NLP.Assignment.1.Jainil.Patel.21070126039

September 9, 2023

1 1. Introduction

2 *PRN: 21070126039 Name: Jainil Patel Batch: AI/ML A2******

2.1 Git Repository : GitRepo

3 1.2. Importing the libraries

[35]:
```
# Importing the libraries
# Preprocessing the data using NLTK

# Importing the libraries/////
import nltk
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer
import pandas as pd
nltk.download('all')
```

```
[nltk_data] Downloading collection 'all'
[nltk_data]    |
[nltk_data]    | Downloading package abc to /usr/share/nltk_data…
[nltk_data]    |   Package abc is already up-to-date!
[nltk_data]    | Downloading package alpino to /usr/share/nltk_data…
[nltk_data]    |   Package alpino is already up-to-date!
[nltk_data]    | Downloading package averaged_perceptron_tagger to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package averaged_perceptron_tagger is already up-
[nltk_data]    |       to-date!
[nltk_data]    | Downloading package averaged_perceptron_tagger_ru to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package averaged_perceptron_tagger_ru is already
[nltk_data]    |       up-to-date!
[nltk_data]    | Downloading package basque_grammars to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package basque_grammars is already up-to-date!
[nltk_data]    | Downloading package bcp47 to /usr/share/nltk_data…
```

```
[nltk_data]     |   Package bcp47 is already up-to-date!
[nltk_data]     | Downloading package biocreative_ppi to
[nltk_data]     |     /usr/share/nltk_data…
[nltk_data]     |   Package biocreative_ppi is already up-to-date!
[nltk_data]     | Downloading package bllip_wsj_no_aux to
[nltk_data]     |     /usr/share/nltk_data…
[nltk_data]     |   Package bllip_wsj_no_aux is already up-to-date!
[nltk_data]     | Downloading package book_grammars to
[nltk_data]     |     /usr/share/nltk_data…
[nltk_data]     |   Package book_grammars is already up-to-date!
[nltk_data]     | Downloading package brown to /usr/share/nltk_data…
[nltk_data]     |   Package brown is already up-to-date!
[nltk_data]     | Downloading package brown_tei to
[nltk_data]     |     /usr/share/nltk_data…
[nltk_data]     |   Package brown_tei is already up-to-date!
[nltk_data]     | Downloading package cess_cat to
[nltk_data]     |     /usr/share/nltk_data…
[nltk_data]     |   Package cess_cat is already up-to-date!
[nltk_data]     | Downloading package cess_esp to
[nltk_data]     |     /usr/share/nltk_data…
[nltk_data]     |   Package cess_esp is already up-to-date!
[nltk_data]     | Downloading package chat80 to /usr/share/nltk_data…
[nltk_data]     |   Package chat80 is already up-to-date!
[nltk_data]     | Downloading package city_database to
[nltk_data]     |     /usr/share/nltk_data…
[nltk_data]     |   Package city_database is already up-to-date!
[nltk_data]     | Downloading package cmudict to
[nltk_data]     |     /usr/share/nltk_data…
[nltk_data]     |   Package cmudict is already up-to-date!
[nltk_data]     | Downloading package comparative_sentences to
[nltk_data]     |     /usr/share/nltk_data…
[nltk_data]     |   Package comparative_sentences is already up-to-
[nltk_data]     |       date!
[nltk_data]     | Downloading package comtrans to
[nltk_data]     |     /usr/share/nltk_data…
[nltk_data]     |   Package comtrans is already up-to-date!
[nltk_data]     | Downloading package conll2000 to
[nltk_data]     |     /usr/share/nltk_data…
[nltk_data]     |   Package conll2000 is already up-to-date!
[nltk_data]     | Downloading package conll2002 to
[nltk_data]     |     /usr/share/nltk_data…
[nltk_data]     |   Package conll2002 is already up-to-date!
[nltk_data]     | Downloading package conll2007 to
[nltk_data]     |     /usr/share/nltk_data…
[nltk_data]     |   Package conll2007 is already up-to-date!
[nltk_data]     | Downloading package crubadan to
[nltk_data]     |     /usr/share/nltk_data…
[nltk_data]     |   Package crubadan is already up-to-date!
```

```
[nltk_data]    | Downloading package dependency_treebank to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package dependency_treebank is already up-to-date!
[nltk_data]    | Downloading package dolch to /usr/share/nltk_data…
[nltk_data]    |   Package dolch is already up-to-date!
[nltk_data]    | Downloading package europarl_raw to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package europarl_raw is already up-to-date!
[nltk_data]    | Downloading package extended_omw to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package extended_omw is already up-to-date!
[nltk_data]    | Downloading package floresta to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package floresta is already up-to-date!
[nltk_data]    | Downloading package framenet_v15 to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package framenet_v15 is already up-to-date!
[nltk_data]    | Downloading package framenet_v17 to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package framenet_v17 is already up-to-date!
[nltk_data]    | Downloading package gazetteers to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package gazetteers is already up-to-date!
[nltk_data]    | Downloading package genesis to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package genesis is already up-to-date!
[nltk_data]    | Downloading package gutenberg to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package gutenberg is already up-to-date!
[nltk_data]    | Downloading package ieer to /usr/share/nltk_data…
[nltk_data]    |   Package ieer is already up-to-date!
[nltk_data]    | Downloading package inaugural to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package inaugural is already up-to-date!
