In [1]:
```
# Team 8(BIA)
import os
from pyspark import SparkConf, SparkContext
```

```
Starting Spark application
```

ID	YARN Application ID	Kind	State	
10	application_1607846600195_0026	pyspark	idle	Link (http://ip-1 128.ec2.internal:20888/proxy/application_16078466001

```
SparkSession available as 'spark'.
```

In [2]:
```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName('breast_cancer_tweet').getOrCreate()
```

▸ Spark Job Progress

In [3]:
```
%%configure -f
{ "conf":{
          "spark.pyspark.python": "python",
          "spark.pyspark.virtualenv.enabled": "true",
          "spark.pyspark.virtualenv.type":"native",
          "spark.pyspark.virtualenv.bin.path":"/usr/bin/virtualenv"
         }
}
```

```
Starting Spark application
```

ID	YARN Application ID	Kind	State	
11	application_1607846600195_0028	pyspark	idle	Link (http://ip-1 128.ec2.internal:20888/proxy/application_16078466001

```
SparkSession available as 'spark'.
```

Current session configs: `{'conf': {'spark.pyspark.python': 'python', 'spark.pyspark.virtualenv.enabled': 'true', 'spark.pyspark.virtualenv.type': 'native', 'spark.pyspark.virtualenv.bin.path': '/usr/bin/virtualenv'}, 'kind': 'pyspark'}`

ID	YARN Application ID	Kind	State	
11	application_1607846600195_0028	pyspark	idle	Link (http://ip-1 128.ec2.internal:20888/proxy/application_16078466001

In [4]:
```
sc.install_pypi_package("pandas")
sc.install_pypi_package("s3fs")
sc.install_pypi_package("matplotlib")
```

