Universidade Federal de São Carlos

Centro de Ciências e de Tecnologia

Departamento de Computação

Aprendizado de Máquina 2 - Trabalho 2 - Análise de Sentimentos

Professor: Murilo Coelho Naldi

Integrantes do grupo

Antonio Caique Santos, 744334, Engenharia de Computação

Jhon Wislin Ribeiro Citron, 776852, Engenharia de Computação

Lucas Henrique Marchiori, 769787, Engenharia de Computação

São Carlos, SP

2024

1. Introdução

O objetivo deste trabalho foi analisar o conjunto de dados (*Dataset*) sobre postagens feitas por diferentes usuários, de forma, a determinar o sentimento que cada mensagem está passando. Para isso, foram aplicadas técnicas de Processamento de Linguagem Natural (PLN) com o objetivo de ser interpretado o sentimento retratado em cada postagem.

2. Conjunto de dados

O conjunto de dados escolhido pelo grupo para a análise, apresenta informações relacionadas a postagens feitas por diferentes usuários de várias plataformas de mídias sócias, totalizando 732 postagens, para determinar o tipo de sentimento que está sendo passado (entusiasmo, admiração, emoção, etc). Os atributos deste *Dataset* incluem o texto da postagem, identificador, nome do usuário, data e hora da postagem, hashtags e outras informações. Esse *Dataset* pode ser consultado no link [1] das bibliografias.

3. Estudo e Tratamento do Conjunto

Todo o código aqui utilizado encontra-se disponível no *Google Colab*[1]

Inicialmente, fazemos a leitura do arquivo .csv que armazena os dados do dataset (sentimentdataset.csv). Após ser feita a leitura, o conjunto de dados pode ser visualizado por meio do comando *df.head(),* que apresenta as 5 primeiras linhas ou dados do dataset. Abaixo (Figura 1), é possível visualizar o conjunto de dados.

	Unnamed: 0.1	Unnamed: 0	Text	Sentiment	Timestamp	User	Platform	Hashtags	Retweets	Likes	Country	Year	Month	Day	Hour
0	0	0	Enjoying a beautiful day at the park! ...	Positive	2023-01-15 12:30:00	User123	Twitter	#Nature #Park	15.0	30.0	USA	2023	1	15	12
1	1	1	Traffic was terrible this morning. ...	Negative	2023-01-15 08:45:00	CommuterX	Twitter	#Traffic #Morning	5.0	10.0	Canada	2023	1	15	8
2	2	2	Just finished an amazing workout! 💪 ...	Positive	2023-01-15 15:45:00	FitnessFan	Instagram	#Fitness #Workout	20.0	40.0	USA	2023	1	15	15
3	3	3	Excited about the upcoming weekend getaway! ...	Positive	2023-01-15 18:20:00	AdventureX	Facebook	#Travel #Adventure	8.0	15.0	UK	2023	1	15	18
4	4	4	Trying out a new recipe for dinner tonight. ...	Neutral	2023-01-15 19:55:00	ChefCook	Instagram	#Cooking #Food	12.0	25.0	Australia	2023	1	15	19

1 https://colab.research.google.com/drive/1VyB4ZmHOAMrNSQZweYqmWYvAAq0ia3fd?usp=sharing#scrollTo=MlW-TU_GkKvF

Figura 1: Dataset

Além disto, outras formas de visualização que ajudaram antes da aplicação do aprendizado de máquina são realizadas através da biblioteca *ydata-profiling*[2] onde a mesma apresenta diversas maneiras de visualização de dados, e de análise exploratória de dados.

Uma das primeiras características extraídas dessa análise são as estatísticas do *dataset utilizado*, ilustrado abaixo (Figura 2). Algo importante a se atentar é que não existem valores faltantes, não ocorrendo necessidade de um tratamento para preencher valores vazios. Outro ponto a se observar é que dos 13 atributos, 5 deles são textos, 6 são numéricos, 1 é do tipo data e 1 categórico. Além disso, é possível notar que existem 19 linhas duplicadas.

Dataset statistics

Number of variables	13
Number of observations	732
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	19
Duplicate rows (%)	2.6%
Total size in memory	74.5 KiB
Average record size in memory	104.2 B

Variable types

Text	5
DateTime	1
Categorical	1
Numeric	6

Figura 2: Estatísticas do Dataset.