[nltk_data]    | Downloading package indian to /usr/share/nltk_data…
[nltk_data]    |   Package indian is already up-to-date!
[nltk_data]    | Downloading package jeita to /usr/share/nltk_data…
[nltk_data]    |   Package jeita is already up-to-date!
[nltk_data]    | Downloading package kimmo to /usr/share/nltk_data…
[nltk_data]    |   Package kimmo is already up-to-date!
[nltk_data]    | Downloading package knbc to /usr/share/nltk_data…
[nltk_data]    |   Package knbc is already up-to-date!
[nltk_data]    | Downloading package large_grammars to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package large_grammars is already up-to-date!
[nltk_data]    | Downloading package lin_thesaurus to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package lin_thesaurus is already up-to-date!
```

```
[nltk_data]    | Downloading package mac_morpho to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package mac_morpho is already up-to-date!
[nltk_data]    | Downloading package machado to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package machado is already up-to-date!
[nltk_data]    | Downloading package masc_tagged to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package masc_tagged is already up-to-date!
[nltk_data]    | Downloading package maxent_ne_chunker to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package maxent_ne_chunker is already up-to-date!
[nltk_data]    | Downloading package maxent_treebank_pos_tagger to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package maxent_treebank_pos_tagger is already up-
[nltk_data]    |       to-date!
[nltk_data]    | Downloading package moses_sample to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package moses_sample is already up-to-date!
[nltk_data]    | Downloading package movie_reviews to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package movie_reviews is already up-to-date!
[nltk_data]    | Downloading package mte_teip5 to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package mte_teip5 is already up-to-date!
[nltk_data]    | Downloading package mwa_ppdb to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package mwa_ppdb is already up-to-date!
[nltk_data]    | Downloading package names to /usr/share/nltk_data…
[nltk_data]    |   Package names is already up-to-date!
[nltk_data]    | Downloading package nombank.1.0 to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package nombank.1.0 is already up-to-date!
[nltk_data]    | Downloading package nonbreaking_prefixes to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package nonbreaking_prefixes is already up-to-date!
[nltk_data]    | Downloading package nps_chat to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package nps_chat is already up-to-date!
[nltk_data]    | Downloading package omw to /usr/share/nltk_data…
[nltk_data]    |   Package omw is already up-to-date!
[nltk_data]    | Downloading package omw-1.4 to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package omw-1.4 is already up-to-date!
[nltk_data]    | Downloading package opinion_lexicon to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package opinion_lexicon is already up-to-date!
[nltk_data]    | Downloading package panlex_swadesh to
```

```
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package panlex_swadesh is already up-to-date!
[nltk_data]    | Downloading package paradigms to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package paradigms is already up-to-date!
[nltk_data]    | Downloading package pe08 to /usr/share/nltk_data…
[nltk_data]    |   Package pe08 is already up-to-date!
[nltk_data]    | Downloading package perluniprops to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package perluniprops is already up-to-date!
[nltk_data]    | Downloading package pil to /usr/share/nltk_data…
[nltk_data]    |   Package pil is already up-to-date!
[nltk_data]    | Downloading package pl196x to /usr/share/nltk_data…
[nltk_data]    |   Package pl196x is already up-to-date!
[nltk_data]    | Downloading package porter_test to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package porter_test is already up-to-date!
[nltk_data]    | Downloading package ppattach to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package ppattach is already up-to-date!
[nltk_data]    | Downloading package problem_reports to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package problem_reports is already up-to-date!
[nltk_data]    | Downloading package product_reviews_1 to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package product_reviews_1 is already up-to-date!
[nltk_data]    | Downloading package product_reviews_2 to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package product_reviews_2 is already up-to-date!
[nltk_data]    | Downloading package propbank to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package propbank is already up-to-date!
[nltk_data]    | Downloading package pros_cons to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package pros_cons is already up-to-date!
[nltk_data]    | Downloading package ptb to /usr/share/nltk_data…
[nltk_data]    |   Package ptb is already up-to-date!
[nltk_data]    | Downloading package punkt to /usr/share/nltk_data…
[nltk_data]    |   Package punkt is already up-to-date!
[nltk_data]    | Downloading package qc to /usr/share/nltk_data…
[nltk_data]    |   Package qc is already up-to-date!
[nltk_data]    | Downloading package reuters to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package reuters is already up-to-date!
[nltk_data]    | Downloading package rslp to /usr/share/nltk_data…
[nltk_data]    |   Package rslp is already up-to-date!
[nltk_data]    | Downloading package rte to /usr/share/nltk_data…