▸ Spark Job Progress

```
Collecting pandas
  Using cached https://files.pythonhosted.org/packages/db/83/7d4008ffc2988066ff37f6a0bb6d7b60822367dcb36ba5e39aa7801fda54/pandas-0.24.2-cp27-cp27mu-manylinux1_x86_64.whl (https://files.pythonhosted.org/packages/db/83/7d4008ffc2988066ff37f6a0bb6d7b60822367dcb36ba5e39aa7801fda54/pandas-0.24.2-cp27-cp27mu-manylinux1_x86_64.whl)
Requirement already satisfied: python-dateutil>=2.5.0 in /usr/lib/python2.7/site-packages (from pandas)
Requirement already satisfied: numpy>=1.12.0 in /usr/lib64/python2.7/site-packages (from pandas)
Collecting pytz>=2011k (from pandas)
  Using cached https://files.pythonhosted.org/packages/12/f8/ff09af6ff61a3efaad5f61ba5facdf17e7722c4393f7d8a66674d2dbd29f/pytz-2020.4-py2.py3-none-any.whl (https://files.pythonhosted.org/packages/12/f8/ff09af6ff61a3efaad5f61ba5facdf17e7722c4393f7d8a66674d2dbd29f/pytz-2020.4-py2.py3-none-any.whl)
Requirement already satisfied: six>=1.5 in /usr/lib/python2.7/site-packages (from python-dateutil>=2.5.0->pandas)
Installing collected packages: pytz, pandas
Successfully installed pandas-0.24.2 pytz-2020.4

Collecting s3fs
Requirement already satisfied: botocore>=1.12.91 in /usr/lib/python2.7/site-packages (from s3fs)
Collecting six>=1.12.0 (from s3fs)
  Using cached https://files.pythonhosted.org/packages/ee/ff/48bde5c0f013094d729fe4b0316ba2a24774b3ff1c52d924a8a4cb04078a/six-1.15.0-py2.py3-none-any.whl (https://files.pythonhosted.org/packages/ee/ff/48bde5c0f013094d729fe4b0316ba2a24774b3ff1c52d924a8a4cb04078a/six-1.15.0-py2.py3-none-any.whl)
Collecting boto3>=1.9.91 (from s3fs)
  Using cached https://files.pythonhosted.org/packages/87/3e/3a4546165383a5fc9f6f7ba15a261c768aee10662bb06105100d859e8940/boto3-1.16.35-py2.py3-none-any.whl (https://files.pythonhosted.org/packages/87/3e/3a4546165383a5fc9f6f7ba15a261c768aee10662bb06105100d859e8940/boto3-1.16.35-py2.py3-none-any.whl)
Requirement already satisfied: jmespath<1.0.0,>=0.7.1 in /usr/lib/python2.7/site-packages (from botocore>=1.12.91->s3fs)
Requirement already satisfied: python-dateutil<3.0.0,>=2.1 in /usr/lib/python2.7/site-packages (from botocore>=1.12.91->s3fs)
Requirement already satisfied: urllib3<1.26,>=1.20 in /usr/lib/python2.7/site-packages (from botocore>=1.12.91->s3fs)
Collecting s3transfer<0.4.0,>=0.3.0 (from boto3>=1.9.91->s3fs)
  Using cached https://files.pythonhosted.org/packages/69/79/e6afb3d8b0b4e96cefbdc690f741d7dd24547ff1f94240c997a26fa908d3/s3transfer-0.3.3-py2.py3-none-any.whl (https://files.pythonhosted.org/packages/69/79/e6afb3d8b0b4e96cefbdc690f741d7dd24547ff1f94240c997a26fa908d3/s3transfer-0.3.3-py2.py3-none-any.whl)
Requirement already satisfied: futures<4.0.0,>=2.2.0; python_version ==
```

```
 "2.7" in /usr/lib/python2.7/site-packages (from s3transfer<0.4.0,>=0.3.0
->boto3>=1.9.91->s3fs)
Installing collected packages: six, s3transfer, boto3, s3fs
  Found existing installation: six 1.9.0
    Not uninstalling six at /usr/lib/python2.7/site-packages, outside env
ironment /tmp/1607851047030-0
  Found existing installation: s3transfer 0.1.12
    Not uninstalling s3transfer at /usr/lib/python2.7/site-packages, outs
ide environment /tmp/1607851047030-0
Successfully installed boto3-1.16.35 s3fs-0.2.2 s3transfer-0.3.3 six-1.1
5.0

Collecting matplotlib
  Using cached https://files.pythonhosted.org/packages/9d/40/5ba7d4a3f80d
39d409f21899972596bf62c8606f1406a825029649eaa439/matplotlib-2.2.5-cp27-cp
27mu-manylinux1_x86_64.whl (https://files.pythonhosted.org/packages/9d/4
0/5ba7d4a3f80d39d409f21899972596bf62c8606f1406a825029649eaa439/matplotlib
-2.2.5-cp27-cp27mu-manylinux1_x86_64.whl)
Requirement already satisfied: numpy>=1.7.1 in /usr/lib64/python2.7/site-
packages (from matplotlib)
Collecting pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 (from matplotlib)
  Using cached https://files.pythonhosted.org/packages/8a/bb/488841f56197
b13700afd5658fc279a2025a39e22449b7cf29864669b15d/pyparsing-2.4.7-py2.py3-
none-any.whl (https://files.pythonhosted.org/packages/8a/bb/488841f56197b
13700afd5658fc279a2025a39e22449b7cf29864669b15d/pyparsing-2.4.7-py2.py3-n
one-any.whl)
Requirement already satisfied: python-dateutil>=2.1 in /usr/lib/python2.
7/site-packages (from matplotlib)
Collecting kiwisolver>=1.0.1 (from matplotlib)
  Using cached https://files.pythonhosted.org/packages/3d/78/cb9248b2289e
c31e301137cedbe4ca503a74ca87f88cdbfd2f8be52323bf/kiwisolver-1.1.0-cp27-cp
27mu-manylinux1_x86_64.whl (https://files.pythonhosted.org/packages/3d/7
8/cb9248b2289ec31e301137cedbe4ca503a74ca87f88cdbfd2f8be52323bf/kiwisolver
-1.1.0-cp27-cp27mu-manylinux1_x86_64.whl)
Collecting cycler>=0.10 (from matplotlib)
  Using cached https://files.pythonhosted.org/packages/f7/d2/e07d3ebb2bd7
af696440ce7e754c59dd546ffe1bbe732c8ab68b9c834e61/cycler-0.10.0-py2.py3-no
ne-any.whl (https://files.pythonhosted.org/packages/f7/d2/e07d3ebb2bd7af6
96440ce7e754c59dd546ffe1bbe732c8ab68b9c834e61/cycler-0.10.0-py2.py3-none-
any.whl)
Collecting subprocess32 (from matplotlib)
Requirement already satisfied: pytz in /mnt/tmp/1607851047030-0/lib/pytho
n2.7/site-packages (from matplotlib)
Requirement already satisfied: six>=1.10 in /mnt/tmp/1607851047030-0/lib/
python2.7/site-packages (from matplotlib)
Collecting backports.functools-lru-cache (from matplotlib)
  Using cached https://files.pythonhosted.org/packages/da/d1/080d2bb13773
803648281a49e3918f65b31b7beebf009887a529357fd44a/backports.functools_lru_
cache-1.6.1-py2.py3-none-any.whl (https://files.pythonhosted.org/package
s/da/d1/080d2bb13773803648281a49e3918f65b31b7beebf009887a529357fd44a/back
ports.functools_lru_cache-1.6.1-py2.py3-none-any.whl)
Requirement already satisfied: setuptools in /mnt/tmp/1607851047030-0/li
b/python2.7/site-packages (from kiwisolver>=1.0.1->matplotlib)
Installing collected packages: pyparsing, kiwisolver, cycler, subprocess3
2, backports.functools-lru-cache, matplotlib
  Found existing installation: pyparsing 1.5.6
    Not uninstalling pyparsing at /usr/lib/python2.7/site-packages, outsi
```

```
de environment /tmp/1607851047030-0
Successfully installed backports.functools-lru-cache-1.6.1 cycler-0.10.0
 kiwisolver-1.1.0 matplotlib-2.2.5 pyparsing-2.4.7 subprocess32-3.5.4
```

