```
from google.colab import drive
drive.mount('/content/drive')

     Mounted at /content/drive
```

Teste com pandas

```
import pandas as pd
```

```
dataset = pd.read_csv("/content/drive/MyDrive/TCC_BANCOS/online-misogyny-eacl2021-main/dat
```

```
dataset.groupby('level_1')['level_1'].count()

     level_1
     Misogynistic        699
     Nonmisogynistic    5868
     Name: level_1, dtype: int64
```

Fim do teste com pandas

```
!pip install pyspark

     Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/p
     Collecting pyspark
       Downloading pyspark-3.3.1.tar.gz (281.4 MB)
          |████████████████████████████████| 281.4 MB 47 kB/s
     Collecting py4j==0.10.9.5
       Downloading py4j-0.10.9.5-py2.py3-none-any.whl (199 kB)
          |████████████████████████████████| 199 kB 52.8 MB/s
     Building wheels for collected packages: pyspark
       Building wheel for pyspark (setup.py) ... done
       Created wheel for pyspark: filename=pyspark-3.3.1-py2.py3-none-any.whl size=281845
       Stored in directory: /root/.cache/pip/wheels/42/59/f5/79a5bf931714dcd201b26025347
     Successfully built pyspark
     Installing collected packages: py4j, pyspark
     Successfully installed py4j-0.10.9.5 pyspark-3.3.1
```

```
from pyspark.sql import SparkSession

spark = SparkSession.builder \
    master('local[*]') \
```

Salvo com sucesso

```
dados = spark.read.csv("/content/drive/MyDrive/TCC_BANCOS/online-misogyny-eacl2021-main/da
                       escape = '\"',
                       inferSchema= True,
                       header = True)
```

```
dados.show()
```

```
+-------------+------------------+------------------+----------+----------------
|     entry_id|           link_id|         parent_id| entry_utc|       subreddit
+-------------+------------------+------------------+----------+----------------
|       exoxn7|         t3_exoxn7|              null|1580652620|badwomensanatomy
|      fgb3bdv|         t3_exoxn7|         t3_exoxn7|1580658139|badwomensanatomy
|Stay hydrated|      it's healthy| you'll look and ...|      null|      17-02-2020
|      fgc6tlu|         t3_exoxn7|         t3_exoxn7|1580669695|badwomensanatomy
|      fge6msg|         t3_exoxn7|        t1_fgc6tlu|1580692566|badwomensanatomy
|      fgawus5|         t3_exoxn7|         t3_exoxn7|1580656280|badwomensanatomy
|      fgctirr|         t3_exoxn7|        t1_fgawus5|1580676096|badwomensanatomy
|    Obviously| the people from ...|              null|17-02-2020|               1
|      fgdomwf|         t3_exoxn7|        t1_fgctirr|1580684792|badwomensanatomy
|      fgbwoi5|         t3_exoxn7|         t3_exoxn7|1580666780|badwomensanatomy
|      fgbxtc0|         t3_exoxn7|        t1_fgbwoi5|1580667138|badwomensanatomy
|      fgdmluh|         t3_exoxn7|         t3_exoxn7|1580684099|badwomensanatomy
|      fgdog3k|         t3_exoxn7|        t1_fgdmluh|1580684716|badwomensanatomy
|      fgdqj28|         t3_exoxn7|        t1_fgdog3k|1580685515|badwomensanatomy
|      fgdowdc|         t3_exoxn7|         t3_exoxn7|1580684907|badwomensanatomy
|      fgay2nh|         t3_exoxn7|         t3_exoxn7|1580656591|badwomensanatomy
|      fgdy0kw|         t3_exoxn7|         t3_exoxn7|1580688242|badwomensanatomy
|      fgj8sxn|         t3_exoxn7|         t3_exoxn7|1580839237|badwomensanatomy
|       exuuxj|         t3_exuuxj|              null|1580676217|badwomensanatomy
|      fgcul27|         t3_exuuxj|         t3_exuuxj|1580676360|badwomensanatomy
+-------------+------------------+------------------+----------+----------------
only showing top 20 rows
```

```
dados=dados[['body','level_1']]
dados.show()
```

```
+--------------------+---------------+
|                body|        level_1|
+--------------------+---------------+
|Do you have the s...|Nonmisogynistic|
|This is taking a ...|           null|
|                   1|           null|
|Honestly my favor...|Nonmisogynistic|
|Source? Doesnt so...|Nonmisogynistic|
|Damn, I saw a mov...|   Misogynistic|
|It's a question o...|           null|
|           (1, 3, 1)|          train|
|Some places have ...|Nonmisogynistic|
|So if I drink eno...|Nonmisogynistic|
|You'll Benjamin B...|Nonmisogynistic|
|Isn't this the pl...|Nonmisogynistic|
|No idea. I watche...|Nonmisogynistic|
                             nistic|
                             nistic|
|What kinda Tom Br...|Nonmisogynistic|
|  Recommended by F.F|Nonmisogynistic|
|Professionals say...|Nonmisogynistic|
|     Virgina Spread |Nonmisogynistic|
|*I can't believe ...|Nonmisogynistic|
+--------------------+---------------+
```

Salvo com sucesso ×

```
only showing top 20 rows
```

```
print(f'Nº de Linhas:{dados.count()}', f'Nº de colunas:{len(dados.columns)}')
```

```
Nº de Linhas:15189 Nº de colunas:2
```

```
dados.printSchema()
```

```
root
 |-- body: string (nullable = true)
 |-- level_1: string (nullable = true)
```

```
print("Negativo")
dados.filter(dados.level_1 == "Nonmisogynistic").select("body", "level_1").show(truncate =
print("Positivo")
dados.filter(dados.level_1 == "Misogynistic").select("body", "level_1").show(truncate = Fa
```

```
Negativo
+--------------------------------------------------------------------------------------
|body
+--------------------------------------------------------------------------------------
|Do you have the skin of a 80 year old grandma? Worry no more, just drink water!
|Honestly my favorite thing about this is that they feel the need to cite beauty prof
|Source? Doesnt sound right to me idk
|Some places have poor water quality. Virgin blood may have been less expensive than
|So if I drink enough water I turn into a baby?
|You'll Benjamin Button yourself, yes.
|Isn't this the plot of Cocoon?
|No idea. I watched the movie 10 years or so ago with my dad and all I can remember a
|In the movie there is this group of elderly that discover that they feel suspiciousl
|One of my friends looks like he's frikkin dead. Been trying for years to pour water
|What kinda Tom Brady nonsense is this
|Recommended by F.F
|Professionals say, that dehydration is caused by not drinking enough water. In other
|Virgina Spread
|*I can't believe it's not* virgina spread open
|I hate you because its the middle of the night and Im silently suffering from laught
|Even if you chug a handle of bad vodka and give this person the benefit of the doubt
|I'll need *way* more than vodka to make sense of this shit
|I think I have a couple more handles in the garage but they're warm, that okay?
|Well now the secrets out
+--------------------------------------------------------------------------------------
only showing top 20 rows

Positivo
Salvo com sucesso    ---------------------------------------------------------------------

                     ---------------------------------------------------------------------
|Damn, I saw a movie in which the old woman bathed in the blood if virgins  to do thi
|There's no way he would want to have sex with those ugly ghetto whores.One of them c
|There's no way he would want to have sex with those ugly ghetto whores.One of them c
|If I knew how to I would open my own subreddit called... Men boycotting woke women.
|Lol I can tell your just a mad women who thinks she is entitled to marriage. Awwww h
|You mean women marry and divorce for financial gain??? Never! /s
```

```
|HAHAHAHAHAHAHAHAHAHAHAHA stupid cunt
|They want our money and peace of mind. I say, come and get it!
|Idk any woman who was sexually assaulted, but I do know a couple of men who were vic
|I literally have to videotape every moment I spend with my daughter because of all c
|If a woman punches herself before lying to the police youre actually better off kil
|And why's that? Your comment strikes me as very integrated feminist TBH. As all fasc
|fucking cunts. that's why tv is dead
|Then, they display the exact opposite of even *those*. Showing the kind of hatred ar
|Isn't this what they wanted ? To kill "traditional gender roles" or it doesn't work
|There aren't many things that are more satisfying than telling a girl, "No."
|It's funny to see the hamster that starts to act up in their little widdle tiny brai
|Let them cry "where have all the good men gone!?" Fucking dumb whores!
|More often than not truth isn't coming out or does way too late. This guy already go
|Gold diggers gotta dig...
+-----------------------------------------------------------------------------------
only showing top 20 rows
```

