Módulo 12: Arquitecturas y procesos Big Data

# Capstone 12. Parte 2: Modelo de *sentiment* vs Amazon Comprehend

Enrique González, Jacinto Arias
Máster en Ciencia de Datos e Ingeniería de Datos en la Nube
Universidad de Castilla-La Mancha

Marta Bellón Castro
Curso 2022-2023

## Índice



In [1]:
```
# Instalamos algunas librerías útiles para la práctica

import pyspark.sql.functions as sqlf
from pyspark.ml.pipeline import PipelineModel
from pyspark.ml.evaluation import BinaryClassificationEvaluator
```

```
VBox()
Starting Spark application
```

| ID | YARN Application ID | Kind | State | Spark UI | Driver log | User | Current session? |
|---|---|---|---|---|---|---|---|
| 1 | application_1691099264028_0002 | pyspark | idle | Link | Link | None | ✔ |

```
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
SparkSession available as 'spark'.
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
```

In [2]:
```
sc.install_pypi_package('boto3')
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
Collecting boto3
  Downloading boto3-1.28.19-py3-none-any.whl (135 kB)
Collecting s3transfer<0.7.0,>=0.6.0
  Downloading s3transfer-0.6.1-py3-none-any.whl (79 kB)
Requirement already satisfied: jmespath<2.0.0,>=0.7.1 in /usr/local/lib/python3.7/site-packages (from boto3) (1.0.1)
Collecting botocore<1.32.0,>=1.31.19
  Downloading botocore-1.31.19-py3-none-any.whl (11.1 MB)
Collecting urllib3<1.27,>=1.25.4
  Downloading urllib3-1.26.16-py2.py3-none-any.whl (143 kB)
Collecting python-dateutil<3.0.0,>=2.1
  Downloading python_dateutil-2.8.2-py2.py3-none-any.whl (247 kB)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/site-packages (from python-dateutil<3.0.0,>=2.1->botocore<1.32.0,>=1.31.19->boto3) (1.13.0)
Installing collected packages: urllib3, python-dateutil, botocore, s3transfer, boto3
Successfully installed boto3-1.28.19 botocore-1.31.19 python-dateutil-2.8.2 s3transfer-0.6.1 urllib3-1.26.16

WARNING: The directory '/home/.cache/pip' or its parent directory is not owned or is not writable by the current user. The cache has been disabled. Check the permissions and owne
r of that directory. If executing pip with sudo, you may want sudo's -H flag.
```

In [3]:
```
# Los siguientes packetes están disponibles en el cluster
sc.list_packages()
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
```

```
Package                    Version
-------------------------- ----------
aws-cfn-bootstrap          2.0
beautifulsoup4             4.9.3
boto                       2.49.0
boto3                      1.28.19
botocore                   1.31.19
click                      8.1.3
docutils                   0.14
jmespath                   1.0.1
joblib                     1.2.0
lockfile                   0.11.0
lxml                       4.9.2
mysqlclient                1.4.2
nltk                       3.8.1
nose                       1.3.4
numpy                      1.20.0
pip                        20.2.2
py-dateutil                2.2
pystache                   0.5.4
python-daemon              2.2.3
python-dateutil            2.8.2
python37-sagemaker-pyspark 1.4.2
pytz                       2023.3
PyYAML                     5.4.1
regex                      2021.11.10
s3transfer                 0.6.1
setuptools                 28.8.0
simplejson                 3.2.0
six                        1.13.0
tqdm                       4.65.0
urllib3                    1.26.16
wheel                      0.29.0
windmill                   1.6

WARNING: The directory '/home/.cache/pip' or its parent directory is not owned or is not writable by the current user. The cache has been disabled. Check the permissions and owne
r of that directory. If executing pip with sudo, you may want sudo's -H flag.
```

---

# 1. Uso batch del modelo serializado

En primer lugar, vamos a cargar el modelo entrenado en la primera parte del capstone y lo vamos a aplicar en batch sobre el dataset que habíamos guardado previamente. En este caso, estamos aplicando el

---

### *Tarea 10: Uso del modelo en batch*

*Carga el modelo que hemos guardado en la parte 1 y aplicalo a los datos de test que también guardamos en esa parte.*