In [5]:
```
# data for September month
import pandas as pd
tweet1 = pd.read_csv('s3://piyudata/Untitled_data_piyu.csv')
print(len(tweet1))
tweet1.head()
```

▸ Spark Job Progress

```
5472
                       created_at  ... classes
0  Mon Sep 07 19:28:38 +0000 2020  ...       1
1  Mon Sep 07 19:28:38 +0000 2021  ...       1
2  Mon Sep 07 19:28:38 +0000 2022  ...       1
3  Mon Sep 07 19:28:38 +0000 2023  ...       1
4  Mon Sep 07 19:28:38 +0000 2024  ...       1

[5 rows x 6 columns]
```

In [6]:
```
tweet1['polarity'] = pd.to_numeric(tweet1['polarity'])
```

▸ Spark Job Progress

In [7]:
```
import re
def clean_tweet(tweet):
        '''
        Utility function to clean tweet text by removing links, special cha
        using simple regex statements.
        '''
        return ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/
cleaned = []
```

▸ Spark Job Progress

In [8]:
```
for t in range(len(tweet1)):
    cleaned.append(clean_tweet(tweet1['original_text'][t]))
tweet1['original_text'] = cleaned
tweet1['original_text'][0]
tweet1.head()
```

▸ Spark Job Progress

	created_at	...	classes
0	Mon Sep 07 19:28:38 +0000 2020	...	1
1	Mon Sep 07 19:28:38 +0000 2021	...	1
2	Mon Sep 07 19:28:38 +0000 2022	...	1
3	Mon Sep 07 19:28:38 +0000 2023	...	1
4	Mon Sep 07 19:28:38 +0000 2024	...	1

[5 rows x 6 columns]

In [9]:
```
# data for October month
import pandas as pd
tweets2 = pd.read_csv('s3://piyudata/Octo1_Data.csv')
print(len(tweets2))
tweets2.head()
```

▸ Spark Job Progress

6884

	created_at	...	classes
0	Fri Oct 04 10:46:09 +0000 2021	...	1
1	Fri Oct 04 10:46:09 +0000 2022	...	0
2	Fri Oct 04 10:46:09 +0000 2023	...	1
3	Fri Oct 04 10:46:09 +0000 2024	...	2
4	Fri Oct 04 10:46:09 +0000 2025	...	0

[5 rows x 6 columns]

In [10]:
```
cleaned = []
for t in range(len(tweets2)):
    cleaned.append(clean_tweet(tweets2['original_text'][t]))
tweets2['original_text'] = cleaned
tweets2['original_text'][len(tweets2)-1]
```

▸ Spark Job Progress

'RT Women with triple negative breast cancer need more options Dr Gursel Aktan shares why she and her team are working hard to h'

In [11]:
```
#Data for November month
import pandas as pd
tweets3 = pd.read_csv('s3://piyudata/Nove1_Data.csv')
print(len(tweets3))
tweets3.head()
```

▸ Spark Job Progress

```
5519
                       created_at  ... classes
0  Wed Nov 09 19:30:00 +0000 2020  ...       2
1  Wed Nov 09 19:30:00 +0000 2021  ...       1
2  Wed Nov 09 19:30:00 +0000 2022  ...       1
3  Wed Nov 09 19:30:00 +0000 2023  ...       1
4  Wed Nov 09 19:30:00 +0000 2024  ...       1

[5 rows x 6 columns]
```

In [12]:
```
cleaned = []
for t in range(len(tweets3)):
    cleaned.append(clean_tweet(tweets3['original_text'][t]))
tweets3['original_text'] = cleaned
tweets3['original_text'][0]
```

