sentimentanalyzer

August 6, 2024

[1]:

```
from pyspark.sql import SparkSession
from pyspark.sql.types import *
import wget
from pyspark.ml.feature import␣
 ↪Bucketizer,RegexTokenizer,StopWordsRemover,CountVectorizer,IDF
from pyspark.sql.functions import *
from pyspark.ml.classification import LogisticRegression
from pyspark.ml import Pipeline,PipelineModel
from pyspark.ml.evaluation import BinaryClassificationEvaluator
```

[2]:

```
#Spark Session creation configured to interact with Kfka and MongoDB
spark = SparkSession.builder.appName("pyspark-notebook").\
config("spark.jars.packages","org.apache.spark:spark-sql-kafka-0-10_2.12:3.0.
 ↪0,org.apache.spark:spark-avro_2.12:3.0.0,org.mongodb.spark:
 ↪mongo-spark-connector_2.12:3.0.0").\
config("spark.mongodb.input.uri","mongodb://ubuntu_mongo_1:27017/twitter_db.
 ↪tweets").\
config("spark.mongodb.output.uri","mongodb://ubuntu_mongo_1:27017/twitter_db.
 ↪tweets").\
getOrCreate()
```

```
Ivy Default Cache set to: /root/.ivy2/cache
The jars for the packages stored in: /root/.ivy2/jars
:: loading settings :: url = jar:file:/usr/local/lib/python3.7/dist-packages/pys
park/jars/ivy-2.4.0.jar!/org/apache/ivy/core/settings/ivysettings.xml
org.apache.spark#spark-sql-kafka-0-10_2.12 added as a dependency
org.apache.spark#spark-avro_2.12 added as a dependency
org.mongodb.spark#mongo-spark-connector_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-
parent-9ca84ec5-5e4e-4ab2-a963-04d8187feb9c;1.0
        confs: [default]
        found org.apache.spark#spark-sql-kafka-0-10_2.12;3.0.0 in central
        found org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.0.0 in
central
        found org.apache.kafka#kafka-clients;2.4.1 in central
        found com.github.luben#zstd-jni;1.4.4-3 in central
        found org.lz4#lz4-java;1.7.1 in central
        found org.xerial.snappy#snappy-java;1.1.7.5 in central
        found org.slf4j#slf4j-api;1.7.30 in central
        found org.spark-project.spark#unused;1.0.0 in central
        found org.apache.commons#commons-pool2;2.6.2 in central
        found org.apache.spark#spark-avro_2.12;3.0.0 in central
        found org.mongodb.spark#mongo-spark-connector_2.12;3.0.0 in central
        found org.mongodb#mongodb-driver-sync;4.0.5 in central
        found org.mongodb#bson;4.0.5 in central
        found org.mongodb#mongodb-driver-core;4.0.5 in central
:: resolution report :: resolve 505ms :: artifacts dl 8ms
        :: modules in use:
        com.github.luben#zstd-jni;1.4.4-3 from central in [default]
        org.apache.commons#commons-pool2;2.6.2 from central in [default]
        org.apache.kafka#kafka-clients;2.4.1 from central in [default]
        org.apache.spark#spark-avro_2.12;3.0.0 from central in [default]
        org.apache.spark#spark-sql-kafka-0-10_2.12;3.0.0 from central in
[default]
        org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.0.0 from central
in [default]
        org.lz4#lz4-java;1.7.1 from central in [default]
        org.mongodb#bson;4.0.5 from central in [default]
        org.mongodb#mongodb-driver-core;4.0.5 from central in [default]
        org.mongodb#mongodb-driver-sync;4.0.5 from central in [default]
        org.mongodb.spark#mongo-spark-connector_2.12;3.0.0 from central in
[default]
        org.slf4j#slf4j-api;1.7.30 from central in [default]
        org.spark-project.spark#unused;1.0.0 from central in [default]
        org.xerial.snappy#snappy-java;1.1.7.5 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   14  |   0   |   0   |   0   ||   14  |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-
parent-9ca84ec5-5e4e-4ab2-a963-04d8187feb9c
        confs: [default]
        0 artifacts copied, 14 already retrieved (0kB/13ms)
24/07/22 04:56:25 WARN NativeCodeLoader: Unable to load native-hadoop library
for your platform… using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
setLogLevel(newLevel).
```

```
[3]: spark.read.json("reviews_Sports_and_Outdoors_5.json.gz").show(35)
```

```
+----------+-------+-------+--------------------+-----------+--------------+----
----------------+--------------------+--------------+
|      asin|helpful|overall|          reviewText| reviewTime|    reviewerID|
reviewerName|             summary|unixReviewTime|
+----------+-------+-------+--------------------+-----------+--------------+----
----------------+--------------------+--------------+
|1881509818| [0, 0]|    5.0|This came in on t…|01 26, 2014|  AIXZKN4ACSKI|