[nltk_data]    |   Package rte is already up-to-date!
```

```
[nltk_data]    | Downloading package sample_grammars to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package sample_grammars is already up-to-date!
[nltk_data]    | Downloading package semcor to /usr/share/nltk_data…
[nltk_data]    |   Package semcor is already up-to-date!
[nltk_data]    | Downloading package senseval to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package senseval is already up-to-date!
[nltk_data]    | Downloading package sentence_polarity to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package sentence_polarity is already up-to-date!
[nltk_data]    | Downloading package sentiwordnet to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package sentiwordnet is already up-to-date!
[nltk_data]    | Downloading package shakespeare to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package shakespeare is already up-to-date!
[nltk_data]    | Downloading package sinica_treebank to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package sinica_treebank is already up-to-date!
[nltk_data]    | Downloading package smultron to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package smultron is already up-to-date!
[nltk_data]    | Downloading package snowball_data to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package snowball_data is already up-to-date!
[nltk_data]    | Downloading package spanish_grammars to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package spanish_grammars is already up-to-date!
[nltk_data]    | Downloading package state_union to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package state_union is already up-to-date!
[nltk_data]    | Downloading package stopwords to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package stopwords is already up-to-date!
[nltk_data]    | Downloading package subjectivity to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package subjectivity is already up-to-date!
[nltk_data]    | Downloading package swadesh to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package swadesh is already up-to-date!
[nltk_data]    | Downloading package switchboard to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package switchboard is already up-to-date!
[nltk_data]    | Downloading package tagsets to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package tagsets is already up-to-date!
[nltk_data]    | Downloading package timit to /usr/share/nltk_data…
```

```
[nltk_data]    |   Package timit is already up-to-date!
[nltk_data]    | Downloading package toolbox to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package toolbox is already up-to-date!
[nltk_data]    | Downloading package treebank to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package treebank is already up-to-date!
[nltk_data]    | Downloading package twitter_samples to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package twitter_samples is already up-to-date!
[nltk_data]    | Downloading package udhr to /usr/share/nltk_data…
[nltk_data]    |   Package udhr is already up-to-date!
[nltk_data]    | Downloading package udhr2 to /usr/share/nltk_data…
[nltk_data]    |   Package udhr2 is already up-to-date!
[nltk_data]    | Downloading package unicode_samples to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package unicode_samples is already up-to-date!
[nltk_data]    | Downloading package universal_tagset to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package universal_tagset is already up-to-date!
[nltk_data]    | Downloading package universal_treebanks_v20 to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package universal_treebanks_v20 is already up-to-
[nltk_data]    |       date!
[nltk_data]    | Downloading package vader_lexicon to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package vader_lexicon is already up-to-date!
[nltk_data]    | Downloading package verbnet to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package verbnet is already up-to-date!
[nltk_data]    | Downloading package verbnet3 to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package verbnet3 is already up-to-date!
[nltk_data]    | Downloading package webtext to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package webtext is already up-to-date!
[nltk_data]    | Downloading package wmt15_eval to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package wmt15_eval is already up-to-date!
[nltk_data]    | Downloading package word2vec_sample to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package word2vec_sample is already up-to-date!
[nltk_data]    | Downloading package wordnet to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package wordnet is already up-to-date!
[nltk_data]    | Downloading package wordnet2021 to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package wordnet2021 is already up-to-date!
```

```
[nltk_data]    | Downloading package wordnet2022 to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package wordnet2022 is already up-to-date!
[nltk_data]    | Downloading package wordnet31 to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package wordnet31 is already up-to-date!
[nltk_data]    | Downloading package wordnet_ic to
[nltk_data]    |     /usr/share/nltk_data…
[nltk_data]    |   Package wordnet_ic is already up-to-date!
[nltk_data]    | Downloading package words to /usr/share/nltk_data…
[nltk_data]    |   Package words is already up-to-date!
[nltk_data]    | Downloading package ycoe to /usr/share/nltk_data…
[nltk_data]    |   Package ycoe is already up-to-date!
[nltk_data]    |
[nltk_data]  Done downloading collection all
```