▸ Spark Job Progress

```
'A new study has just been released and the results state that some postm
enopausal women with a common breastcancer'
```

In [13]:
```
#creating the September dataset for analysis
df = spark.createDataFrame(tweet1)
df.printSchema()
```

▸ Spark Job Progress

```
root
 |-- created_at: string (nullable = true)
 |-- original_text: string (nullable = true)
 |-- polarity: double (nullable = true)
 |-- lang: string (nullable = true)
 |-- place: string (nullable = true)
 |-- classes: long (nullable = true)
```

In [14]:
```
from pyspark.ml.feature import HashingTF, IDF, Tokenizer, CountVectorizer
from pyspark.ml.feature import StringIndexer
from pyspark.ml import Pipeline
from pyspark.ml.classification import LogisticRegression
from pyspark.ml.evaluation import BinaryClassificationEvaluator
from pyspark.ml.feature import HashingTF, IDF, Tokenizer
from pyspark.ml.feature import StringIndexer
from pyspark.ml import Pipeline
from pyspark.mllib.regression import LabeledPoint
from pyspark.mllib.classification import NaiveBayes, NaiveBayesModel
from pyspark.mllib.classification import LogisticRegressionWithLBFGS, Logis
from pyspark.mllib.classification import SVMWithSGD, SVMModel
from pyspark.mllib.tree import DecisionTree, DecisionTreeModel
```

▸ Spark Job Progress

In [15]:
```
(train_set, val_set) = df.randomSplit([0.7, 0.3], seed = 2000)
tokenizer = Tokenizer(inputCol="original_text", outputCol="tokens")
hashtf = HashingTF(numFeatures=2**16, inputCol="tokens", outputCol='tf')
idf = IDF(inputCol='tf', outputCol="features", minDocFreq=5) #minDocFreq: r
label_stringIdx = StringIndexer(inputCol = "classes", outputCol = "label")
pipeline = Pipeline(stages=[tokenizer, hashtf, idf, label_stringIdx])
pipelineFit = pipeline.fit(train_set)
train_df = pipelineFit.transform(train_set)
val_df = pipelineFit.transform(val_set)
train_df.show(10)
```

▸ Spark Job Progress

```
+------------------+------------------+------------------+----+-----
--------+-------+--------------------+--------------------+-------------
------+-----+
|          created_at|       original_text|           polarity|lang|
place|classes|              tokens|                  tf|            featu
res|label|
+------------------+------------------+------------------+----+-----
--------+-------+--------------------+--------------------+-------------
------+-----+
|Mon Sep 07 19:28:...|Thanks for sharin...|0.35714285700000004|  en|Nashv
ille, TN|      1|[thanks, for, sha...|(65536,[463,9639,...|(65536,[463,96
39,...|  0.0|
|Mon Sep 07 19:28:...|RT The Pink Ribbo...|        0.116071429|  en|Nashv
ille, TN|      1|[rt, the, pink, r...|(65536,[13142,163...|(65536,[13142,
163...|  0.0|
|Mon Sep 07 19:28:...|Taselisib or Plac...|              -0.05|  en|Nashv
ille, TN|      1|[taselisib, or, p...|(65536,[4486,9235...|(65536,[4486,9
235...|  0.0|
|Mon Sep 07 19:28:...|Hey EvilRegals Tr...|-0.7142857140000001|  en|Nashv
ille, TN|      1|[hey, evilregals,...|(65536,[594,5199,...|(65536,[594,51
99,...|  0.0|
|Mon Sep 07 19:28:...|I had a fantastic...|                0.7|  en|Nashv
ille, TN|      1|[i, had, a, fanta...|(65536,[3975,5232...|(65536,[3975,5
232...|  0.0|
|Mon Sep 07 19:28:...|Cancer affects us...|                0.0|  en|Nashv
ille, TN|      1|[cancer, affects,...|(65536,[4488,9639...|(65536,[4488,9
639...|  0.0|
|Mon Sep 07 19:28:...|New research publ...|0.21666666699999998|  en|Nashv
ille, TN|      2|[new, research, p...|(65536,[8436,9863...|(65536,[8436,9
863...|  1.0|
|Mon Sep 07 19:28:...|Thank you for pro...|              -0.2|  en|  Sea
ttle, WA|      0|[thank, you, for,...|(65536,[1386,9639...|(65536,[1386,9
639...|  2.0|
|Mon Sep 07 19:28:...|Excited to see th...|                0.0|  en|Nashv
ille, TN|      1|[excited, to, see...|(65536,[666,7830,...|(65536,[666,78
30,...|  0.0|
|Mon Sep 07 19:28:...|RT Our scientist ...|                0.0|  en|Nashv
ille, TN|      1|[rt, our, scienti...|(65536,[9616,1143...|(65536,[9616,1
143...|  0.0|
+------------------+------------------+------------------+----+-----
--------+-------+--------------------+--------------------+-------------
```

```
------+-----+
only showing top 10 rows
```