David Briner|      Woks very good|    1390694400|
|1881509818| [1, 1]|    5.0|I had a factory G…| 02 2, 2012|A1L5P841VIO02V|
Jason A. Kramer|Works as well as …|    1328140800|
|1881509818| [2, 2]|    4.0|If you don't have…|02 28, 2012| AB2W04NI4OEAD|
J. Fernald|It's a punch, tha…|    1330387200|
|1881509818| [0, 0]|    4.0|This works no bet…| 02 5,
2012|A148SVSWKTJKU6|Jusitn A. Watts "…|It's a punch with…|    1328400000|
|1881509818| [0, 0]|    4.0|I purchased this …|04 23, 2013| AAAWJ6LW9WMOO|
Material Man|Ok,tool does what…|    1366675200|
|1881509818| [0, 0]|    5.0|Needed this tool …| 11 2, 2012|A2XX2A4OJCDNLZ|
RatherLiveInKeyWest|Glock punch tool …|    1351814400|
|1881509818| [0, 0]|    5.0|If u don't have i…|06 10, 2014|A283UOBQRUNM4Q|
Thomas Dragon|          Great tool|    1402358400|
|2094869245| [0, 0]|    4.0|This light will n…|08 31, 2013| AWG3H90WVZ0Z1|
Alec Nelson|             Bright!|    1377907200|
|2094869245| [0, 1]|    5.0|Light and laser t…|05 27, 2013|A3V52OTJHKIJZX|A.
Saenz Jr. "Bet…|             Be seen|    1369612800|
|2094869245| [0, 0]|    5.0|Does everything i…| 11 2, 2013|A3SZBE5F3UQ9EC|
ChasRat "ChasRat"|Bicycle rear tail…|    1383350400|
|2094869245| [0, 0]|    4.0|Very bright.  I w…| 05 7, 2014|A2HVMUMOKOGCQ9|
G. Inman|          Great lite|    1399420800|
|2094869245| [0, 0]|    3.0|It's cheaply made…| 01 7, 2014|A21AJ9GNCM89MK|
Greg|It's worth the pr…|    1389052800|
|2094869245| [0, 0]|    5.0|Mine arrived with…|01 14, 2014|A1OX9ME6R66JDX|
Hugo M. M. Rabson|For $11, it's a b…|    1389657600|
|2094869245| [0, 0]|    4.0|It works great it…|12 20,
2013|A2I7K5OIEXUI6R|Lswieckitay "Lswi…|               Bulky|    1387497600|
|2094869245| [0, 0]|    5.0|I love this light…|09 18, 2013|A2RCMHV3MHEBDP|
Micah Chan|            Love it!|    1379462400|
|2094869245| [0, 0]|    5.0|Bit bulky. One bu…|01 16, 2014|A2A26KED39175E|
Pudknocker71|       Bulky but…|    1389830400|
|2094869245| [0, 0]|    5.0|it is bright and …| 12 7, 2013| ANKZUDSZFUMNZ|
ronald|     rear bike light|    1386374400|
|2094869245| [0, 0]|    4.0|A mice bright lig…| 11 4, 2013|A2M93OC5AOMMM3|
Vette71|Needed a little m…|    1383523200|
|2094869245| [0, 0]|    4.0|Had one ride on t…|11 12, 2013| AO3MOAXLLOAGW|
Vic D "Cope"|Good light for th…|    1384214400|
|7245456259| [0, 0]|    2.0|So it worked well…|03 28, 2014|A2NFEGCOY2TO1Q|
Adam|resistance was go…|    1395964800|
|7245456259| [0, 0]|    5.0|My girlfriend is …| 09 5, 2013|A16VC5E75E3KT4|
Dan B.|Girlfriend loves …|    1378339200|
|7245456259| [0, 0]|    5.0|I worked out once…| 07 6, 2014| ACH40GDEWZRJS|
Dark Harden|I'm not opposed t…|    1404604800|
|7245456259| [0, 1]|    5.0|I bought the purp…|07 13, 2013| AHRQOLXJE4CBV|
JACK LOBO "ljb926"|I BOUGHT THE PURP…|    1373673600|
|7245456259| [0, 2]|    5.0|Well I have had t…|06 27,
2013|A3AFVG8GJRVFM1|Melchor Orozco-Ma…|good resistance band|    1372291200|
|7245456259| [0, 0]|    5.0|Works just as adv…|04 17, 2014|A27M3YPDI1YU5B|
R. Davis|       Great device.|    1397692800|
|7245456259| [0, 0]|    5.0|This band works w…| 03 9, 2013|A1YYDO3HZHED58|Ryan
Nathaniel White|Great product, ev…|    1362787200|
|7245456313| [0, 0]|    5.0|These bands were …|01 24,
2014|A1CJ6O4N4ZWGGR|Afrikan "Freedom …|   Excellent product|    1390521600|
|7245456313| [0, 0]|    5.0|If you are creati…|04 12, 2013|A29NO61G1WH74V|
Alan Smithee|         Gym in a bag|    1365724800|
|7245456313| [0, 0]|    4.0|I like it but I h…| 07 6, 2014|A2ZWD1RON75HX3|
Al|          Great set.|    1404604800|
|7245456313| [0, 0]|    5.0|Love the wide var…|12 15, 2012|A1XLWJPVB4WEMP|
Amazon Customer|             Love it|    1355529600|
|7245456313| [0, 0]|    5.0|You can vary your…|10 19, 2013| AE4LMEMGK2TI8|
Angela Lunn|           Versatile|    1382140800|
|7245456313| [0, 1]|    2.0|I have several di…|04 17, 2014|A1A45IW850QZBT|
AnneMarie|  They are too short|    1397692800|
|7245456313| [0, 0]|    5.0|the bands are gre…|12 25, 2011|A2AECU5QSJ6UB7|
arnold|               super|    1324771200|
|7245456313| [1, 1]|    5.0|I absolutely love…|07 14, 2013|A1FKGHZYT64TYG|
Ashley Y.|       Lovin' these!|    1373760000|
|7245456313| [0, 0]|    5.0|My wife and I use…|07 22, 2013|A3V65EQUFDQ1FL|
Austen Hayes|     Excellent bands|    1374451200|
+----------+-------+-------+--------------------+-----------+--------------+----
----------------+--------------------+--------------+
only showing top 35 rows
```