Outra estatística apresentada é uma matriz de correlação, onde apresenta a relação entre todos os atributos do Dataset, aplicando uma escala de cores de -1 a 1. Observando a matriz e analisando as correlações entre os atributos, foi possível observar (Figura 3) que likes e Retweets possuem uma correlação de basicamente 100%.

[2] https://ydata-profiling.ydata.ai/docs/master/

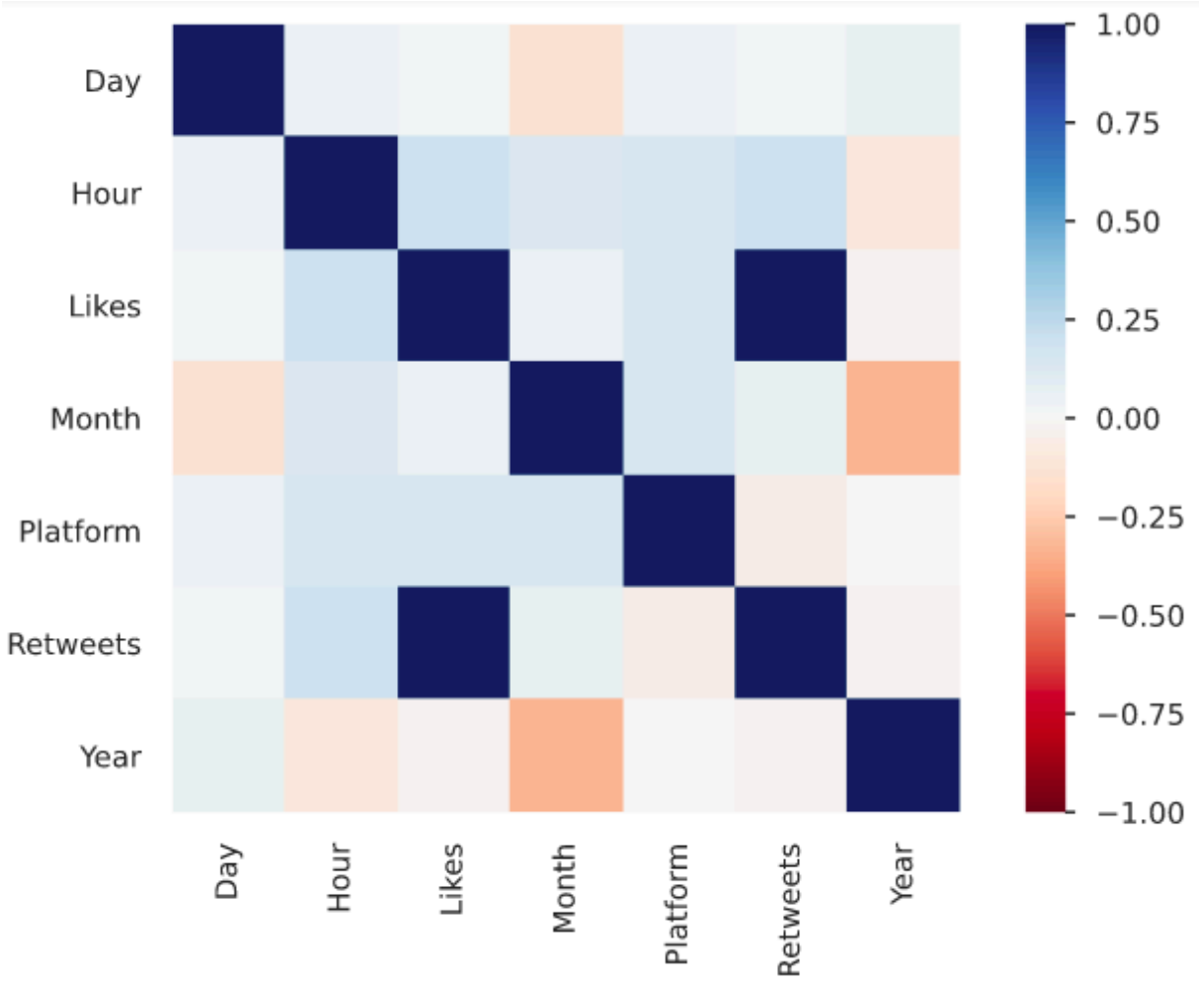

Figura 3: Heatmap de Correlação.