In [16]:
```
#timeing for LogisticRegression in September month analysis
from pyspark.ml.classification import LogisticRegression
import time
start_time = time.time()
lr = LogisticRegression(featuresCol = 'features', labelCol = 'label', maxIt
lrModel = lr.fit(train_df)
print("--- %s seconds ---" % (time.time() - start_time))
predictions = lrModel.transform(val_df)
from pyspark.ml.evaluation import BinaryClassificationEvaluator
evaluator = BinaryClassificationEvaluator(rawPredictionCol="rawPrediction")
evaluator.evaluate(predictions)
```

▸ Spark Job Progress

```
--- 17.0746881962 seconds ---
0.5301214573613384
```

In [17]:
```
#accuracy for LogisticRegression in September month analysis
accuracy = predictions.filter(predictions.label == predictions.prediction).
accuracy
```

▸ Spark Job Progress

```
0.3500619578686493
```

In [18]:
```
#creating the October dataset for analysis
df = spark.createDataFrame(tweets2)
df.printSchema()
```

▸ Spark Job Progress

```
root
 |-- created_at: string (nullable = true)
 |-- original_text: string (nullable = true)
 |-- Polarity: double (nullable = true)
 |-- lang: string (nullable = true)
 |-- place: string (nullable = true)
 |-- classes: long (nullable = true)
```

In [19]:
```
(train_set, val_set) = df.randomSplit([0.7, 0.3], seed = 2000)
tokenizer = Tokenizer(inputCol="original_text", outputCol="tokens")
hashtf = HashingTF(numFeatures=2**16, inputCol="tokens", outputCol='tf')
idf = IDF(inputCol='tf', outputCol="features", minDocFreq=5) #minDocFreq: r
label_stringIdx = StringIndexer(inputCol = "classes", outputCol = "label")
pipeline = Pipeline(stages=[tokenizer, hashtf, idf, label_stringIdx])
pipelineFit = pipeline.fit(train_set)
train_df = pipelineFit.transform(train_set)
val_df = pipelineFit.transform(val_set)
train_df.show(10)
```

▸ Spark Job Progress

```
+------------------+------------------+----------+----+------------
+-------+------------------+--------------------+--------------------+-
----+
|          created_at|       original_text|   Polarity|lang|       place
|classes|              tokens|                  tf|            features|l
abel|
+------------------+------------------+----------+----+------------
+-------+------------------+--------------------+--------------------+-
----+
|Fri Oct 04 10:46:...|Know more what ou...|       0.0|  en|Nashville, TN
|      1|[know, more, what...|(65536,[2432,7146...|(65536,[2432,7146...|
1.0|
|Fri Oct 04 10:46:...|RT I am knocking ...|    -0.125|  en|Nashville, TN
|      0|[rt, i, am, knock...|(65536,[210,4427,...|(65536,[210,4427,...|
2.0|
|Fri Oct 04 10:46:...|RT More than 60 c...|       0.0|  en|Nashville, TN
|      1|[rt, more, than, ...|(65536,[3191,5945...|(65536,[3191,5945...|
1.0|
|Fri Oct 04 10:46:...|Agree Really exci...|0.333333333|  en|Nashville, TN
|      2|[agree, really, e...|(65536,[14,8226,8...|(65536,[14,8226,8...|
0.0|
|Fri Oct 04 10:46:...|Woman dying of br...|    -0.125|  en|Nashville, TN
|      0|[woman, dying, of...|(65536,[739,8436,...|(65536,[739,8436,...|
2.0|
|Fri Oct 04 10:46:...|RT I am knocking ...|    -0.125|  en|Nashville, TN
|      0|[rt, i, am, knock...|(65536,[210,4427,...|(65536,[210,4427,...|
2.0|
|Fri Oct 04 10:46:...|RT Fossette Went ...|    -0.125|  en|Nashville, TN
|      0|[rt, fossette, we...|(65536,[5055,8436...|(65536,[5055,8436...|
2.0|
|Fri Oct 04 10:46:...|Take Part in a Re...|    -0.125|  en|  Seattle, WA
|      0|[take, part, in, ...|(65536,[7006,8804...|(65536,[7006,8804...|
2.0|
|Fri Oct 04 10:46:...|RT I am knocking ...|      0.25|  en|Nashville, TN
|      2|[rt, i, am, knock...|(65536,[210,4427,...|(65536,[210,4427,...|
0.0|
|Fri Oct 04 10:46:...|With the help of ...|    -0.125|  en|Nashville, TN
|      0|[with, the, help,...|(65536,[666,4697,...|(65536,[666,4697,...|
2.0|
+------------------+------------------+----------+----+------------
+-------+------------------+--------------------+--------------------+-
```

```
----+
only showing top 10 rows
```