[4]:

```python
#Download dataset if not exists and read it as spark dataframe
try:
    df0 = spark.read.json("reviews_Sports_and_Outdoors_5.json.gz")
except Exception as e:
    url = "http://snap.stanford.edu/data/amazon/productGraph/categoryFiles/
 ↪reviews_Sports_and_Outdoors_5.json.gz"
    wget.download(url)
    df0 = spark.read.json("reviews_Sports_and_Outdoors_5.json.gz")

df = df0.withColumn("text",concat(col("summary"), lit(" "),col("reviewText")))\
 .drop("helpful")\
 .drop("reviewerID")\
 .drop("reviewerName")\
 .drop("reviewTime")
```

```
df.count()
```

[4]: 296337

[5]:

```
print(spark.version)
```

```
3.0.0
```

[6]:

```
df.describe("overall").show()
```

```
[Stage 5:>                                                          (0 + 1) / 1]

+-------+------------------+
|summary|           overall|
+-------+------------------+
|  count|            296337|
|   mean| 4.393450699710128|
| stddev|0.9869053992908551|
|    min|               1.0|
|    max|               5.0|
+-------+------------------+
```

[7]:

```
#Bucketize data and create labels 0 if overall rating is in (1.0,2.0),␣
 ↪otherwise 1
df1 = df.filter("overall !=3")

splits = [-float("inf"), 4.0, float("inf")]

bucketizer = Bucketizer(splits=splits, inputCol="overall", outputCol="label")

df2= bucketizer.transform(df1)

df2.groupBy("overall","label").count().show()
```

```
+-------+-----+------+
|overall|label| count|
+-------+-----+------+
|    2.0|  0.0| 10204|
|    5.0|  1.0|188208|
|    1.0|  0.0|  9045|
|    4.0|  1.0| 64809|
+-------+-----+-------+
```

[8]:

```
#take sample to create train and test dataset
fractions = {1.0 : .1, 0.0 : 1.0}
df3 = df2.stat.sampleBy("label", fractions, 36)
df3.groupBy("label").count().show()
```

```
+-----+-----+
|label|count|
+-----+-----+
|  0.0|19249|
|  1.0|25224|
+-----+-----+
```

[9]:

```
#Split data as 80-20% Train and Test dataset
splitSeed = 5043
trainingData, testData = df3.randomSplit([0.8, 0.2], splitSeed)
```

[10]:

```
#Tokenize
tokenizer =␣
 ↪RegexTokenizer(inputCol="text",outputCol="reviewTokensUf",pattern="\\s+|[,.
 ↪()\"]")

remover = StopWordsRemover(stopWords=StopWordsRemover.
 ↪loadDefaultStopWords("english"),inputCol="reviewTokensUf",outputCol="reviewTokens")
```

[11]:

```
#converts word documents to vectors of token counts
cv = CountVectorizer(inputCol="reviewTokens",outputCol="cv",vocabSize=296337)
```

[12]:

```
#IDF model
idf = IDF(inputCol="cv",outputCol="features")
```

[13]:

```
lr = LogisticRegression(maxIter=100,regParam=0.02,elasticNetParam=0.3)
```

[14]:

```
#Creates a pipeline
steps =  [tokenizer, remover, cv, idf,lr]
pipeline = Pipeline(stages=steps)
```

[15]:

```
model = pipeline.fit(trainingData)
```

```
24/07/22 04:57:47 WARN DAGScheduler: Broadcasting large task binary with size
1969.4 KiB
24/07/22 04:57:53 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
```

```
24/07/22 04:57:53 WARN BLAS: Failed to load implementation from:
com.github.fommil.netlib.NativeSystemBLAS
24/07/22 04:57:53 WARN BLAS: Failed to load implementation from:
com.github.fommil.netlib.NativeRefBLAS
24/07/22 04:57:53 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:54 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:54 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:54 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:54 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:55 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:55 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:55 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:55 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:55 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:55 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:56 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:56 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:56 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:56 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:56 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:57 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:57 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:57 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:57 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:57 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:57 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
```

```
24/07/22 04:57:58 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:58 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:58 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:58 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:58 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:59 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:59 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:59 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:59 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:57:59 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:00 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:00 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:00 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:00 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:00 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:00 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:01 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:01 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:01 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:01 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:01 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:02 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:02 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:02 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
```

```
24/07/22 04:58:02 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:02 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:02 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:03 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:03 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:03 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:03 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:03 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:03 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:04 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:04 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:04 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:04 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:04 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:04 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:05 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:05 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:05 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:05 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:05 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:05 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:06 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:06 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:06 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
```

```
24/07/22 04:58:06 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:06 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:06 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:07 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:07 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:07 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:07 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:07 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:07 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:08 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:08 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:08 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:08 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:08 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:08 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:09 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:09 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:09 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:09 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:09 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:09 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:10 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:10 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:10 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
```

```
24/07/22 04:58:10 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:10 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:10 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
24/07/22 04:58:11 WARN DAGScheduler: Broadcasting large task binary with size
1968.6 KiB
```

[16]:
```
#collecting all metrics
vocabulary = model.stages[2].vocabulary
weights = model.stages[-1].coefficients.toArray()
weights = [float(weight) for weight in weights]
```

[17]:
```
schema = StructType([StructField('word', StringType()),
                     StructField('weight', FloatType())
                     ])
cdf = spark.createDataFrame(zip(vocabulary, weights), schema)
```