Por conseguinte, filtramos as classes de sentimentos que possuem uma frequência menor que 5, e aplicamos a classificação de “Others” para elas, para reduzir o número de sentimentos que podem ser classificados, mantendo apenas os mais relevantes. Com isso, abaixo (Figura 4), é possível observar o número de dados para cada sentimento, onde, “Others” ficou com 274 dados pertencentes a ele, e apenas os sentimentos com um número de dados maior ou igual a 5 foi mantido.

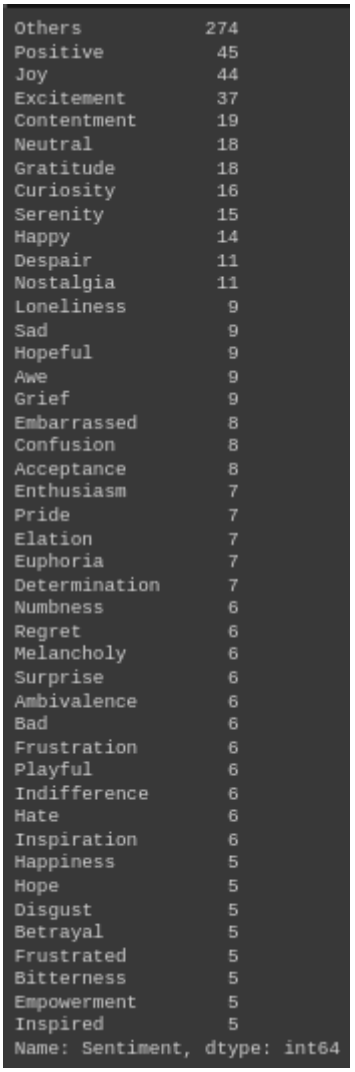

Figura 4: Quantidade de dados por sentimento

Após isso, é feito um mapeamento entre os rótulos da classe sentimento em inteiros utilizando o dicionário class_to_int. Assim, para cada sentimento foi atribuído um valor inteiro de 0 a 43, sendo 44 o número total de sentimentos. Esse mapeamento foi passado para o atributo “Label”. Esse processo foi feito para facilitar o treinamento dos modelos que necessitam que os rótulos sejam numéricos. Após esse processo, é feita a plotagem do gráfico (Figura 5) da frequência de cada um dos sentimentos. Observando o gráfico, é possível notar que “Others” é o sentimento que possui uma maior frequência, isso devido a ele ser um aglomerado de sentimentos. Desconsiderando “Others”, é possível notar que os três sentimentos mais frequentes são, alegria, positivo e excitação.

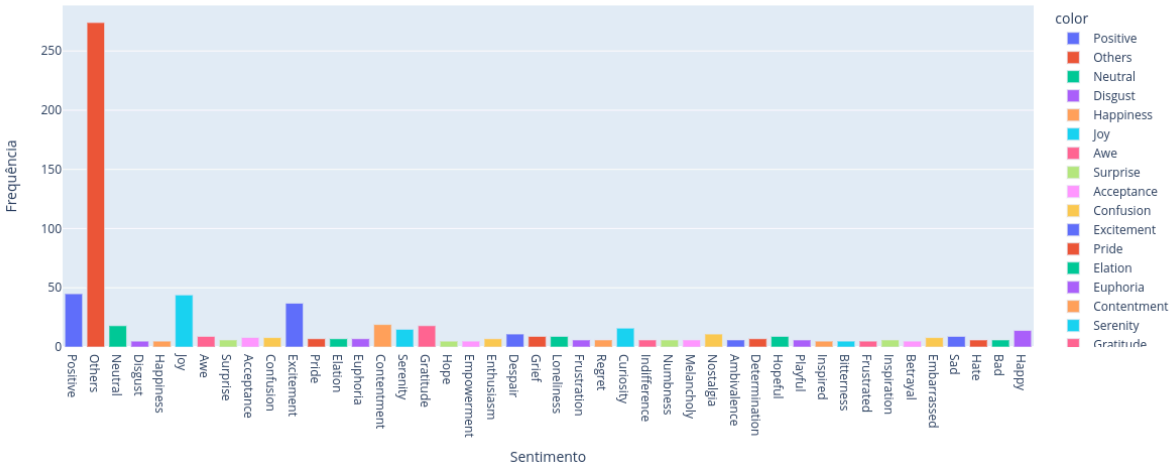

Figura 5: Frequência dos sentimentos

Abaixo (Figura 6), é possível observar um histograma em relação à distribuição do tamanho dos textos. Observado o histograma, é possível notar que a maioria dos textos postados (97 postagens) possuíam tamanhos em torno de 50 a 54. Avaliando o histograma de forma geral, é possível notar que a medida que o texto é maior, a quantidade de postagens se torna menor.