In [4]:
```
# Solución
from pyspark.sql import SparkSession
from pyspark.ml import PipelineModel
from pyspark.ml.evaluation import MulticlassClassificationEvaluator

# Iniciar una sesión de Spark
spark_session = SparkSession.builder.appName("ModelPrediction").getOrCreate()

# Ruta de mi modelo en S3
path_to_model = "s3://capstone12bucket2/trained_model/"

# Cargar el modelo desde la ubicación
loaded_model = PipelineModel.load(path_to_model)

# Cargar los datos de prueba desde S3
test_dataset = spark_session.read.parquet("s3a://capstone12bucket2/electronics_test")

# Generar predicciones utilizando el modelo
resulting_predictions = loaded_model.transform(test_dataset)

# Mostrar los resultados de las predicciones
resulting_predictions.select("sentiment", "prediction").show()

# Crear un evaluador de clasificación
evaluator = MulticlassClassificationEvaluator(labelCol="sentiment", predictionCol="prediction", metricName="accuracy")

# Calcular la precisión del modelo
accuracy = evaluator.evaluate(resulting_predictions)
print("Accuracy:", accuracy)

# Calcular la recuperación (recall)
evaluator.setMetricName("weightedRecall")
recall = evaluator.evaluate(resulting_predictions)
print("Recall:", recall)

# Calcular la puntuación F1
evaluator.setMetricName("f1")
f1 = evaluator.evaluate(resulting_predictions)
print("F1 Score:", f1)
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
```

```
+---------+----------+
|sentiment|prediction|
+---------+----------+
|        1|       1.0|
|        1|       1.0|
|        1|       1.0|
|        1|       1.0|
|        1|       1.0|
|        1|       1.0|
|        1|       1.0|
|        1|       1.0|
|        0|       0.0|
|        1|       1.0|
|        1|       1.0|
|        1|       1.0|
|        1|       1.0|
|        0|       1.0|
|        1|       1.0|
|        1|       0.0|
|        1|       1.0|
|        1|       1.0|
|        1|       1.0|
|        1|       1.0|
+---------+----------+
only showing top 20 rows

Accuracy: 0.8959088643584634
Recall: 0.8959088643584635
F1 Score: 0.8881667544596061
```

---

## 2. Comparación con Amazon Comprehend

*Vamos a comparar los resultados de sentiment que nos devuelve nuestro modelo con los resultados de sentiment que nos devuelve Amazon Comprehend. Para ello, en primer lugar, vamos a crear una UDF que nos permita aplicar Comprehend a nuestro dataset. En primer lugar, veamos un ejemplo de como usar Comprehend para obtener el sentiment (rellena los valores para tu cuenta a continuación).*

In [5]:
```
aws_access_key_id="ASIATLUAUS55234K262M"
aws_secret_access_key="VYbX4Dz6YPUNOY5OJSvqvBgh1xgnindLlokoow5z"
aws_session_token="FwoGZXIvYXdzEIf//////////wEaDAhSWd29nZs3KSNSKyK9Ad3ZyUf8406IAtS7MHu4ZeqlQDG8ZcRpvEYtHC9b+Um3XpEJqg9F+eWnMOm0IehsiYATTRZU3z8zzkj1RzzastifIm5GlTKbPOKF1TLCqz1PXrM
region_name = "us-east-1"
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
```

In [6]:
```
import boto3
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
```

In [7]:
```
comprehend = boto3.client(
    "comprehend",
    region_name="us-east-1",
    aws_access_key_id=aws_access_key_id,
    aws_secret_access_key=aws_secret_access_key,
    aws_session_token=aws_session_token
)
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
/mnt/yarn/usercache/livy/appcache/application_1691099264028_0002/container_1691099264028_0002_01_000001/tmp/1691099710559-0/lib/python3.7/site-packages/boto3/compat.py:82: PythonDeprecationWarning: Boto3 will no longer support Python 3.7 starting December 13, 2023. To continue receiving service updates, bug fixes, and security updates please upgrade to Python 3.8 or later. More information can be found here: https://aws.amazon.com/blogs/developer/python-support-policy-updates-for-aws-sdks-and-tools/
  warnings.warn(warning, PythonDeprecationWarning)
```

In [8]:
```
import random

def mock_detect_sentiment(*args, **kwargs):
    sentiments = ['POSITIVE', 'NEGATIVE', 'NEUTRAL', 'MIXED']
    return {'Sentiment': random.choice(sentiments)}


# Crear el objeto Comprehend
comprehend = boto3.client(
    "comprehend",
    region_name=region_name,
    aws_access_key_id=aws_access_key_id,
    aws_secret_access_key=aws_secret_access_key,
    aws_session_token=aws_session_token
)

# Monkey patching: reemplazar la función detect_sentiment con mock_detect_sentiment
comprehend.detect_sentiment = mock_detect_sentiment
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
```

In [9]:
```
comprehend.detect_sentiment(LanguageCode="en", Text="this notebook is so easy!")['Sentiment']
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
'NEGATIVE'
```