In [20]:
```
#timeing for LogisticRegression in october month analysis
start_time = time.time()
lr = LogisticRegression(featuresCol = 'features', labelCol = 'label', maxIt
lrModel = lr.fit(train_df)
print("--- %s seconds ---" % (time.time() - start_time))
predictions = lrModel.transform(val_df)
evaluator = BinaryClassificationEvaluator(rawPredictionCol="rawPrediction")
evaluator.evaluate(predictions)
```

▸ Spark Job Progress

```
--- 12.658066988 seconds ---
0.5085093108182144
```

In [21]:
```
#accuracy for LogisticRegression for october month analysis
accuracy = predictions.filter(predictions.label == predictions.prediction).
accuracy
```

▸ Spark Job Progress

```
0.35136476426799007
```

In [22]:
```
#creating the November dataset for analysis
df = spark.createDataFrame(tweets3)
df.printSchema()
```

▸ Spark Job Progress

```
root
 |-- created_at: string (nullable = true)
 |-- original_text: string (nullable = true)
 |-- polarity: double (nullable = true)
 |-- lang: string (nullable = true)
 |-- place: string (nullable = true)
 |-- classes: long (nullable = true)
```

In [23]:
```
(train_set, val_set) = df.randomSplit([0.7, 0.3], seed = 2000)
tokenizer = Tokenizer(inputCol="original_text", outputCol="tokens")
hashtf = HashingTF(numFeatures=2**16, inputCol="tokens", outputCol='tf')
idf = IDF(inputCol='tf', outputCol="features", minDocFreq=5) #minDocFreq: r
label_stringIdx = StringIndexer(inputCol = "classes", outputCol = "label")
pipeline = Pipeline(stages=[tokenizer, hashtf, idf, label_stringIdx])
pipelineFit = pipeline.fit(train_set)
train_df = pipelineFit.transform(train_set)
val_df = pipelineFit.transform(val_set)
train_df.show(10)
```

▸ Spark Job Progress

```
+--------------------+--------------------+--------+----+-------------+--
-----+--------------------+--------------------+--------------------+----
-+
|          created_at|       original_text|polarity|lang|        place|cl
asses|              tokens|                  tf|            features|labe
l|
+--------------------+--------------------+--------+----+-------------+--
-----+--------------------+--------------------+--------------------+----
-+
|Wed Nov 09 19:30:...|A new study has j...|     0.5|  en|Nashville, TN|
2|[a, new, study, h...|(65536,[12297,152...|(65536,[12297,152...|  1.0|
|Wed Nov 09 19:30:...|After breastcance...|     0.0|  en|Nashville, TN|
1|[after, breastcan...|(65536,[666,9639,...|(65536,[666,9639,...|  0.0|
|Wed Nov 09 19:30:...|RT General sessio...|     0.0|  en|Nashville, TN|
1|[rt, general, ses...|(65536,[2833,9639...|(65536,[2833,9639...|  0.0|
|Wed Nov 09 19:30:...|RT We re looking ...|     0.0|  en|Nashville, TN|
1|[rt, we, re, look...|(65536,[8315,8436...|(65536,[8315,8436...|  0.0|
|Wed Nov 09 19:30:...|General session 2...|     0.0|  en|Nashville, TN|
1|[general, session...|(65536,[2833,9639...|(65536,[2833,9639...|  0.0|
|Wed Nov 09 19:30:...|RT member effby i...|    0.05|  en|Nashville, TN|
2|[rt, member, effb...|(65536,[6585,9802...|(65536,[6585,9802...|  1.0|
|Wed Nov 09 19:30:...|We re looking for...|    -0.2|  en|Nashville, TN|
0|[we, re, looking,...|(65536,[8315,8436...|(65536,[8315,8436...|  2.0|
|Wed Nov 09 19:30:...|Advocate chat rec...|     0.0|  en|  Seattle, WA|
1|[advocate, chat, ...|(65536,[7644,9639...|(65536,[7644,9639...|  0.0|
|Wed Nov 09 19:30:...|RT Check out PS11...|     0.0|  en|Nashville, TN|
1|[rt, check, out, ...|(65536,[4889,9318...|(65536,[4889,9318...|  0.0|
|Wed Nov 09 19:30:...|RT Looking forwar...|     0.0|  en|Nashville, TN|
1|[rt, looking, for...|(65536,[2071,8315...|(65536,[2071,8315...|  0.0|
+--------------------+--------------------+--------+----+-------------+--
-----+--------------------+--------------------+--------------------+----
-+
only showing top 10 rows
```