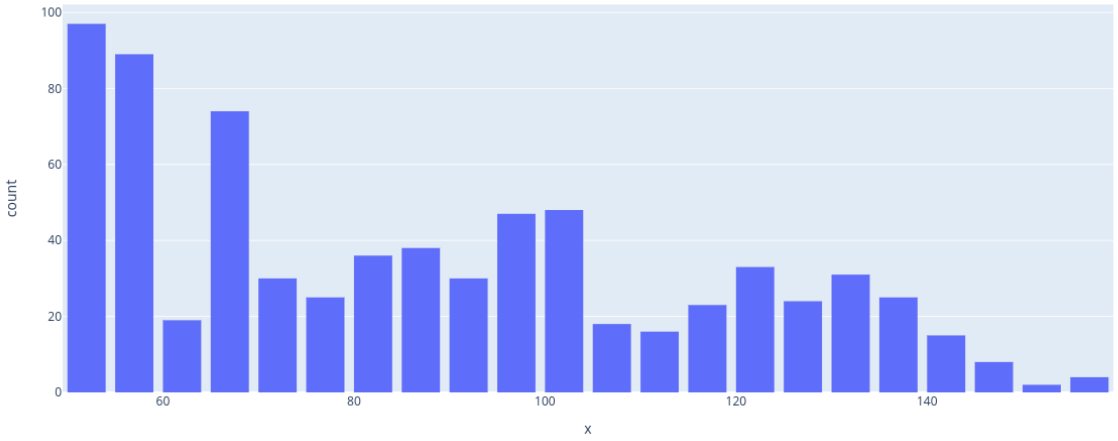

Figura 6: Distribuição de tamanho

Em seguida, para preparar os dados para ser utilizado o PLN, é inicialmente definido o modelo BERT e carregado o tokenizador (tokenizer) o qual é responsável por dividir o texto em tokens. Após isso, os atributos “text” e “label” são selecionados em relação ao dataset original e o conjunto de dados é dividido em conjunto de treinamento (train_df), validação (val_df) e teste (test_df).

Por conseguinte, o tokenizador é aplicado para dividir os textos em tokens, onde, os IDs (input_ids_train e input_ids_val), máscaras de atenção (attention_masks_train e attention_masks_val) e rótulos dos dados (labels_train e labels_val) tokenizados são extraídos e criados dois datasets (dataset_train e dataset_val). Em seguida, são criados dataloaders (dataloader_train e dataloader_validation) utilizando as classes RandomSampler e SequentialSampler para amostragem aleatória e sequencial.

A classe **SentimentClassifier** é definida, ao qual utilizou a arquitetura BERT pré-treinada disponibilizada pelo hugging face[3]. Essa classe adicionou uma camada linear para produzir as saídas de classificação e uma camada de dropout para regularização. Além disso, vale ser ressaltado que o número de neurônios foi definido para 44 (Mesmo número de classes de sentimento). O treinamento foi feito por meio de um loop de épocas, onde, o otimizador **AdamW** foi utilizado para fazer o ajuste dos pesos do modelo. Após esse processo, a função **eval_model** avaliou o desempenho do modelo no conjunto de validação após a ocorrência de cada época. Durante o treinamento, um agendador linear com warm-up foi utilizado para controlar a taxa de aprendizado, e a função **train_model** atuou sobre o conjunto de treinamento ajustando os gradientes, fazendo a atualização dos pesos do modelo e propagação. Com o intuito de se encontrar o melhor estado do modelo, para cada época, o modelo era salvo caso sua acurácia fosse superior à melhor obtida até o momento atual. Neste caso, a melhor acurácia foi de 0,55.

