituosas nas redes sociais, por isso 88.89% dos textos rotulados que contém essa palavra-chave foi considerado preconceituoso.

Devido a existência de palavras-chave com baixas ocorrências na base, o gráfico abaixo mostra as 5 maiores palavras-chaves que possuem maior quantidade de textos rotulados na base:

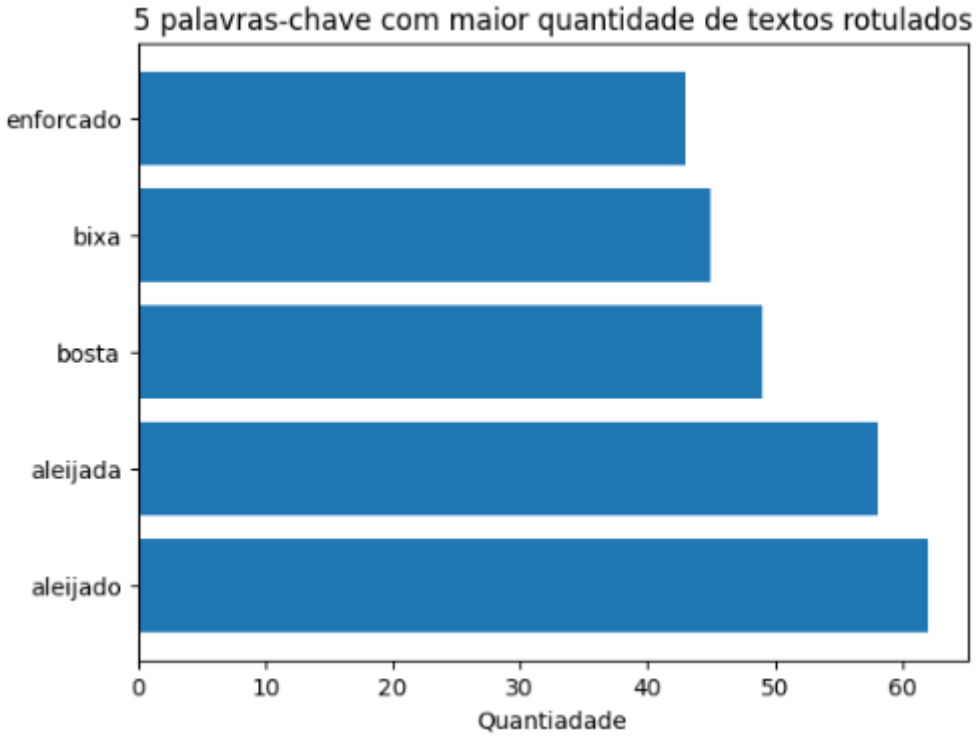

Figura 8. Palavras-chave com maior quantidade de textos rotulados

- aleijado: 62
- aleijada: 58
- bosta: 49
- bixa: 45
- enforcado: 43

Por fim, o gráfico abaixo mostra a quantidade de rótulos 0/1 por palavra-chave. Lembrando que 0 é usado para textos sem preconceito e 1 para textos com preconceito.

