```
import os
# os.system("pip3 install torch")
# os.system("pip3 install transformers")
# os.system("pip3 install urllib3==1.26.15")
# #
import re
import torch
from pyspark.sql import SparkSession
from pyspark.sql.types import StringType
from transformers import pipeline
from dateutil.parser import parse
from pyspark.sql.functions import col, udf, to_timestamp, to_date, collect_list

sentiment_classifier = pipeline("sentiment-analysis", model="cardiffnlp/twitter-
roberta-base-sentiment")

start_date = "2017-01-01"
end_date = "2019-01-01"
spark = SparkSession.builder.appName("DSW_Final_Project").getOrCreate()
# spark.conf.set("spark.dynamicAllocation.enabled", "false")

df = spark.read.csv('s3://dsw-bittweet-bucket/tweets.csv', header=True, sep = ',')

df = df.withColumn("Date",to_timestamp(col("Date"))).na.drop(subset=['Date'])

df = df.withColumn("Date",to_date(col("Date"),"yyyy-MM-
dd")).na.drop(subset=['Date'])
df = df[(df.Date > start_date) & (df.Date < end_date)]
df = df.groupBy('date').agg(collect_list('text'))
df = df.withColumnRenamed( 'collect_list(text)', 'text_list')


# print(df.show(10))
def remove_urls(text):
# Regular expression to match URLs
      url_regex = r'http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[!*\(\),]|(?:%[0-9a-fA-
F][0-9a-fA-F]))+'
      return re.sub(url_regex, '', text)

remove_urls_udf = udf(lambda x:remove_urls(x),StringType())




# # print('Loaded sentiment sentiment_classifier')
def extract_sentiment(text):
      m = min(5, len(text))
      text = text[:m]
      text = list(map(lambda x: remove_urls(x), text))
      score = sentiment_classifier.model(**sentiment_classifier.tokenizer(text,
truncation=True, padding=True, return_tensors = 'pt')).logits.softmax(1)
[:,2].mean().item()
      return score

extract_sentiment_udf = udf(lambda x:extract_sentiment(x),StringType())
df = df.withColumn('sentiment', extract_sentiment_udf(df.text_list))
df = df.drop('text_list')
```

```
df.repartition(1).write.csv('s3://dsw-bittweet-bucket/preprocessed_tweets.csv',
header=True, sep = ',')
# #

# # print('Aggregated sentiments')
# # print(df.show(10))
# # print('Writing processed sentiments')
```