Limpeza dos dados nulos

```
dados = dados.dropna(subset="body")
dados.show()
```

```
+--------------------+---------------+
|                body|        level_1|
+--------------------+---------------+
|Do you have the s...|Nonmisogynistic|
|This is taking a ...|           null|
|                   1|           null|
|Honestly my favor...|Nonmisogynistic|
|Source? Doesnt so...|Nonmisogynistic|
|Damn, I saw a mov...|   Misogynistic|
|It's a question o...|           null|
|           (1, 3, 1)|          train|
|Some places have ...|Nonmisogynistic|
|So if I drink eno...|Nonmisogynistic|
|You'll Benjamin B...|Nonmisogynistic|
|Isn't this the pl...|Nonmisogynistic|
|No idea. I watche...|Nonmisogynistic|
|In the movie ther...|Nonmisogynistic|
|One of my friends...|Nonmisogynistic|
|What kinda Tom Br...|Nonmisogynistic|
|  Recommended by F.F|Nonmisogynistic|
|Professionals say...|Nonmisogynistic|
|     Virgina Spread |Nonmisogynistic|
|*I can't believe ...|Nonmisogynistic|
+--------------------+---------------+
only showing top 20 rows
```

Salvo com sucesso ×

```
print(f'N° de Linhas:{dados.count()}', f'N° de colunas:{len(dados.columns)}')
```

```
N° de Linhas:9005 N° de colunas:2
```

```
dados = dados.dropna(subset="level_1")
```

```
dados.show()
```

```
+--------------------+---------------+
|                body|        level_1|
+--------------------+---------------+
|Do you have the s...|Nonmisogynistic|
|Honestly my favor...|Nonmisogynistic|
|Source? Doesnt so...|Nonmisogynistic|
|Damn, I saw a mov...|   Misogynistic|
|           (1, 3, 1)|          train|
|Some places have ...|Nonmisogynistic|
|So if I drink eno...|Nonmisogynistic|
|You'll Benjamin B...|Nonmisogynistic|
|Isn't this the pl...|Nonmisogynistic|
|No idea. I watche...|Nonmisogynistic|
|In the movie ther...|Nonmisogynistic|
|One of my friends...|Nonmisogynistic|
|What kinda Tom Br...|Nonmisogynistic|
|  Recommended by F.F|Nonmisogynistic|
|Professionals say...|Nonmisogynistic|
|     Virgina Spread |Nonmisogynistic|
|*I can't believe ...|Nonmisogynistic|
|I hate you becaus...|Nonmisogynistic|
|Even if you chug ...|Nonmisogynistic|
|I'll need *way* m...|Nonmisogynistic|
+--------------------+---------------+
only showing top 20 rows
```

```
print(f'N° de Linhas:{dados.count()}', f'N° de colunas:{len(dados.columns)}')
```

```
N° de Linhas:5226 N° de colunas:2
```

Verificar como limpa uma coluna com dados específicos, para a coluna Body quando existir o número '1' e para a coluna level_1 quando existir a palavra 'train'

```
dados = dados.filter(dados.body != "1").select("body", "level_1")
dados.show(truncate = False)
```

```
+------------------------------------------------------------------------------------------
|body
+------------------------------------------------------------------------------------------
|Do you have the skin of a 80 year old grandma? Worry no more, just drink water!
|Honestly my favorite thing about this is that they feel the need to cite beauty prof
|Source? Doesnt sound right to me idk
|Damn, I saw a movie in which the old woman bathed in the blood if virgins  to do thi
|(1, 3, 1)
                              ality. Virgin blood may have been less expensive than
                              urn into a baby?
                              f, yes.
|Isn't this the plot of Cocoon?
|No idea. I watched the movie 10 years or so ago with my dad and all I can remember a
|In the movie there is this group of elderly that discover that they feel suspiciousl
|One of my friends looks like he's frikkin dead. Been trying for years to pour water
|What kinda Tom Brady nonsense is this
|Recommended by F.F
```

Salvo com sucesso ×

```
|Professionals say, that dehydration is caused by not drinking enough water. In other
|Virgina Spread
|*I can't believe it's not* virgina spread open
|I hate you because its the middle of the night and Im silently suffering from laught
|Even if you chug a handle of bad vodka and give this person the benefit of the doubt
|I'll need *way* more than vodka to make sense of this shit
+------------------------------------------------------------------------------------
only showing top 20 rows
```

```
dados = dados.filter(dados.level_1 != "train").select("body", "level_1")
dados.show(truncate = False)
```

```
+------------------------------------------------------------------------------------
|body
+------------------------------------------------------------------------------------
|Do you have the skin of a 80 year old grandma? Worry no more, just drink water!
|Honestly my favorite thing about this is that they feel the need to cite beauty prof
|Source? Doesnt sound right to me idk
|Damn, I saw a movie in which the old woman bathed in the blood if virgins  to do thi
|Some places have poor water quality. Virgin blood may have been less expensive than
|So if I drink enough water I turn into a baby?
|You'll Benjamin Button yourself, yes.
|Isn't this the plot of Cocoon?
|No idea. I watched the movie 10 years or so ago with my dad and all I can remember a
|In the movie there is this group of elderly that discover that they feel suspiciousl
|One of my friends looks like he's frikkin dead. Been trying for years to pour water
|What kinda Tom Brady nonsense is this
|Recommended by F.F
|Professionals say, that dehydration is caused by not drinking enough water. In other
|Virgina Spread
|*I can't believe it's not* virgina spread open
|I hate you because its the middle of the night and Im silently suffering from laught
|Even if you chug a handle of bad vodka and give this person the benefit of the doubt
|I'll need *way* more than vodka to make sense of this shit
|I think I have a couple more handles in the garage but they're warm, that okay?
+------------------------------------------------------------------------------------
only showing top 20 rows
```

```
print(f'N° de Linhas:{dados.count()}', f'N° de colunas:{len(dados.columns)}')
```

```
N° de Linhas:4783 N° de colunas:2
```

```
dados.filter(dados.level_1 == "Nonmisogynistic").select("body", "level_1").groupBy('level_
dados.filter(dados.level_1 == "Misogynistic").select("body", "level_1").groupBy('level_1')
```

Salvo com sucesso ×

```
|level_1        |count|
+---------------+-----+
|Nonmisogynistic|4246 |
+---------------+-----+

+------------+-----+
|level_1     |count|
+-----------+-----+
|Misogynistic|317  |
+-----------+-----+
```

Criação de uma coluna índice (index)

```
from pyspark.sql import SparkSession, functions as F
from pyspark import SparkConf
conf = SparkConf()

spark = SparkSession.builder.config(conf=conf).appName('Dataframe with Indexes').getOrCrea


df = dados

rdd_df = df.rdd.zipWithIndex()
df_final = rdd_df.toDF()

df_final = df_final.withColumn('body', df_final['_1'].getItem("body"))
df_final = df_final.withColumn('level_1', df_final['_1'].getItem("level_1"))


df_final = df_final.withColumnRenamed("_2","index")


dados=df_final[['index','body','level_1']]
dados.show()
```

```
+-----+--------------------+---------------+
|index|                body|        level_1|
+-----+--------------------+---------------+
|    0|Do you have the s...|Nonmisogynistic|
|    1|Honestly my favor...|Nonmisogynistic|
|    2|Source? Doesnt so...|Nonmisogynistic|
|    3|Damn, I saw a mov...|   Misogynistic|
|    4|Some places have ...|Nonmisogynistic|
|    5|So if I drink eno...|Nonmisogynistic|
|    6|You'll Benjamin B...|Nonmisogynistic|
|    7|Isn't this the pl...|Nonmisogynistic|
|    8|No idea. I watche...|Nonmisogynistic|
|    9|In the movie ther...|Nonmisogynistic|
|   10|One of my friends...|Nonmisogynistic|
|   11|What kinda Tom Br...|Nonmisogynistic|
|   12|  Recommended by F.F|Nonmisogynistic|
|   13|Professionals say...|Nonmisogynistic|
|   14|     Virgina Spread |Nonmisogynistic|
                               misogynistic|
                               misogynistic|
                               misogynistic|
|   18|I'll need *way* m...|Nonmisogynistic|
|   19|I think I have a ...|Nonmisogynistic|
+-----+--------------------+---------------+
only showing top 20 rows
```