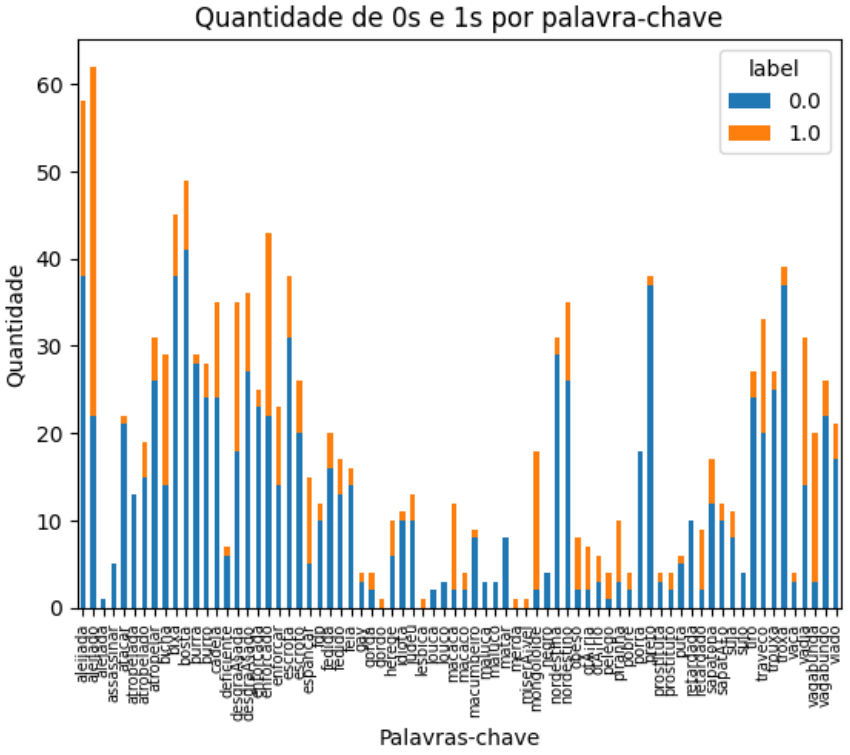

Figura 9. Distribuição dos rótulos por palavra-chave

6.1. Matriz de Correlação

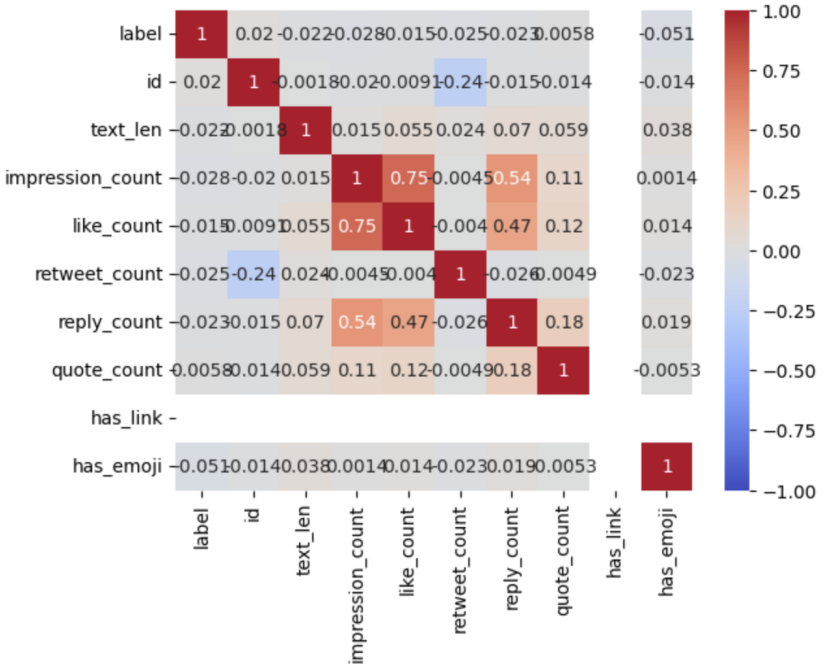

Figura 10. Matriz de correlação entre as variáveis

As variáveis numéricas que mais tem correlação são as contagens de impressão, like e reply. É comum que tweets que possuem um maior alcance (impression), tenham um maior engajamento, como um número meio de likes e reply.

6. Análise de Componentes Principais

A partir do PCA não se aplica em nosso contexto, devido a todas as variáveis presentes na base de dados serem do tipo categórica.

References

MITTELSTADT, B. D. et al. The ethics of algorithms: mapping the debate. 2016. Disponível em: https://journals.sagepub.com/doi/full/10.1177/2053951716679679. Acesso em: 20 mai. 2023.

ROSSETTI, R.; ANGELUCI, A. Ética Algorítmica: questões e desafios éticos do avanço tecnológico da sociedade da informação. 2021. Disponível em: https://www.scielo.br/j/gal/a/R9F45HyqFZMpQp9BGTfZnyr/?lang=pt&format=pdf. Acesso em: 19 mai. 2023.