```
from pyspark.sql import SparkSession
from pyspark.ml.feature import IDF,StopWordsRemover,Tokenizer,HashingTF
from pyspark.sql.functions import when, col
from pyspark.ml.classification import LogisticRegression
spark = SparkSession.builder.appName("Twitter").getOrCreate()
df = spark.read.csv("/content/twitter.csv",inferSchema=True,header=True)
df = df.filter(col("tweet").isNotNull()).filter(col("label").isNotNull())
df.show()
tokenizer = Tokenizer(inputCol="tweet",outputCol="words")
filtered_words = StopWordsRemover(inputCol="words",outputCol="filtered_words")
term_freq = HashingTF(inputCol="filtered_words",outputCol="term_freq", numFeatures=5000)
idf = IDF(inputCol="term_freq",outputCol="idf")
tokenized_df = tokenizer.transform(df)
filtered = filtered_words.transform(tokenized_df)
term_freqency = term_freq.transform(filtered)
idf_model=idf.fit(term_freqency)
idf = idf_model.transform(term_freqency)
idf = idf.withColumn("label", when(col("label") == -1, 3).otherwise(col("label")))
train,test = idf.randomSplit([0.8,0.2],seed=42)
LR = LogisticRegression(featuresCol="idf",labelCol="label")
model=LR.fit(train)
predictions = model.transform(test)
predictions.show()
```