[18]:
```
cdf.orderBy(desc("weight")).show(10)
```

```
+---------+----------+
|     word|    weight|
+---------+----------+
|    great| 0.5876225|
|   thoses|  0.325535|
|  perfect|0.32343474|
|     easy| 0.2615016|
|   highly|0.25427502|
|     love|0.23299988|
|excellent|0.22146676|
|     nice|0.21586789|
|     good|0.20862874|
|    works|0.20269535|
+---------+----------+
only showing top 10 rows
```

[19]:
```
cdf.orderBy("weight").show(10)
```

```
+-------------+-----------+
|         word|     weight|
+-------------+-----------+
|     returned|-0.38842562|
|         poor|-0.33077022|
|      useless|-0.30299458|
|        waste|-0.27846226|
|        broke|-0.26966578|
|         junk| -0.2493974|
|       return|-0.24831308|
|disappointing|-0.22999014|
|    returning|-0.21706156|
| disappointed|-0.21414408|
+-------------+-----------+
only showing top 10 rows
```

[20]:

```
predictions = model.transform(testData)
```

[21]:

```
evaluator = BinaryClassificationEvaluator()
areaUnderROC = evaluator.evaluate(predictions)
```

```
24/07/22 04:58:13 WARN DAGScheduler: Broadcasting large task binary with size
1986.0 KiB
```