Por fim, após os processos anteriores, foi gerado um relatório de classificação (Figura 7) que apresenta as métricas de precisão, revocação, suporte e f1-score em relação a cada sentimento. Observando a imagem abaixo (Figura 7), é possível notar que boa parte das métricas obteve valor 0, isso é devido ao fato que o número de ocorrência desses sentimentos foi baixo. Vale ser ressaltado que apenas os sentimentos com IDs 5, 17, 20 e 40 obtiveram uma precisão de 100%.

[3] https://huggingface.co/

	precision	recall	f1-score	support
0	0.05	0.67	0.10	27
1	0.98	0.88	0.93	186
2	0.00	0.00	0.00	0
3	0.00	0.00	0.00	0
4	0.00	0.00	0.00	0
5	1.00	0.20	0.41	60
6	0.00	0.00	0.00	0
7	0.00	0.00	0.00	0
8	0.25	0.17	0.20	6
9	0.60	0.50	0.55	6
10	0.00	0.00	0.00	3
12	0.75	0.60	0.67	5
13	0.50	1.00	0.67	2
14	0.00	0.00	0.00	0
15	0.00	0.00	0.00	0
16	0.00	0.00	0.00	0
17	1.00	1.00	1.00	1
18	0.00	0.00	0.00	0
19	0.50	1.00	0.67	1
20	1.00	0.21	0.35	14
21	0.00	0.00	0.00	0
22	0.00	0.00	0.00	3
23	0.00	0.00	0.00	0
24	0.00	0.00	0.00	1
25	0.00	0.00	0.00	0
26	0.00	0.00	0.00	0
27	0.00	0.00	0.00	2
28	0.00	0.00	0.00	0
29	0.00	0.00	0.00	0
30	0.00	0.00	0.00	0
31	0.00	0.00	0.00	1
32	0.00	0.00	0.00	0
33	0.00	0.00	0.00	0
34	0.00	0.00	0.00	0
35	0.00	0.00	0.00	0
36	0.00	0.00	0.00	0
37	0.00	0.00	0.00	2
38	0.00	0.00	0.00	1
39	0.00	0.00	0.00	0
40	1.00	0.40	0.57	5
41	0.33	0.50	0.40	2
42	0.00	0.00	0.00	3
43	0.00	0.00	0.00	0
accuracy			0.50	257
macro avg	0.21	0.17	0.17	257
weighted avg	0.88	0.50	0.64	257

Figura 7: Relatório de classificação

Os processos descritos anteriormente são repetidos, mas sem ser feita nenhuma filtragem em relação à frequência para cada sentimento. Dessa forma, a classe “Others” não foi criada, assim sendo considerando todos os sentimentos totalizando 191, como é possível observar na figura abaixo (Figura 8). Como “Others” não foi criado, o sentimento com mais frequência é o positivo, com 45 postagens classificadas.

```
Positive          45
Joy               44
Excitement        37
Contentment       19
Neutral           18
                  ..
LostLove           1
EmotionalStorm     1
Suffering          1
Bittersweet        1
Intrigue           1
Name: Sentiment, Length: 191, dtype: int64
```

Figura 8: Frequência sem filtragem

Abaixo (Figura 9), é possível observar um novo gráfico para a frequência dos sentimentos, onde é possível visualizar de uma melhor forma quais sentimentos estão mais e menos frequentes nas postagens, além de apresentar todos os sentimentos sem estarem filtrados em “Others”.