Salvo com sucesso ×

Tentativas de groupBy

```
dados\
    .select('level_1')\
    .groupBy('level_1')\
    .count()\
    .show()
```

```
+--------------------+-----+
|             level_1|count|
+--------------------+-----+
| which they would...|    1|
| feel BIGGER than...|    1|
|           (5, 2, 1)|    1|
|             (27, 3)|    1|
|                   8|    1|
|My primary slave ...|    1|
| BLS and some of ...|    1|
|Sexual_or_physica...|    3|
| I KNEW that I wa...|    1|
| you'll be abused...|    1|
|So of course she ...|    1|
|           Hypergamy|    1|
|               Stacy|    1|
| instilling certa...|    1|
| thin with a bit ...|    5|
|I have even had g...|    1|
| I didn't take my...|    1|
|               sluts|    2|
|Luckily, feminist...|    1|
|           (3, 1, 1)|    1|
+--------------------+-----+
only showing top 20 rows
```

```
dados.groupBy('level_1').count().show()
```

```
+--------------------+-----+
|             level_1|count|
+--------------------+-----+
| which they would...|    1|
| feel BIGGER than...|    1|
|           (5, 2, 1)|    1|
|             (27, 3)|    1|
|                   8|    1|
|My primary slave ...|    1|
| BLS and some of ...|    1|
```

Salvo com sucesso ×

```
| you'll be abused...|    1|
|So of course she ...|    1|
|           Hypergamy|    1|
|               Stacy|    1|
| instilling certa...|    1|
| thin with a bit ...|    5|
|I have even had g...|    1|
| I didn't take my...|    1|
|               sluts|    2|
|Luckily, feminist...|    1|
|           (3, 1, 1)|    1|
+--------------------+-----+
only showing top 20 rows
```

Fim das tentativas

```
dados.limit(10).show()
```

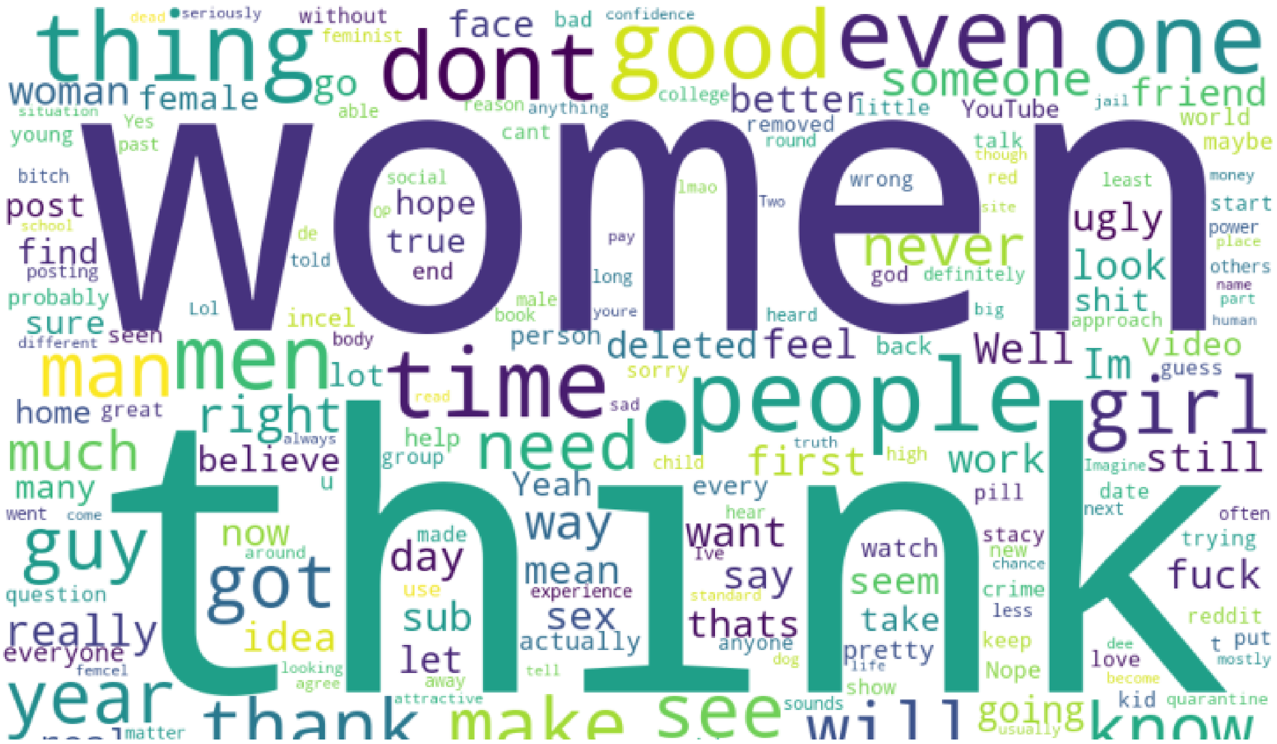
```
+-----+--------------------+---------------+
|index|                body|        level_1|
+-----+--------------------+---------------+
|    0|Do you have the s...|Nonmisogynistic|
|    1|Honestly my favor...|Nonmisogynistic|
|    2|Source? Doesnt so...|Nonmisogynistic|
|    3|Damn, I saw a mov...|   Misogynistic|
|    4|Some places have ...|Nonmisogynistic|
|    5|So if I drink eno...|Nonmisogynistic|
|    6|You'll Benjamin B...|Nonmisogynistic|
|    7|Isn't this the pl...|Nonmisogynistic|
|    8|No idea. I watche...|Nonmisogynistic|
|    9|In the movie ther...|Nonmisogynistic|
+-----+--------------------+---------------+
```

Criação da nuvem de palavras

```
from wordcloud import WordCloud
import matplotlib.pyplot as plt


amostra = dados.select('body').sample(fraction = 0.1, seed = 101)
tudo = [texto['body'] for texto in amostra.collect()]


wordcloud = WordCloud(background_color = 'white',
                      width = 1000,
                      height = 600,
                      collocations = False,
                      prefer_horizontal = 1).generate(str(tudo))


plt.figure(figsize=(20,8))
```

Salvo com sucesso ×

```
plt.show()
```



Limpeza dos caracteres especiais

```
import string
string.punctuation
```

```
'!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'
```

```
import pyspark.sql.functions as f
```

```
dados = dados.withColumn("texto_regex", f.regexp_replace("body", "[\$#,\"!%&'()*+-./:;<=>?
```

```
dados.limit(5).show(truncate=False)
```

```
+-----+-------------------------------------------------------------------------------
|index|body
+-----+-------------------------------------------------------------------------------
|0    |Do you have the skin of a 80 year old grandma? Worry no more, just drink water
|1    |Honestly my favorite thing about this is that they feel the need to cite beaut
|2    |Source? Doesnt sound right to me idk
|3    |Damn, I saw a movie in which the old woman bathed in the blood if virgins  to
|4    |Some places have poor water quality. Virgin blood may have been less expensive
+-----+-------------------------------------------------------------------------------
```

Salvo com sucesso ×

```
o", f.trim(dados.texto_regex) )
```

```
dados.limit(2).show()
```

```
+-----+--------------------+--------------+--------------------+-------------------
|index|                body|       level_1|         texto_regex|         texto_limpo
```

```
+-----+-------------------+--------------+-------------------+-------------------
|    0|Do you have the s...|Nonmisogynistic|Do you have the s...|Do you have the s..
|    1|Honestly my favor...|Nonmisogynistic|Honestly my favor...|Honestly my favor..
+-----+-------------------+--------------+-------------------+-------------------
```

Tokenização do texto

```
from pyspark.ml.feature import Tokenizer

tokenizer = Tokenizer(inputCol = "texto_limpo", outputCol = "tokens")
tokenizado = tokenizer.transform(dados)
```

```
tokenizado.select("texto_limpo", "tokens").show()
```

```
+--------------------+--------------------+
|         texto_limpo|              tokens|
+--------------------+--------------------+
|Do you have the s...|[do, you, have, t...|
|Honestly my favor...|[honestly, my, fa...|
|Source Doesnt sou...|[source, doesnt, ...|
|Damn I saw a movi...|[damn, i, saw, a,...|
|Some places have ...|[some, places, ha...|
|So if I drink eno...|[so, if, i, drink...|
|Youll Benjamin Bu...|[youll, benjamin,...|
|Isnt this the plo...|[isnt, this, the,...|
|No idea I watched...|[no, idea, i, wat...|
|In the movie ther...|[in, the, movie, ...|
|One of my friends...|[one, of, my, fri...|
|What kinda Tom Br...|[what, kinda, tom...|
|   Recommended by FF|[recommended, by,...|
|Professionals say...|[professionals, s...|
|      Virgina Spread|   [virgina, spread]|
|I cant believe it...|[i, cant, believe...|
|I hate you becaus...|[i, hate, you, be...|
|Even if you chug ...|[even, if, you, c...|
|Ill need way more...|[ill, need, way, ...|
|I think I have a ...|[i, think, i, hav...|
+--------------------+--------------------+
only showing top 20 rows
```