[22]:

```
predictions.show()
```

```
24/07/22 04:58:17 WARN DAGScheduler: Broadcasting large task binary with size
2003.5 KiB
[Stage 143:>                                                        (0 + 1) / 1]

+----------+-------+--------------------+--------------------+--------------+---
-----------------+-----+--------------------+--------------------+--------------
------+--------------------+--------------------+--------------------+----------
+
|      asin|overall|          reviewText|             summary|unixReviewTime|
text|label|      reviewTokensUf|        reviewTokens|                  cv|
features|       rawPrediction|         probability|prediction|
+----------+-------+--------------------+--------------------+--------------+---
-----------------+-----+--------------------+--------------------+--------------
------+--------------------+--------------------+--------------------+----------
+
|7245456313|    1.0|I wish I would ha…|Defective - Be Ca…|
1354492800|Defective - Be Ca…|  0.0|[defective, -, be…|[defective, -, ca…|
(71899,[0,11,15,1…|(71899,[0,11,15,1…|[1.99668098749145…|[0.88044816229074
…|       0.0|
|7245456313|    5.0|I bought this ban…|Great product, aw…|
1400112000|Great product, aw…|  1.0|[great, product, …|[great, product, …|
(71899,[0,1,2,4,5…|(71899,[0,1,2,4,5…|[-3.3010460840682…|[0.03553531986812
…|       1.0|
|7245456313|    5.0|I used to be a pe…|GREAT product for…|
1304899200|GREAT product for…|  1.0|[great, product, …|[great, product, …|
(71899,[1,8,16,20…|(71899,[1,8,16,20…|[-1.4400664803236…|[0.19153505391365
…|       1.0|
|7245456313|    5.0|My arms are burni…|Love Love Love th…|
1358985600|Love Love Love th…|  1.0|[love, love, love…|[love, love, love…|
(71899,[6,10,13,3…|(71899,[6,10,13,3…|[-3.2229472289436…|[0.03831125139765
…|       1.0|
|B00000IURU|    5.0|Use this at pre s…|Toddlers love thi…|
1400803200|Toddlers love thi…|  1.0|[toddlers, love, …|[toddlers, love, …|
(71899,[4,18,38,3…|(71899,[4,18,38,3…|[-0.6913907181625…|[0.33372377250877
…|       1.0|
|B00000J6JO|    1.0|As I write this r…|Very Cheaply made…|
1369699200|Very Cheaply made…|  0.0|[very, cheaply, m…|[cheaply, made, p…|
(71899,[4,8,14,18…|(71899,[4,8,14,18…|[1.19648991575829…|[0.76789977056535
…|       0.0|
|B00000J6JO|    4.0|I saw a lot of ne…|Really good gift …|
1401148800|Really good gift …|  1.0|[really, good, gi…|[really, good, gi…|
(71899,[0,2,3,4,5…|(71899,[0,2,3,4,5…|[-0.4114566537646…|[0.39856289439000
…|       1.0|
|B0000224UE|    5.0|I was given this …|   Always by my side|
1361404800|Always by my side…|  1.0|[always, by, my, …|[always, side, gi…|
(71899,[0,4,5,6,1…|(71899,[0,4,5,6,1…|[0.07039114776984…|[0.51759052424809
…|       0.0|
|B0000224UE|    5.0|The victor inbox …|Victorinox Multi-…|
1361923200|Victorinox Multi-…|  1.0|[victorinox, mult…|[victorinox, mult…|