In [24]:
```
#timeing for LogisticRegression in November month analysis
start_time = time.time()
lr = LogisticRegression(featuresCol = 'features', labelCol = 'label', maxIt
lrModel = lr.fit(train_df)
print("--- %s seconds ---" % (time.time() - start_time))
predictions = lrModel.transform(val_df)
evaluator = BinaryClassificationEvaluator(rawPredictionCol="rawPrediction")
evaluator.evaluate(predictions)
```

▸ Spark Job Progress

```
--- 12.220386982 seconds ---
0.48921489635859405
```

In [25]:
```
#accuracy for LogisticRegression for November month analysis
accuracy = predictions.filter(predictions.label == predictions.prediction).
accuracy
```

▸ Spark Job Progress

```
0.43478260869565216
```

In [26]:
```
#creating the final dataset for sentiment analysis
df_final = tweet1.append(tweets2)
df_final = df_final.append(tweets3)
df = spark.createDataFrame(df_final)
df.printSchema()
```

▸ Spark Job Progress

```
root
 |-- Polarity: double (nullable = true)
 |-- classes: long (nullable = true)
 |-- created_at: string (nullable = true)
 |-- lang: string (nullable = true)
 |-- original_text: string (nullable = true)
 |-- place: string (nullable = true)
 |-- polarity: double (nullable = true)

/tmp/1607851047030-0/lib/python2.7/site-packages/pandas/core/frame.py:669
2: FutureWarning: Sorting because non-concatenation axis is not aligned.
A future version
of pandas will change to not sort by default.

To accept the future behavior, pass 'sort=False'.

To retain the current behavior and silence the warning, pass 'sort=True'.

  sort=sort)
```

In [27]:
```
#count total number of data from all three month(tweet1, tweets2, tweets3)
df.count()
```

▸ Spark Job Progress

```
17875
```

In [28]:
```
(train_set, val_set) = df.randomSplit([0.3, 0.7], seed = 2000)
tokenizer = Tokenizer(inputCol="original_text", outputCol="tokens")
hashtf = HashingTF(numFeatures=2**16, inputCol="tokens", outputCol='tf')
idf = IDF(inputCol='tf', outputCol="features", minDocFreq=5) #minDocFreq: r
label_stringIdx = StringIndexer(inputCol = "classes", outputCol = "label")
pipeline = Pipeline(stages=[tokenizer, hashtf, idf, label_stringIdx])
pipelineFit = pipeline.fit(train_set)
train_df = pipelineFit.transform(train_set)
val_df = pipelineFit.transform(val_set)
train_df.show(10)
```

▸ Spark Job Progress

```
+--------+-------+--------------------+----+--------------------+--------
-------+------------+--------------------+--------------------+----------
----------+-----+
|Polarity|classes|          created_at|lang|       original_text|
place|    polarity|              tokens|                  tf|
features|label|
+--------+-------+--------------------+----+--------------------+--------
-------+------------+--------------------+--------------------+----------
----------+-----+
|     NaN|      0|Mon Sep 07 19:28:...|  en|Thank you for pro...|    Seat
tle, WA|        -0.2|[thank, you, for,...|(65536,[1386,9639...|(65536,[13
86,9639...|  2.0|
|     NaN|      0|Mon Sep 07 19:28:...|  en|RT Excited to mod...|  Nashvi
lle, TN|     -0.0625|[rt, excited, to,...|(65536,[3280,8436...|(65536,[32
80,8436...|  2.0|
|     NaN|      0|Mon Sep 07 19:28:...|  en|Thank you for rai...|Santa Cl
ara, CA|-0.566666667|[thank, you, for,...|(65536,[1386,6698...|(65536,[13
86,6698...|  2.0|
|     NaN|      0|Mon Sep 07 19:28:...|  en|RT As someone who...|Santa Cl
ara, CA|     -0.1625|[rt, as, someone,...|(65536,[8315,1588...|(65536,[83
15,1588...|  2.0|
|     NaN|      0|Mon Sep 07 19:28:...|  en|RT Last week on B...|Santa Cl
ara, CA|        -0.1|[rt, last, week, ...|(65536,[338,5381,...|(65536,[33
8,5381,...|  2.0|
|     NaN|      0|Mon Sep 07 19:28:...|  en|RT Why is triple ...|Santa Cl
ara, CA|     -0.1625|[rt, why, is, tri...|(65536,[6949,8461...|(65536,[69
49,8461...|  2.0|
|     NaN|      0|Mon Sep 07 19:28:...|  en|Pink ribbon for B...|    Hous
ton, TX|       -0.05|[pink, ribbon, fo...|(65536,[4488,8741...|(65536,[44
88,8741...|  2.0|
|     NaN|      0|Mon Sep 07 19:28:...|  en|Announcement FRC ...|Santa Cl
ara, CA|        -0.5|[announcement, fr...|(65536,[9514,1633...|(65536,[95
14,1633...|  2.0|
|     NaN|      0|Mon Sep 07 19:28:...|  en|Death rates from ...|    Hous
ton, TX|        -0.7|[death, rates, fr...|(65536,[4179,5595...|(65536,[41
79,5595...|  2.0|
|     NaN|      0|Mon Sep 07 19:28:...|  en|Finally going thr...|      Mi
ami, FL|     -0.1625|[finally, going, ...|(65536,[4991,6293...|(65536,[49
91,6293...|  2.0|
+--------+-------+--------------------+----+--------------------+--------
-------+------------+--------------------+--------------------+----------
```

```
----------+-----+
only showing top 10 rows
```