Contagem dos tokens

```
[illegible]rType
```

Salvo com sucesso ✕

```
countTokens = f.udf(lambda tokens: len(tokens), IntegerType())
tokenizado.select("texto_limpo", "tokens").withColumn("Freq_tokens", countTokens(f.col("to
```

```
+--------------------+--------------------+-----------+
|         texto_limpo|              tokens|Freq_tokens|
+--------------------+--------------------+-----------+
|Do you have the s...|[do, you, have, t...|         17|
|Honestly my favor...|[honestly, my, fa...|         29|
|Source Doesnt sou...|[source, doesnt, ...|          7|
|Damn I saw a movi...|[damn, i, saw, a,...|         32|
|Some places have ...|[some, places, ha...|         16|
|So if I drink eno...|[so, if, i, drink...|         11|
|Youll Benjamin Bu...|[youll, benjamin,...|          5|
|Isnt this the plo...|[isnt, this, the,...|          6|
|No idea I watched...|[no, idea, i, wat...|         29|
|In the movie ther...|[in, the, movie, ...|         35|
|One of my friends...|[one, of, my, fri...|         38|
|What kinda Tom Br...|[what, kinda, tom...|          7|
|   Recommended by FF|[recommended, by,...|          3|
|Professionals say...|[professionals, s...|         21|
|      Virgina Spread|   [virgina, spread]|          2|
|I cant believe it...|[i, cant, believe...|          8|
|I hate you becaus...|[i, hate, you, be...|         19|
|Even if you chug ...|[even, if, you, c...|         19|
|Ill need way more...|[ill, need, way, ...|         12|
|I think I have a ...|[i, think, i, hav...|         16|
+--------------------+--------------------+-----------+
only showing top 20 rows
```

Retirada das stop words

```
#teste (nltk)
import nltk
nltk.download("stopwords")

from nltk.corpus import stopwords
stop_nltk = stopwords.words("english")
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

```
from pyspark.ml.feature import StopWordsRemover

stop = StopWordsRemover.loadDefaultStopWords("english")
```

```
from pyspark.ml.feature import Tokenizer
```

```
tokenizer = Tokenizer(inputCol = "texto_limpo", outputCol = "tokens" )
tokenized = tokenizer.transform(dados)
```

Salvo com sucesso ×

```
remover = StopWordsRemover(inputCol = "tokens", outputCol = "texto_final", stopWords= stop
feature_data = remover.transform(tokenizado)
```

```
feature_data.select("tokens", "texto_final").limit(11).show(truncate = False)
```

```
+------------------------------------------------------------------------------
```

```
|tokens
+------------------------------------------------------------------------------------
|[do, you, have, the, skin, of, a, 80, year, old, grandma, worry, no, more, just, dri
|[honestly, my, favorite, thing, about, this, is, that, they, feel, the, need, to, ci
|[source, doesnt, sound, right, to, me, idk]
|[damn, i, saw, a, movie, in, which, the, old, woman, bathed, in, the, blood, if, vir
|[some, places, have, poor, water, quality, virgin, blood, may, have, been, less, exp
|[so, if, i, drink, enough, water, i, turn, into, a, baby]
|[youll, benjamin, button, yourself, yes]
|[isnt, this, the, plot, of, cocoon]
|[no, idea, i, watched, the, movie, 10, years, or, so, ago, with, my, dad, and, all,
|[in, the, movie, there, is, this, group, of, elderly, that, discover, that, they, fe
|[one, of, my, friends, looks, like, hes, frikkin, dead, been, trying, for, years, to
+------------------------------------------------------------------------------------
```

```
countTokens = f.udf(lambda tokens: len(tokens), IntegerType())
tokenizado.select("texto_limpo", "tokens").withColumn("Freq_tokens", countTokens(f.col("to
```

```
+--------------------+--------------------+-----------+
|         texto_limpo|              tokens|Freq_tokens|
+--------------------+--------------------+-----------+
|Do you have the s...|[do, you, have, t...|         17|
|Honestly my favor...|[honestly, my, fa...|         29|
|Source Doesnt sou...|[source, doesnt, ...|          7|
|Damn I saw a movi...|[damn, i, saw, a,...|         32|
|Some places have ...|[some, places, ha...|         16|
|So if I drink eno...|[so, if, i, drink...|         11|
|Youll Benjamin Bu...|[youll, benjamin,...|          5|
|Isnt this the plo...|[isnt, this, the,...|          6|
|No idea I watched...|[no, idea, i, wat...|         29|
|In the movie ther...|[in, the, movie, ...|         35|
|One of my friends...|[one, of, my, fri...|         38|
|What kinda Tom Br...|[what, kinda, tom...|          7|
|   Recommended by FF|[recommended, by,...|          3|
|Professionals say...|[professionals, s...|         21|
|      Virgina Spread|   [virgina, spread]|          2|
|I cant believe it...|[i, cant, believe...|          8|
|I hate you becaus...|[i, hate, you, be...|         19|
|Even if you chug ...|[even, if, you, c...|         19|
|Ill need way more...|[ill, need, way, ...|         12|
|I think I have a ...|[i, think, i, hav...|         16|
+--------------------+--------------------+-----------+
only showing top 20 rows
```

```
feature_data.select("tokens", "texto_final")\
          .withColumn("Freq_tokens", countTokens(f.col("tokens")))\
                                   impos",  countTokens(f.col("texto_final"))).show()
```

Salvo com sucesso ×

```
+--------------------+--------------------+-----------+------------------+
|              tokens|         texto_final|Freq_tokens|Freq_tokens_limpos|
+--------------------+--------------------+-----------+------------------+
|[do, you, have, t...|[skin, 80, year, ...|         17|                 8|
|[honestly, my, fa...|[honestly, favori...|         29|                15|
|[source, doesnt, ...|[source, doesnt, ...|          7|                 5|
|[damn, i, saw, a,...|[damn, saw, movie...|         32|                14|
|[some, places, ha...|[places, poor, wa...|         16|                 11|
|[so, if, i, drink...|[drink, enough, w...|         11|                  5|
|[youll, benjamin,...|[youll, benjamin,...|          5|                  4|
|[isnt, this, the,...|[isnt, plot, cocoon]|          6|                  3|
|[no, idea, i, wat...|[idea, watched, m...|         29|                 12|
|[in, the, movie, ...|[movie, group, el...|         35|                 15|
|[one, of, my, fri...|[one, friends, lo...|         38|                 24|
|[what, kinda, tom...|[kinda, tom, brad...|          7|                  4|
|[recommended, by,...|   [recommended, ff]|          3|                  2|
|[professionals, s...|[professionals, s...|         21|                 12|
|   [virgina, spread]|   [virgina, spread]|          2|                  2|
|[i, cant, believe...|[cant, believe, v...|          8|                  5|
|[i, hate, you, be...|[hate, middle, ni...|         19|                  8|
|[even, if, you, c...|[even, chug, hand...|         19|                  9|
|[ill, need, way, ...|[ill, need, way, ...|         12|                  7|
|[i, think, i, hav...|[think, couple, h...|         16|                  7|
+--------------------+--------------------+-----------+-------------------+
only showing top 20 rows
```

```
from pyspark.ml.feature import CountVectorizer
cv = CountVectorizer(inputCol="texto_final", outputCol="CountVec")
model = cv.fit(feature_data)
countVectorizer_features = model.transform(feature_data)

countVectorizer_features.select('texto_final','CountVec').limit(5).show()#truncate=False
```

```
+--------------------+--------------------+
|         texto_final|            CountVec|
+--------------------+--------------------+
|[skin, 80, year, ...|(9929,[173,190,36...|
|[honestly, favori...|(9929,[27,31,37,2...|
|[source, doesnt, ...|(9929,[32,48,510,...|
|[damn, saw, movie...|(9929,[1,9,25,161...|
|[places, poor, wa...|(9929,[134,167,36...|
+--------------------+--------------------+
```

```
from pyspark.ml.feature import HashingTF

hashingTF = HashingTF(inputCol="texto_final", outputCol="hashingTF")
hashingTF.setNumFeatures(1000)

HTFfeaturizedData = hashingTF.transform(countVectorizer_features)
```

Salvo com sucesso ×

```
l", "hashingTF").limit(5).show()
```

```
+--------------------+--------------------+
|         texto_final|           hashingTF|
+--------------------+--------------------+
|[skin, 80, year, ...|(1000,[83,292,343...|
|[honestly, favori...|(1000,[74,86,115,...|
|[source, doesnt, ...|(1000,[166,721,79...|
|[damn, saw, movie...|(1000,[83,129,162...|
|[places, poor, wa...|(1000,[103,160,19...|
+--------------------+--------------------+
```

```
from pyspark.ml.feature import IDF
idf = IDF(inputCol="hashingTF", outputCol="features")
idfModel = idf.fit(HTFfeaturizedData)
TFIDFfeaturizedData = idfModel.transform(HTFfeaturizedData)
```