(71899,[6,14,16,1…|(71899,[6,14,16,1…|[-1.1410191572624…|[0.24213329163981
…|       1.0|
|B000030056|    1.0|Cheap product!  W…|       Cheap product|
1309564800|Cheap product Che…|  0.0|[cheap, product, …|[cheap, product, …|
(71899,[4,8,14,17…|(71899,[4,8,14,17…|[1.16342300609821…|[0.76195413639682
…|       0.0|
|B00003CYPK|    5.0|Trac Ball is just…|One of the best b…|    1226188800|One
of the best b…|  1.0|[one, of, the, be…|[one, best, backy…|(71899,[0,3,5,1
3,…|(71899,[0,3,5,13,…|[-3.1076942546118…|[0.04279098786357…|       1.0|
|B00004NKIQ|    5.0|This net is great…|excellent net for…|
1341532800|excellent net for…|  1.0|[excellent, net, …|[excellent, net, …|
(71899,[6,14,21,2…|(71899,[6,14,21,2…|[-2.1941902835827…|[0.10027341808034
…|       1.0|
|B00004SQM7|    2.0|This must be more…|          Didn't Fit|
1307404800|Didn't Fit This m…|  0.0|[didn't, fit, thi…|[fit, must, ideal…|
(71899,[9,39,41,4…|(71899,[9,39,41,4…|[2.05943139774295…|[0.88689714564144
…|       0.0|
|B00004SQM9|    1.0|This lock jammed …|            Not good|    1234656000|Not
good This loc…|  0.0|[not, good, this,…|[good, lock, jamm…|(71899,[2,60,11
3,…|(71899,[2,60,113,…|[0.27827736694020…|[0.56912384699701…|       0.0|
|B00004SQM9|    2.0|Works great on fi…|doesn't work well…|
1272326400|doesn't work well…|  0.0|[doesn't, work, w…|[work, well, leve…|
(71899,[1,5,6,11,…|(71899,[1,5,6,11,…|[-0.9750220790220…|[0.27388062786153
…|       1.0|
|B00004SQM9|    4.0|This is a great a…|Works for multipl…|
1369180800|Works for multipl…|  1.0|[works, for, mult…|[works, multiple,…|
(71899,[0,1,5,15,…|(71899,[0,1,5,15,…|[-1.4518081953470…|[0.18972343897225
…|       1.0|
|B00004SQM9|    5.0|I like the combin…|Very good trigger…|
1342051200|Very good trigger…|  1.0|[very, good, trig…|[good, trigger, l…|
(71899,[2,3,26,44…|(71899,[2,3,26,44…|[-0.9615364235615…|[0.27657068276025
…|       1.0|
|B00004T1JW|    5.0|These bases are v…|Very nice set of …|
1357084800|Very nice set of …|  1.0|[very, nice, set,…|[nice, set, bases…|
(71899,[6,13,15,1…|(71899,[6,13,15,1…|[-1.5148134864671…|[0.18022652998300
…|       1.0|
|B00004THDC|    4.0|Excellent optics…|    Excellent Optics|
1386288000|Excellent Optics …|  1.0|[excellent, optic…|[excellent, optic…|
(71899,[6,20,22,2…|(71899,[6,20,22,2…|[-2.9582846383313…|[0.04934641382990
…|       1.0|
|B00004TQ2P|    5.0|Good for kids and…|  Good for my nephew|
1382832000|Good for my nephe…|  1.0|[good, for, my, n…|[good, nephew, go…|
(71899,[2,31,47,7…|(71899,[2,31,47,7…|[-0.5040760931476…|[0.37658325100325
…|       1.0|
+----------+-------+--------------------+--------------------+--------------+---
-----------------+-----+--------------------+--------------------+--------------
------+--------------------+--------------------+--------------------+----------
+
only showing top 20 rows
```