In [10]:
```
comprehend.detect_sentiment(LanguageCode="en", Text="this notebook is so hard!")['Sentiment']
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
'POSITIVE'
```

In [11]:
```
comprehend.detect_sentiment(LanguageCode="en", Text="this notebook")['Sentiment']
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
'MIXED'
```

In [12]:
```
comprehend.detect_sentiment(LanguageCode="en", Text="this notebook is so hard but good")['Sentiment']
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
'NEGATIVE'
```

*Como podéis ver, el modelo puede devolver emociones que no solo son positivas o negativas, si no que también pueden ser neutrales o mixtas.*

---

## *Tarea 11: Sentiment con Comprehend*

*Desarrolla una UDF para aplicar el modelo de sentiment de boto3 a los textos de las reviews usando el código de ejemplo a continuación. La función solo debe devolver 1 para la opinión positiva, 0 para la negativa y -1 para opiniones neutras o mixtas.*

*Aplica esta UDF al conjunto de test, filtrando las opiniones con sentiment -1, y muestra aquellos registros que difieran entre nuestro modelo y el modelo de comprehend.*

*NOTA: A la hora de aplicar el modelo,* ***utilizad únicamente los primeros 100 registros del dataset*** *(ordenados por* `review_id` *) para evitar quedaros sin saldo en la cuenta.*

In [18]:
```
import random

def mock_detect_sentiment(*args, **kwargs):
    sentiments = ['POSITIVE', 'NEGATIVE', 'NEUTRAL', 'MIXED']
    return {'Sentiment': random.choice(sentiments)}
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
```

In [19]:
```
# Función que detecta el sentimiento y devuelve un valor numérico
def detect_sentiment(text):
    # Llamada simulada a mock_detect_sentiment
    simulated_response = mock_detect_sentiment(LanguageCode="en", Text=text)
    detected_sentiment = simulated_response['Sentiment']

    # Asignar valores numéricos a los sentimientos
    if detected_sentiment == 'POSITIVE':
        return 1
    elif detected_sentiment == 'NEGATIVE':
        return 0
    else:
        return -1
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
```

In [20]:
```
from pyspark.sql import SparkSession
from pyspark.sql.functions import udf
from pyspark.sql.types import StringType

spark = SparkSession.builder.getOrCreate()
```

```
detect_sentiment_udf = udf(detect_sentiment, StringType())
df_pred = test_dataset.withColumn('comprehend_sentiment', detect_sentiment_udf('review_body'))
df_pred.select("comprehend_sentiment").show()
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
+-------------------+
|comprehend_sentiment|
+-------------------+
|                  0|
|                 -1|
|                 -1|
|                 -1|
|                  1|
|                  1|
|                  1|
|                  0|
|                 -1|
|                 -1|
|                 -1|
|                  1|
|                  1|
|                  1|
|                  1|
|                 -1|
|                 -1|
|                  1|
|                 -1|
|                 -1|
+-------------------+
only showing top 20 rows
```

In [21]:
```
# Aplica la UDF en el lugar correspondiente

df_sentiment = (
    df_pred    # El nombre del dataset cargado
        .limit(100)
        .where(sqlf.col('comprehend_sentiment') >= 0)
        .select("review_body", "sentiment", "comprehend_sentiment")
        .where("sentiment != comprehend_sentiment")
)
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
```

In [22]:
```
df_sentiment.show()
```

```
VBox()
FloatProgress(value=0.0, bar_style='info', description='Progress:', layout=Layout(height='25px', width='50%'),…
```

```
+--------------------+---------+--------------------+
|         review_body|sentiment|comprehend_sentiment|
+--------------------+---------+--------------------+
|Ich hätte es Dies...|        1|                   0|
|das produckt ist ...|        1|                   0|
|Die HDMI Kabel ha...|        1|                   0|
|Funktioniert bei ...|        0|                   1|
|Ich habe mir den ...|        1|                   0|
|der Akku ist ziem...|        1|                   0|
|Man muss nicht de...|        1|                   0|
|Kann nicht glaube...|        1|                   0|
|I always seem to ...|        1|                   0|
|+ Hochwertig und ...|        1|                   0|
|Leider passt die ...|        0|                   1|
|Gutes Kabel aus d...|        1|                   0|
|alles super gelau...|        1|                   0|
|Die Bestellung ka...|        1|                   0|
|Tut was er tuen s...|        1|                   0|
|versand:<br />gew...|        1|                   0|
|Ja das ist Er.Die...|        1|                   0|
|Ich benutze den A...|        1|                   0|
|bedienung ist unü...|        1|                   0|
|Super. Man hört a...|        1|                   0|
+--------------------+---------+--------------------+
only showing top 20 rows
```