```
TFIDFfeaturizedData.select('texto_final', 'features').limit(5).show(truncate = False)
```

```
+----------------------------------------------------------------------------------
|texto_final
+----------------------------------------------------------------------------------
|[skin, 80, year, old, grandma, worry, drink, water]
|[honestly, favorite, thing, feel, need, cite, beauty, professionals, order, prove,
|[source, doesnt, sound, right, idk]
|[damn, saw, movie, old, woman, bathed, blood, virgins, , , one, tell, needed, water]
|[places, poor, water, quality, virgin, blood, may, less, expensive, imported, water]
+----------------------------------------------------------------------------------
```

```
TFIDFfeaturizedData.groupBy('level_1').count().show()
```

```
+--------------------+-----+
|             level_1|count|
+--------------------+-----+
| which they would...|    1|
| feel BIGGER than...|    1|
|           (5, 2, 1)|    1|
|             (27, 3)|    1|
|                   8|    1|
|My primary slave ...|    1|
| BLS and some of ...|    1|
|Sexual_or_physica...|    3|
| I KNEW that I wa...|    1|
| you'll be abused...|    1|
|So of course she ...|    1|
|           Hypergamy|    1|
|               Stacy|    1|
| instilling certa...|    1|
| thin with a bit ...|    5|
|I have even had g...|    1|
| I didn't take my...|    1|
|               sluts|    2|
+--------------------+-----+
only showing top 20 rows
```

Salvo com sucesso ×

```
from pyspark.ml.feature import StringIndexer
```

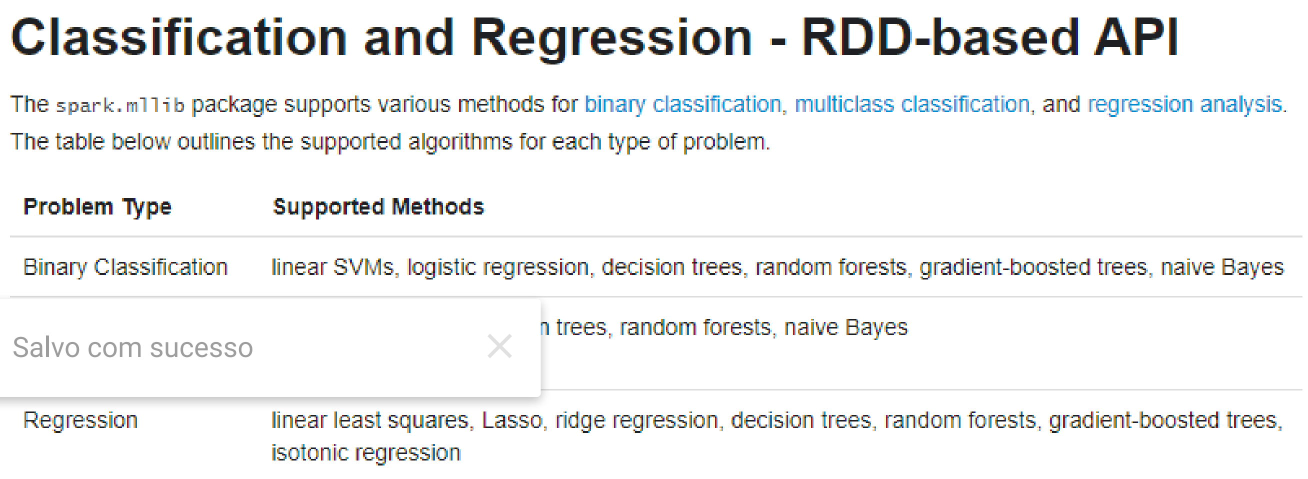
```
stringindexer = StringIndexer(inputCol="level_1", outputCol="label")
dados = stringindexer.fit(dados).transform(dados)
```

```
dados.groupBy(['level_1','label']).count().show()
```

```
+--------------------+-----+-----+
|             level_1|label|count|
+--------------------+-----+-----+
|             beckies| 69.0|    1|
|               becky| 70.0|    1|
|Luckily, feminist...| 62.0|    1|
|So of course she ...| 66.0|    1|
|                (8,)| 53.0|    1|
|     Nonmisogynistic|  0.0| 4246|
|Nature of the abu...|  7.0|    4|
|             (47, 6)| 15.0|    2|
|          (55, 4, 1)| 48.0|    1|
|*,ÄúYou just remi...| 55.0|    1|
|          Hypergamy| 18.0|    1|
|             (10, 4)| 42.0|    1|
|Women do not like...| 68.0|    1|
| but normies have...| 25.0|    1|
|I vet them and th...| 60.0|    1|
| to escalate and ...| 37.0|    1|
|women are likely ...| 77.0|    1|
|                test|  2.0|  112|
|                   8| 57.0|    1|
|   The Rational Male|  8.0|    3|
+--------------------+-----+-----+
only showing top 20 rows
```

Definição dos dados de treino(train) e teste(test)

```
train, test = dados.randomSplit([0.7, 0.3], seed = 101)
```



Classification and Regression - RDD-based API

The `spark.mllib` package supports various methods for binary classification, multiclass classification, and regression analysis. The table below outlines the supported algorithms for each type of problem.

Problem Type	Supported Methods
Binary Classification	linear SVMs, logistic regression, decision trees, random forests, gradient-boosted trees, naive Bayes
	n trees, random forests, naive Bayes
Regression	linear least squares, Lasso, ridge regression, decision trees, random forests, gradient-boosted trees, isotonic regression

Salvo com sucesso ×

https://spark.apache.org/docs/2.2.0/mllib-classification-regression.html

Árvore de Decisão

```
from pyspark.ml import Pipeline
from pyspark.ml.classification import DecisionTreeClassifier

tokenizer = Tokenizer(inputCol="texto_limpo", outputCol="tokens")
stopwords = StopWordsRemover(inputCol="tokens", outputCol="texto_final")
hashingTF = HashingTF(inputCol=stopwords.getOutputCol(), outputCol="HTF", numFeatures=1000
tfidf = IDF(inputCol="HTF", outputCol="features")
dt = DecisionTreeClassifier(featuresCol='features', labelCol='label', maxDepth=10)

pipeline_arvore = Pipeline(stages = [tokenizer,stopwords, hashingTF, tfidf, dt])
```

```
dados_transformados = pipeline_arvore.fit(dados).transform(dados)
dados_transformados.limit(5).show()
```

```
+-----+--------------------+--------------+--------------------+-------------------
|index|                body|       level_1|          texto_regex|         texto_limpo
+-----+--------------------+--------------+--------------------+-------------------
|    0|Do you have the s...|Nonmisogynistic|Do you have the s...|Do you have the s..
|    1|Honestly my favor...|Nonmisogynistic|Honestly my favor...|Honestly my favor..
|    2|Source? Doesnt so...|Nonmisogynistic|Source Doesnt sou...|Source Doesnt sou..
|    3|Damn, I saw a mov...|   Misogynistic|Damn I saw a movi...|Damn I saw a movi..
|    4|Some places have ...|Nonmisogynistic|Some places have ...|Some places have ..
+-----+--------------------+--------------+--------------------+-------------------
```

```
dt_model_treino = pipeline_arvore.fit(train)
predictions_treino_arvore = dt_model_treino.transform(train)
```

```
dt_model_teste = pipeline_arvore.fit(test)
predictions_teste_arvore = dt_model_teste.transform(test)
```

```
predictions_teste_arvore.show()
```

```
+-----+--------------------+--------------+--------------------+-------------------
|index|                body|       level_1|          texto_regex|         texto_limpo
+-----+--------------------+--------------+--------------------+-------------------
|    4|Some places have ...|Nonmisogynistic|Some places have ...|Some places have ..
|    5|So if I drink eno...|Nonmisogynistic|So if I drink eno...|So if I drink eno..
|   14|     Virgina Spread |Nonmisogynistic|     Virgina Spread |      Virgina Spread
|   15|*I can't believe ...|Nonmisogynistic|I cant believe it...|I cant believe it..
                                 misogynistic|Ill need way more...|Ill need way more..
                                 misogynistic|Well now the secr...|Well now the secr..
                                 misogynistic|       rihadastroke |        rihadastroke
|   23|When you think it...|Nonmisogynistic|When you think it...|When you think it..
|   33|at first i was re...|Nonmisogynistic|at first i was re...|at first i was re..
|   35|Given that he wan...|Nonmisogynistic|Given that he wan...|Given that he wan..
|   36|My two favorite t...|Nonmisogynistic|My two favorite t...|My two favorite t..
|   42|"Vogoncel".That m...|Nonmisogynistic|VogoncelThat made...|VogoncelThat made..
|   47|But remember, the...|Nonmisogynistic|But remember thei...|But remember thei..
|   50|Like listening to...|Nonmisogynistic|Like listening to...|Like listening to..
```

Salvo com sucesso ×