In [29]:
```
#timing for LogisticRegression in the final dataset for sentiment analysis
start_time = time.time()
lr = LogisticRegression(featuresCol = 'features', labelCol = 'label')
lrModel = lr.fit(train_df)
print("--- %s seconds ---" % (time.time() - start_time))
predictions = lrModel.transform(val_df)
evaluator = BinaryClassificationEvaluator(rawPredictionCol="rawPrediction")
evaluator.evaluate(predictions)
```

▸ Spark Job Progress

```
--- 12.2592139244 seconds ---
0.5032332332222572
```

In [30]:
```
#accuracy for LogisticRegression in the final dataset for sentiment analysi
accuracy = predictions.filter(predictions.label == predictions.prediction).
accuracy
```

▸ Spark Job Progress

```
0.3669739478957916
```

In [31]:
```
from pyspark.ml.classification import DecisionTreeClassifier
from pyspark.ml.classification import RandomForestClassifier
```

▸ Spark Job Progress

In [32]:
```
# timing for Decision Tree Classifier in the final dataset for sentiment an
start_time = time.time()
dt = DecisionTreeClassifier()
dtModel = dt.fit(train_df)
print("--- %s seconds ---" % (time.time() - start_time))
predictions = dtModel.transform(val_df)
evaluator = BinaryClassificationEvaluator(rawPredictionCol="rawPrediction")
evaluator.evaluate(predictions)
```

▸ Spark Job Progress

```
--- 11.210490942 seconds ---
0.5010155742090517
```

In [33]:
```
# accuracy for Decision Tree Classifier in the final dataset for sentiment
accuracy = predictions.filter(predictions.label == predictions.prediction).
accuracy
```

▸ Spark Job Progress

```
0.436312625250501
```

In [34]:
```
# timing for Random Forest Classifier in the final dataset for sentiment an
start_time = time.time()
rf = RandomForestClassifier(featuresCol = 'features', labelCol = 'label')
rfModel = rf.fit(train_df)
print("--- %s seconds ---" % (time.time() - start_time))
predictions = rfModel.transform(val_df)
evaluator = BinaryClassificationEvaluator(rawPredictionCol="rawPrediction")
evaluator.evaluate(predictions)
```

▸ Spark Job Progress

```
--- 10.5688681602 seconds ---
0.5021665670150791
```

In [35]:
```
# accuracy for Random Forest Classifier in the final dataset for sentiment
accuracy = predictions.filter(predictions.label == predictions.prediction).
accuracy
```

▸ Spark Job Progress

```
0.4377555110220441
```

In [36]:
```
# timing for Naive Bayes Classifier in the final dataset for sentiment anal
from pyspark.ml.classification import NaiveBayes
nb = NaiveBayes()
model = nb.fit(train_df)
predictions = model.transform(val_df)
evaluator = BinaryClassificationEvaluator(rawPredictionCol="rawPrediction")
evaluator.evaluate(predictions)
```

▸ Spark Job Progress

```
0.5086761627293456
```

In [37]:
```
# accuracy for Naive Bayes Classifier in the final dataset for sentiment an
accuracy = predictions.filter(predictions.label == predictions.prediction).
accuracy
```

▸ Spark Job Progress

```
0.40857715430861724
```