[23]:

```
#model evaluation
lp = predictions.select("label", "prediction")
counttotal = predictions.count()
correct = lp.filter(col("label") == col("prediction")).count()
wrong = lp.filter(~(col("label") == col("prediction"))).count()
ratioWrong = float(wrong) / float(counttotal)
lp = predictions.select(  "prediction","label")
counttotal = float(predictions.count())
correct = lp.filter(col("label") == col("prediction")).count()
wrong = lp.filter("label != prediction").count()
ratioWrong=wrong/counttotal
ratioCorrect=correct/counttotal
trueneg =( lp.filter(col("label") == 0.0).filter(col("label") ==␣
  ↪col("prediction")).count()) /counttotal
truepos = (lp.filter(col("label") == 1.0).filter(col("label") ==␣
  ↪col("prediction")).count())/counttotal
falseneg = (lp.filter(col("label") == 0.0).filter(~(col("label") ==␣
  ↪col("prediction"))).count())/counttotal
falsepos = (lp.filter(col("label") == 1.0).filter(~(col("label") ==␣
  ↪col("prediction"))).count())/counttotal
```

```

precision= truepos / (truepos + falsepos)
recall= truepos / (truepos + falseneg)
#fmeasure= 2  precision  recall / (precision + recall)
accuracy=(truepos + trueneg) / (truepos + trueneg + falsepos + falseneg)
```

```
24/07/22 04:58:22 WARN DAGScheduler: Broadcasting large task binary with size
1983.3 KiB
24/07/22 04:58:25 WARN DAGScheduler: Broadcasting large task binary with size
1983.3 KiB
24/07/22 04:58:31 WARN DAGScheduler: Broadcasting large task binary with size
1983.3 KiB
24/07/22 04:58:34 WARN DAGScheduler: Broadcasting large task binary with size
1983.3 KiB
24/07/22 04:58:37 WARN DAGScheduler: Broadcasting large task binary with size
1983.5 KiB
24/07/22 04:58:40 WARN DAGScheduler: Broadcasting large task binary with size
1983.5 KiB
24/07/22 04:58:43 WARN DAGScheduler: Broadcasting large task binary with size
1983.6 KiB
24/07/22 04:58:46 WARN DAGScheduler: Broadcasting large task binary with size
1983.6 KiB
```

[24]:

```
print('counttotal   :', counttotal     )
print('correct      :', correct        )
print('wrong        :', wrong          )
print('ratioWrong   :', ratioWrong     )
print('ratioCorrect :', ratioCorrect   )
print('truen        :', trueneg          )
print('truep        :', truepos          )
print('falsen       :', falseneg         )
print('falsep       :', falsepos         )
print('precision    :', precision      )
print('recall       :', recall         )
#print('fmeasure     :', fmeasure       )
print('accuracy     :', accuracy       )
```

```
counttotal   : 9003.0
correct      : 7776
wrong        : 1227
ratioWrong   : 0.13628790403198934
ratioCorrect : 0.8637120959680107
truen        : 0.3361101854937243
truep        : 0.5276019104742864
falsen       : 0.08863712095968011
falsep       : 0.04765078307230923
precision    : 0.9171654759606103
recall       : 0.8561643835616438
```

```
accuracy     : 0.8637120959680107
```

[25]:
```
predictions.filter(col("prediction") == 0.0)\
.select("summary","reviewTokens","overall","prediction")\
.orderBy(desc("rawPrediction")).show(5)
```

```
24/07/22 04:58:49 WARN DAGScheduler: Broadcasting large task binary with size
1996.3 KiB
[Stage 164:>                                                        (0 + 1) / 1]

+--------------------+--------------------+-------+----------+
|             summary|        reviewTokens|overall|prediction|
+--------------------+--------------------+-------+----------+
|Buyer Beware - Yo…|[buyer, beware, -…|    2.0|       0.0|
|Awful Phone and T…|[awful, phone, te…|    1.0|       0.0|
|DO NOT BUY HERE I…|[buy, need, custo…|    1.0|       0.0|
|                JUNK|[junk, well, rece…|    1.0|       0.0|
|Poor 3-9x40 Hamme…|[poor, 3-9x40, ha…|    1.0|       0.0|
+--------------------+--------------------+-------+----------+
only showing top 5 rows
```