```
|   51|God is the reason...|Nonmisogynistic|God is the reason...|God is the reason..
|   52|Personally I'd sa...|Nonmisogynistic|Personally Id say...|Personally Id say..
|   56|Aren't you super ...|Nonmisogynistic|Arent you super h...|Arent you super h..
|   58|I'm personally an...|Nonmisogynistic|Im personally an ...|Im personally an ..
|   61|Yes.  Rights are ...|Nonmisogynistic|Yes  Rights are l...|Yes  Rights are l..
|   63|"Oh, if only I ha...|Nonmisogynistic|Oh if only I had ...|Oh if only I had ..
+-----+--------------------+---------------+--------------------+-------------------
only showing top 20 rows
```

```
predictions_teste_arvore.select(['label','prediction']).show()
```

```
+-----+----------+
|label|prediction|
+-----+----------+
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
+-----+----------+
only showing top 20 rows
```

```
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
evaluator = MulticlassClassificationEvaluator(labelCol='label', predictionCol='prediction'
```

```
print("Acuracia = %f" % evaluator.evaluate(predictions_teste_arvore, {evaluator.metricName
print("Precisão = %f" % evaluator.evaluate(predictions_teste_arvore, {evaluator.metricName
print("Recall = %f" % evaluator.evaluate(predictions_teste_arvore, {evaluator.metricName:'
print("F1 = %f" % evaluator.evaluate(predictions_teste_arvore, {evaluator.metricName:'fMea
```

Salvo com sucesso ×

```
Recall = 0.998423
F1 = 0.962006
```

Random Forest

```
from pyspark.ml.regression import RandomForestRegressor

tokenizer = Tokenizer(inputCol="texto_limpo", outputCol="tokens")
stopwords = StopWordsRemover(inputCol="tokens", outputCol="texto_final")
hashingTF = HashingTF(inputCol=stopwords.getOutputCol(), outputCol="HTF", numFeatures=1000
tfidf = IDF(inputCol="HTF", outputCol="features")
rfr = RandomForestRegressor(featuresCol='features', labelCol='label', maxDepth=10, numTree

pipeline_randomforest = Pipeline(stages=[tokenizer, stopwords, hashingTF, tfidf, rfr])
```

```
rfr_model_treino = pipeline_randomforest.fit(train)
predictions_treino_ranomforest = rfr_model_treino.transform(train)
```

```
rfr_model_teste = pipeline_randomforest.fit(test)
predictions_teste_randomforest = rfr_model_teste.transform(test)
```

```
predictions_teste_randomforest.show()
```

```
+-----+--------------------+--------------+--------------------+-------------------
|index|                body|       level_1|         texto_regex|         texto_limpo
+-----+--------------------+--------------+--------------------+-------------------
|    4|Some places have ...|Nonmisogynistic|Some places have ...|Some places have ..
|    5|So if I drink eno...|Nonmisogynistic|So if I drink eno...|So if I drink eno..
|   14|     Virgina Spread |Nonmisogynistic|     Virgina Spread |      Virgina Spread
|   15|*I can't believe ...|Nonmisogynistic|I cant believe it...|I cant believe it..
|   18|I'll need *way* m...|Nonmisogynistic|Ill need way more...|Ill need way more..
|   20|Well now the secr...|Nonmisogynistic|Well now the secr...|Well now the secr..
|   22|  r/ihadastroke ...?|Nonmisogynistic|       rihadastroke |       rihadastroke
|   23|When you think it...|Nonmisogynistic|When you think it...|When you think it..
|   33|at first i was re...|Nonmisogynistic|at first i was re...|at first i was re..
|   35|Given that he wan...|Nonmisogynistic|Given that he wan...|Given that he wan..
|   36|My two favorite t...|Nonmisogynistic|My two favorite t...|My two favorite t..
|   42|"Vogoncel".That m...|Nonmisogynistic|VogoncelThat made...|VogoncelThat made..
|   47|But remember, the...|Nonmisogynistic|But remember thei...|But remember thei..
|   50|Like listening to...|Nonmisogynistic|Like listening to...|Like listening to..
|   51|God is the reason...|Nonmisogynistic|God is the reason...|God is the reason..
|   52|Personally I'd sa...|Nonmisogynistic|Personally Id say...|Personally Id say..
|   56|Aren't you super ...|Nonmisogynistic|Arent you super h...|Arent you super h..
|   58|I'm personally an...|Nonmisogynistic|Im personally an ...|Im personally an ..
|   61|Yes.  Rights are ...|Nonmisogynistic|Yes  Rights are l...|Yes  Rights are l..
|   63|"Oh, if only I ha...|Nonmisogynistic|Oh if only I had ...|Oh if only I had ..
+-----+--------------------+--------------+--------------------+-------------------
only showing top 20 rows
```

Salvo com sucesso ×

```
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
evaluator = MulticlassClassificationEvaluator(labelCol='label', predictionCol='prediction'
```

```
print("Acuracia = %f" % evaluator.evaluate(predictions_teste_randomforest, {evaluator.metr
print("Precisão = %f" % evaluator.evaluate(predictions_teste_randomforest, {evaluator.metr
print("Recall = %f" % evaluator.evaluate(predictions_teste_randomforest, {evaluator.metric
print("F1 = %f" % evaluator.evaluate(predictions_teste_randomforest, {evaluator.metricName
```

```
Acuracia = 0.000700
Precisão = 1.000000
Recall = 0.000789
F1 = 0.001576
```

Regressão Logistica

```
from pyspark.ml.classification import LogisticRegression

tokenizer = Tokenizer(inputCol="texto_limpo", outputCol="tokens")
stopwords = StopWordsRemover(inputCol="tokens", outputCol="texto_final")
hashingTF = HashingTF(inputCol=stopwords.getOutputCol(), outputCol="HTF", numFeatures=1000
tfidf = IDF(inputCol="HTF", outputCol="features")
lr = LogisticRegression(featuresCol='features', labelCol='label', maxIter=10, regParam=0.0

pipeline_logisticregression = Pipeline(stages=[tokenizer, stopwords, hashingTF, tfidf, lr]
```

```
lr_model_treino = pipeline_logisticregression.fit(train)
predictions_treino_logisticregression = lr_model_treino.transform(train)
```

```
lr_model_teste = pipeline_logisticregression.fit(test)
predictions_teste_logisticregression = lr_model_teste.transform(test)
```

```
predictions_teste_logisticregression.show()
```

```
+-----+--------------------+--------------+--------------------+--------------------
|index|                body|       level_1|         texto_regex|         texto_limpo
+-----+--------------------+--------------+--------------------+--------------------
|    4|Some places have ...|Nonmisogynistic|Some places have ...|Some places have ..
|    5|So if I drink eno...|Nonmisogynistic|So if I drink eno...|So if I drink eno..
|   14|     Virgina Spread |Nonmisogynistic|     Virgina Spread |      Virgina Spread
|   15|*I can't believe ...|Nonmisogynistic|I cant believe it...|I cant believe it..
|   18|I'll need *way* m...|Nonmisogynistic|Ill need way more...|Ill need way more..
|   20|Well now the secr...|Nonmisogynistic|Well now the secr...|Well now the secr..
|   22|  r/ihadastroke ...?|Nonmisogynistic|       rihadastroke |        rihadastroke
|   23|When you think it...|Nonmisogynistic|When you think it...|When you think it..
|   33|at first i was re...|Nonmisogynistic|at first i was re...|at first i was re..
|   35|Given that he wan...|Nonmisogynistic|Given that he wan...|Given that he wan..
|   36|My two favorite t...|Nonmisogynistic|My two favorite t...|My two favorite t..
|   42|"Vogoncel".That m...|Nonmisogynistic|VogoncelThat made...|VogoncelThat made..
                             misogynistic|But remember thei...|But remember thei..
                             misogynistic|Like listening to...|Like listening to..
                             misogynistic|God is the reason...|God is the reason..
|   52|Personally I'd sa...|Nonmisogynistic|Personally Id say...|Personally Id say..
|   56|Aren't you super ...|Nonmisogynistic|Arent you super h...|Arent you super h..
|   58|I'm personally an...|Nonmisogynistic|Im personally an ...|Im personally an ..
|   61|Yes.  Rights are ...|Nonmisogynistic|Yes  Rights are l...|Yes  Rights are l..
|   63|"Oh, if only I ha...|Nonmisogynistic|Oh if only I had ...|Oh if only I had ..
+-----+--------------------+--------------+--------------------+--------------------
```

Salvo com sucesso ×