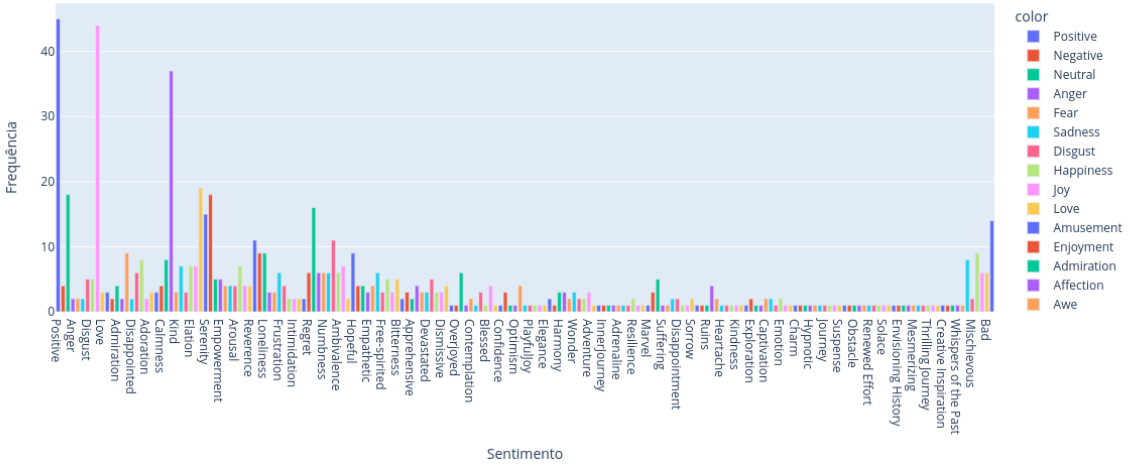

Figura 9: Frequência dos sentimentos (Sem Filtro)

Por fim, após os processos anteriores, foi gerado um relatório de classificação novamente (Figura 10) que apresenta as métricas de precisão, revocação, suporte e f1-score em relação a cada sentimento. Observando a imagem abaixo (Figura 10), é possível notar que praticamente todos os resultados obtidos foram 0, além disso, a acurácia obtida foi extremamente baixa (0,0077). Vale ser ressaltado que a presença de muitas classes apenas diminui a acurácia para 0.

150	0.00	0.00	0.00	18
152	0.00	0.00	0.00	0
154	0.00	0.00	0.00	0
157	0.00	0.00	0.00	0
158	0.00	0.00	0.00	0
159	0.00	0.00	0.00	3
166	0.00	0.00	0.00	0
171	0.00	0.00	0.00	0
173	0.00	0.00	0.00	0
174	0.00	0.00	0.00	0
179	0.00	0.00	0.00	0
183	0.00	0.00	0.00	0
184	0.00	0.00	0.00	0
185	0.00	0.00	0.00	0
186	0.00	0.00	0.00	0
187	0.00	0.00	0.00	0
188	0.00	0.00	0.00	0
189	0.00	0.00	0.00	0
190	0.00	0.00	0.00	0
accuracy			0.00	257
macro avg	0.00	0.00	0.00	257
weighted avg	0.05	0.00	0.01	257

Figura 11: Relatório de classificação (Sem Filtragem)

Por fim, os processos anteriores foram novamente repetidos, fazendo uma filtragem das classes para aquelas com uma frequência maior ou igual a 15, como é possível observar abaixo (Figura 12). É possível notar que após a filtragem apenas 8 classes restaram.

```
Positive       45
Joy            44
Excitement     37
Contentment    19
Neutral        18
Gratitude      18
Curiosity      16
Serenity       15
Name: Sentiment, dtype: int64
```

Figura 12: Frequência igual ou maior que 15

Assim, conforme todo o processo foi aplicado novamente, foi obtida uma acurácia de 0,25 como sendo a melhor após a passagem das épocas. Além disso, o relatório de classificação foi novamente gerado (Figura 13), onde, é possível notar que os resultados obtidos não foram muito diferentes em relação ao sem filtro.

0	0.75	0.25	0.38	48
1	0.00	0.00	0.00	0
2	0.00	0.00	0.00	0
3	0.00	0.00	0.00	0
4	0.00	0.00	0.00	21
5	0.00	0.00	0.00	1
6	0.00	0.00	0.00	5
7	0.00	0.00	0.00	0
accuracy			0.16	75
macro avg	0.09	0.03	0.05	75
weighted avg	0.48	0.16	0.24	75

Figura 13: Relatório de classificação (Filtragem para maior ou igual a 15)

4. Conclusão

Após a aplicação em diferentes configurações para a filtragem da frequência dos sentimentos, foi possível determinar que utilizar todas as classes de sentimento sem nenhuma filtragem acabou reduzindo a acurácia, isso devido ao fato que a dimensão do problema ficou muito grande. Em contrapartida, utilizar como filtro uma frequência de sentimentos muito alta não foi muito benéfico pelo fato que boa parte das labels foram adicionados em “Others” então houve uma discrepância muito grande entre cada label. Dessa forma, a primeira configuração com uma filtragem de frequência para maior ou igual a 5, obteve a melhor acurácia (0,55), isso devido ao fato que apenas os sentimentos menos frequentes foram filtrados.

5. Bibliografia

[1]
https://www.kaggle.com/datasets/kashishparmar02/social-media-sentiments-analysis-dataset?resource=download