[26]:
```
predictions.filter(col("prediction")== 1.0)\
.select("summary","reviewTokens","overall","prediction")\
.orderBy("rawPrediction").show(5)
```

```
24/07/22 04:58:53 WARN DAGScheduler: Broadcasting large task binary with size
1996.2 KiB
[Stage 165:>                                                        (0 + 1) / 1]

+--------------------+--------------------+-------+----------+
|             summary|        reviewTokens|overall|prediction|
+--------------------+--------------------+-------+----------+
|My DROID Story an…|[droid, story, co…|    5.0|       1.0|
| great trucker phone|[great, trucker, …|    5.0|       1.0|
|    Favorite EDC Bag|[favorite, edc, b…|    4.0|       1.0|
|One of My Favorit…|[one, favorites!!…|    4.0|       1.0|
|Best Hopper I've …|[best, hopper, us…|    4.0|       1.0|
+--------------------+--------------------+-------+----------+
only showing top 5 rows
```

[27]:
```
dir = "sentiment/"
model.write().overwrite().save(dir)
```

```
24/07/22 04:58:58 WARN TaskSetManager: Stage 170 contains a task of very large
```

```
size (1385 KiB). The maximum recommended task size is 1000 KiB.
24/07/22 04:58:59 WARN TaskSetManager: Stage 173 contains a task of very large
size (1153 KiB). The maximum recommended task size is 1000 KiB.
```

[33]:

```python
dir = "sentiment/"
model = PipelineModel.load(dir)
```

[44]:

```python
df = spark.read.format("mongo").load()
df.printSchema()
```

```
root
 |-- _id: string (nullable = true)
 |-- prediction: double (nullable = true)
 |-- text: string (nullable = true)
 |-- timestamp_ms: string (nullable = true)


24/07/22 06:17:20 WARN MongoInferSchema: Field '_id' contains conflicting types
converting to StringType
```

[52]:

```python
df = spark.read.format("mongo").load().select("timestamp_ms","text")
# Kiểm tra số lượng giá trị null trong cột 'text'
null_count = df.filter(df.text.isNull()).count()
print(f"Số lượng giá trị null trong cột 'text': {null_count}")
```

```
Số lượng giá trị null trong cột 'text': 1

24/07/22 06:20:38 WARN MongoInferSchema: Field '_id' contains conflicting types
converting to StringType
```

[53]:

```python
# Loại bỏ các hàng có giá trị null trong cột 'text'
df = df.na.drop(subset=["text"])
```

[54]:

```python
df.printSchema()
```

```
root
 |-- timestamp_ms: string (nullable = true)
 |-- text: string (nullable = true)
```

[67]:

```python
splits = [-float("inf"), 0, float("inf")]
#bucketizer = Bucketizer(inputCol="timestamp_ms",outputCol="sentiment",splits=splits)

#df5= bucketizer.transform(df)
predictions = model.transform(df)
result_df = predictions.select('text', 'prediction')

# Chuyển đổi sang Pandas DataFrame
```

```
result_pd = result_df.toPandas()
```

```
24/07/22 06:34:39 WARN DAGScheduler: Broadcasting large task binary with size
1951.2 KiB
```

[83]:
```
# Định dạng và hiển thị kết quả
styled_result = result_pd.style.set_table_attributes('style="width: 80%;␣
 ↪border-collapse: collapse;"') \
                                .set_caption("Dự đoán Sentiment") \
                                .highlight_max(color='lightgreen') \
                                .highlight_min(color='lightcoral') \
                                    .set_properties(**{'text-align': 'left'})


# Hiển thị kết quả
styled_result
```

[83]:
```
<pandas.io.formats.style.Styler at 0x71521a769a90>
```

[ ]:
```

```