```
    only showing top 20 rows
```

```
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
evaluator = MulticlassClassificationEvaluator(labelCol='label', predictionCol='prediction'


print("Acuracia = %f" % evaluator.evaluate(predictions_teste_logisticregression, {evaluato
print("Precisão = %f" % evaluator.evaluate(predictions_teste_logisticregression, {evaluato
print("Recall = %f" % evaluator.evaluate(predictions_teste_logisticregression, {evaluator.
print("F1 = %f" % evaluator.evaluate(predictions_teste_logisticregression, {evaluator.metr
```

```
    Acuracia = 0.992997
    Precisão = 0.996072
    Recall = 1.000000
    F1 = 0.998032
```

```
y_true = predictions_teste_logisticregression.select(['label']).collect()
y_pred = predictions_teste_logisticregression.select(['prediction']).collect()

from sklearn.metrics import classification_report, confusion_matrix
print(confusion_matrix(y_true, y_pred))
```

```
y_true = predictions_teste_logisticregression.select(['label']).collect()
y_pred = predictions_teste_logisticregression.select(['prediction']).collect()

from sklearn.metrics import classification_report, confusion_matrix
print(classification_report(y_true, y_pred))
```

```
                  precision    recall  f1-score   support

             0.0       1.00      1.00      1.00      1268
             1.0       1.00      0.99      0.99        95
             2.0       0.88      0.95      0.91        38
             3.0       1.00      1.00      1.00         1
             4.0       1.00      1.00      1.00         1
             6.0       1.00      1.00      1.00         2
             9.0       1.00      1.00      1.00         1
            10.0       1.00      1.00      1.00         2
            11.0       1.00      1.00      1.00         1
            12.0       1.00      1.00      1.00         1
            13.0       1.00      1.00      1.00         2
            16.0       0.00      0.00      0.00         1
            18.0       1.00      1.00      1.00         1
            23.0       1.00      1.00      1.00         1
            24.0       1.00      1.00      1.00         1
                                    .00      1.00         1
                                    .00      1.00         1
            40.0       1.00      1.00      1.00         1
            42.0       0.00      0.00      0.00         1
            43.0       0.00      0.00      0.00         1
            52.0       1.00      1.00      1.00         1
            56.0       1.00      1.00      1.00         1
            57.0       1.00      1.00      1.00         1
            60.0       0.00      0.00      0.00         1
```

Salvo com sucesso ×

```
        62.0       0.00      0.00      0.00         1
        70.0       0.00      0.00      0.00         1
        76.0       0.00      0.00      0.00         1

    accuracy                           0.99      1428
   macro avg       0.74      0.74      0.74      1428
weighted avg       0.99      0.99      0.99      1428

/usr/local/lib/python3.7/dist-packages/sklearn/metrics/_classification.py:1318: Unde
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.7/dist-packages/sklearn/metrics/_classification.py:1318: Unde
  _warn_prf(average, modifier, msg_start, len(result))
/usr/local/lib/python3.7/dist-packages/sklearn/metrics/_classification.py:1318: Unde
  _warn_prf(average, modifier, msg_start, len(result))
```

Naive Bayes

```
from pyspark.ml.classification import NaiveBayes

tokenizer = Tokenizer(inputCol="texto_limpo", outputCol="tokens")
stopwords = StopWordsRemover(inputCol="tokens", outputCol="texto_final")
hashingTF = HashingTF(inputCol=stopwords.getOutputCol(), outputCol="HTF", numFeatures=1000
tfidf = IDF(inputCol="HTF", outputCol="features")
nb = NaiveBayes(featuresCol='features', labelCol='label', smoothing=1.0, modelType="multin

pipeline_naive = Pipeline(stages=[tokenizer, stopwords, hashingTF, tfidf,nb])
```

```
naive_model_treino = pipeline_naive.fit(train)
predictions_treino_naive = naive_model_treino.transform(train)
```

```
naive_model_teste = pipeline_naive.fit(test)
predictions_teste_naive = naive_model_teste.transform(test)
```

```
predictions_teste_naive.show()
```

```
+-----+--------------------+---------------+--------------------+--------------------
|index|                body|        level_1|         texto_regex|         texto_limpo
+-----+--------------------+---------------+--------------------+--------------------
|    4|Some places have ...|Nonmisogynistic|Some places have ...|Some places have ..
|    5|So if I drink eno...|Nonmisogynistic|So if I drink eno...|So if I drink eno..
|   14|     Virgina Spread |Nonmisogynistic|     Virgina Spread |      Virgina Spreac
|   15|*I can't believe ...|Nonmisogynistic|I cant believe it...|I cant believe it..
|   18|I'll need *way* m...|Nonmisogynistic|Ill need way more...|Ill need way more..
|   20|Well now the secr...|Nonmisogynistic|Well now the secr...|Well now the secr..
                              misogynistic|       rihadastroke |        rihadastroke
                              misogynistic|When you think it...|When you think it..
|   33|at first i was re...|Nonmisogynistic|at first i was re...|at first i was re..
|   35|Given that he wan...|Nonmisogynistic|Given that he wan...|Given that he wan..
|   36|My two favorite t...|Nonmisogynistic|My two favorite t...|My two favorite t..
|   42|"Vogoncel".That m...|Nonmisogynistic|VogoncelThat made...|VogoncelThat made..
|   47|But remember, the...|Nonmisogynistic|But remember thei...|But remember thei..
|   50|Like listening to...|Nonmisogynistic|Like listening to...|Like listening to..
|   51|God is the reason...|Nonmisogynistic|God is the reason...|God is the reason..
```

Salvo com sucesso ×

```
|   52|Personally I'd sa...|Nonmisogynistic|Personally Id say...|Personally Id say..
|   56|Aren't you super ...|Nonmisogynistic|Arent you super h...|Arent you super h..
|   58|I'm personally an...|Nonmisogynistic|Im personally an ...|Im personally an ..
|   61|Yes.  Rights are ...|Nonmisogynistic|Yes  Rights are l...|Yes  Rights are l..
|   63|"Oh, if only I ha...|Nonmisogynistic|Oh if only I had ...|Oh if only I had ..
+-----+--------------------+---------------+--------------------+-------------------
only showing top 20 rows
```

```
print("Acuracia = %f" % evaluator.evaluate(predictions_teste_naive, {evaluator.metricName:
print("Precisão = %f" % evaluator.evaluate(predictions_teste_naive, {evaluator.metricName:
print("Recall = %f" % evaluator.evaluate(predictions_teste_naive, {evaluator.metricName:'r
print("F1 = %f" % evaluator.evaluate(predictions_teste_naive, {evaluator.metricName:'fMeas
```

```
Acuracia = 0.815126
Precisão = 0.990431
Recall = 0.816246
F1 = 0.894942
```

Continuação utilizando Scikit-learn

```
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer, TfidfTransformer
from sklearn.svm import LinearSVC
from sklearn.pipeline import Pipeline
from sklearn.base import BaseEstimator, TransformerMixin
from sklearn.model_selection import train_test_split
from sklearn import metrics
```

```
dataset = pd.read_csv("/content/drive/MyDrive/TCC_BANCOS/online-misogyny-eacl2021-main/dat
```

```
dataset=dataset[['body','level_1']]
dataset.head()
```

	body	level_1
0	Do you have the skin of a 80 year old grandma?...	Nonmisogynistic
1	This is taking a grain of truth and extrapolat...	Nonmisogynistic
2	Honestly my favorite thing about this is that ...	Nonmisogynistic
3	Source? Doesnt sound right to me idk	Nonmisogynistic
4	Damn, I saw a movie in which the old woman bat...	Misogynistic

Salvo com sucesso ×

```
dataset.isnull().sum()
```

```
body       12
level_1     0
dtype: int64
```

```
dataset.dropna(inplace=True)
```

```
dataset.isnull().sum()
```

```
body       0
level_1    0
dtype: int64
```

```
dataset.dtypes
```

```
body       object
level_1    object
dtype: object
```

```
dataset.groupby('level_1')['level_1'].count()
```

```
level_1
Misogynistic        699
Nonmisogynistic    5856
Name: level_1, dtype: int64
```

```
dataset['body'] = dataset['body'].astype(str)
```

```
a_trocar = {
    'Nonmisogynistic': 0,
    'Misogynistic': 1
}
dataset.level_1 = dataset.level_1.map(a_trocar)
dataset.head()
```

	body	level_1
0	Do you have the skin of a 80 year old grandma?...	0
1	This is taking a grain of truth and extrapolat...	0
2	Honestly my favorite thing about this is that ...	0
3	Source? Doesnt sound right to me idk	0
4	Damn, I saw a movie in which the old woman bat...	1

```
class TColumns(BaseEstimator, TransformerMixin):
```

Salvo com sucesso ×

```
    return self

  def transform(self, X):

    dataset = X.copy()
    dataset['body'] = dataset['body'].str.replace('[,.:;!?]+', ' ', regex=True).copy()
    dataset['body'] = dataset['body'].str.replace('[/<>()|\+\-\$%&#@\'\"]+', ' ', regex=Tr
    dataset['body'] = dataset['body'].str.replace('[0-9]+', '', regex=True)

    return dataset.body
```

KNN (K-nearest neighbours)

```
from sklearn.neighbors import KNeighborsClassifier
```

```
tco = TColumns()

cvt = CountVectorizer(strip_accents='ascii', lowercase=True, stop_words=stop)

tfi = TfidfTransformer(use_idf=True)

knn = KNeighborsClassifier(n_neighbors=3)

knn_pipeline = Pipeline(steps=[('Transformer', tco),
                              ('CountVectorizer', cvt),
                              ('TfidfTransformer', tfi),
                              ('Model', knn)])
```

```
entrada = dataset[['body']]
saida = dataset['level_1']
X_train, X_test, y_train, y_test = train_test_split(entrada,
                                                    saida,
                                                    test_size=0.3)
knn_pipeline.fit(X_train, y_train)
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/feature_extraction/text.py:401: UserWa
  % sorted(inconsistent)
Pipeline(steps=[('Transformer', TColumns()),
                ('CountVectorizer',
                 CountVectorizer(stop_words=['i', 'me', 'my', 'myself', 'we',
                                             'our', 'ours', 'ourselves', 'you',
                                             'your', 'yours', 'yourself',
                                             'yourselves', 'he', 'him', 'his',
                                             'himself', 'she', 'her', 'hers',
                                             'herself', 'it', 'its', 'itself',
                                             'they', 'them', 'their', 'theirs',
                                             'themselves', 'what', ...],
                                 strip_accents='ascii')),
                ('TfidfTransformer', TfidfTransformer()),
                ('Model', KNeighborsClassifier(n_neighbors=3))])
```

Salvo com sucesso ×

```
.predict(X_test)
```

```
print("Acurácia: {}".format(metrics.accuracy_score(y_test, predictions_teste_knn)))
print("Precision: {}".format(metrics.precision_score(y_test, predictions_teste_knn)))
print("Recall: {}".format(metrics.recall_score(y_test, predictions_teste_knn)))
print("F1: {}".format(metrics.f1_score(y_test, predictions_teste_knn)))
```

```
Acurácia: 0.9201830198271479
Precision: 0.972972972972973
Recall: 0.1875
F1: 0.314410480349345
```

```
from sklearn.metrics import accuracy_score
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
from sklearn.metrics import f1_score
```

SVM (Support Vector Machine)

```
from sklearn import svm
```

```
tco = TColumns()

cvt = CountVectorizer(strip_accents='ascii', lowercase=True, stop_words=stop)

tfi = TfidfTransformer(use_idf=True)

svm = svm.SVC()

svm_pipeline = Pipeline(steps=[('Transformer', tco),
                               ('CountVectorizer', cvt),
                               ('TfidfTransformer', tfi),
                               ('Model', svm)])
```

```
entrada = dataset[['body']]
saida = dataset['level_1']
X_train, X_test, y_train, y_test = train_test_split(entrada,
                                                     saida,
                                                     test_size=0.3)
svm_pipeline.fit(X_train, y_train)
```

```
/usr/local/lib/python3.7/dist-packages/sklearn/feature_extraction/text.py:401: UserWa
  % sorted(inconsistent)
Pipeline(steps=[('Transformer', TColumns()),
                ('CountVectorizer',
                 CountVectorizer(stop_words=['i', 'me', 'my', 'myself', 'we',
                                             'our', 'ours', 'ourselves', 'you',
                                             'your', 'yours', 'yourself',
                                             'yourselves', 'he', 'him', 'his',
                                             'himself', 'she', 'her', 'hers',
                                             'herself', 'it', 'its', 'itself',
                                             'they', 'them', 'their', 'theirs',
                                             'themselves', 'what', ...],
                                 strip_accents='ascii')),
                ('TfidfTransformer', TfidfTransformer()), ('Model', SVC())])
```

Salvo com sucesso ×

```
predictions_teste_svm = knn_pipeline.predict(X_test)
```

```
print("Acurácia: {}".format(metrics.accuracy_score(y_test, predictions_teste_svm)))
print("Precision: {}".format(metrics.precision_score(y_test, predictions_teste_svm)))
print("Recall: {}".format(metrics.recall_score(y_test, predictions_teste_svm)))
print("F1: {}".format(metrics.f1_score(y_test, predictions_teste_svm)))
```

```
Acurácia: 0.9344178952719878
Precision: 0.9733333333333334
Recall: 0.365
F1: 0.5309090909090909
```

BERT (unweighted)

BERT (weighted)

Tabela de resultado de cada modelo

```
print("====================================")
print("Árvore de Decisão")
print("====================================")
print("Acuracia = %f" % evaluator.evaluate(predictions_teste_arvore, {evaluator.metricName
print("Precisão = %f" % evaluator.evaluate(predictions_teste_arvore, {evaluator.metricName
print("Recall = %f" % evaluator.evaluate(predictions_teste_arvore, {evaluator.metricName:'
print("F1 = %f" % evaluator.evaluate(predictions_teste_arvore, {evaluator.metricName:'fMea

print("====================================")
print("Random Forest")
print("====================================")
print("Acuracia = %f" % evaluator.evaluate(predictions_teste_randomforest, {evaluator.metr
print("Precisão = %f" % evaluator.evaluate(predictions_teste_randomforest, {evaluator.metr
print("Recall = %f" % evaluator.evaluate(predictions_teste_randomforest, {evaluator.metric
print("F1 = %f" % evaluator.evaluate(predictions_teste_randomforest, {evaluator.metricName

print("====================================")
print("Regressão Logistica")
print("====================================")
print("Acuracia = %f" % evaluator.evaluate(predictions_teste_logisticregression, {evaluato
print("Precisão = %f" % evaluator.evaluate(predictions_teste_logisticregression, {evaluato
print("Recall = %f" % evaluator.evaluate(predictions_teste_logisticregression, {evaluator.
print("F1 = %f" % evaluator.evaluate(predictions_teste_logisticregression, {evaluator.metr

print("====================================")
```

Salvo com sucesso ×

```
=======")
print("Acuracia = %f" % evaluator.evaluate(predictions_teste_naive, {evaluator.metricName:
print("Precisão = %f" % evaluator.evaluate(predictions_teste_naive, {evaluator.metricName:
print("Recall = %f" % evaluator.evaluate(predictions_teste_naive, {evaluator.metricName:'r
print("F1 = %f" % evaluator.evaluate(predictions_teste_naive, {evaluator.metricName:'fMeas
print("====================================")
print("KNN")
print("====================================")
print("Acurácia: {}".format(metrics.accuracy_score(y_test, predictions_teste_knn)))
print("Precision: {}".format(metrics.precision_score(y_test, predictions_teste_knn)))
print("Recall: {}".format(metrics.recall_score(y_test, predictions_teste_knn)))
print("F1: {}".format(metrics.f1_score(y_test, predictions_teste_knn)))


print("====================================")
print("SVM")
print("====================================")
print("Acurácia: {}".format(metrics.accuracy_score(y_test, predictions_teste_svm)))
print("Precision: {}".format(metrics.precision_score(y_test, predictions_teste_svm)))
print("Recall: {}".format(metrics.recall_score(y_test, predictions_teste_svm)))
print("F1: {}".format(metrics.f1_score(y_test, predictions_teste_svm)))
print("====================================")
```

```
====================================
Árvore de Decisão
====================================
Acuracia = 0.929972
Precisão = 0.928152
Recall = 0.998423
F1 = 0.962006
====================================
Random Forest
====================================
Acuracia = 0.000700
Precisão = 1.000000
Recall = 0.000789
F1 = 0.001576
====================================
Regressão Logistica
====================================
Acuracia = 0.992997
Precisão = 0.996072
Recall = 1.000000
F1 = 0.998032
====================================
Naive Bayes
====================================
Acuracia = 0.815126
Precisão = 0.990431
Recall = 0.816246
F1 = 0.894942
====================================
KNN
====================================
```

Salvo com sucesso ×

```
F1: 0.008438818565400845
====================================
SVM
====================================
Acurácia: 0.9344178952719878
Precision: 0.9733333333333334
```

```
Recall: 0.365
F1: 0.53090909090909
=================================
```