[35]:
```
True
```

4 1.3. Importing the dataset

[37]:
```
import re
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Embedding, LSTM, Dense
from sklearn.metrics import classification_report
```

[50]:
```
df = pd.read_csv(r"/kaggle/input/multilingual-lyrics-for-genre-classification/
 ↪train.csv")
df.head()# reading and showing dataset
```

[50]:

	Artist	Song	Genre	Language	Lyrics
0	12 stones	world so cold	Rock	en	It starts with pain, followed by hate\nFueled …
1	12 stones	broken	Rock	en	Freedom!\nAlone again again alone\nPatiently w…
2	12 stones	3 leaf loser	Rock	en	Biting the hand that feeds you, lying to the v…
3	12 stones	anthem for the underdog	Rock	en	You say you know just who I am\nBut you can't …
4	12 stones	adrenaline	Rock	en	My heart is beating faster can't control these…

5 2. Data Preprocessing

[51]:
```
df.dropna(inplace=True) # preprocessing
```

[52]:
```
df['Genre'].value_counts()
```

[52]:
```
Genre
Rock          121390
Pop           108693
Metal          20286
Jazz           13545
Folk            8644
Indie           8449
R&B             2793
Hip-Hop         2240
Electronic      2213
Country         1890
Name: count, dtype: int64
```

[53]:
```
df = df[df['Genre'].isin(['Rock', 'Jazz', 'Hip-Hop', 'Metal', 'Country'])]
df = df.dropna()  # Remove empty rows
df = df.drop_duplicates()  # Remove duplicates
```

[54]:
```
df.shape
```

[54]:
```
(147771, 5)
```

6 2.1. Lemmatization and Tokenization

[42]:
```
import nltk
nltk.download('stopwords')
nltk.download('wordnet')
!unzip /usr/share/nltk_data/corpora/wordnet.zip -d /usr/share/nltk_data/corpora/
```

```
[nltk_data] Downloading package stopwords to /usr/share/nltk_data…
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to /usr/share/nltk_data…
[nltk_data]   Package wordnet is already up-to-date!
Archive:  /usr/share/nltk_data/corpora/wordnet.zip
replace /usr/share/nltk_data/corpora/wordnet/lexnames? [y]es, [n]o, [A]ll,
[N]one, [r]ename: ^C
```

[55]:
```
lemmatizer = WordNetLemmatizer()

def preprocess_text(text):
    # Tokenization and Lemmatization
```

```
    tokens = word_tokenize(text)
    tokens = [lemmatizer.lemmatize(token) for token in tokens]

    # Data Cleansing
    text = ' '.join(tokens)
    text = re.sub(r'http\S+', '', text)  # Remove URLs
    text = re.sub(r'[^a-zA-Z\s]', '', text)  # Remove symbols
    text = re.sub(r' +', ' ', text)  # Remove excess whitespaces
    text = text.lower()  # Lowercase text

    return text

df['Lyrics'] = df['Lyrics'].apply(preprocess_text)
```

[56]:
```
df["Lyrics"].head()
```

[56]:
```
0    it start with pain followed by hate fueled by …
1    freedom alone again again alone patiently wait…
2    biting the hand that feed you lying to the voi…
3    you say you know just who i am but you ca nt i…
4    my heart is beating faster ca nt control these…
Name: Lyrics, dtype: object
```

7 3. Data Cleaning

7.1 3.1 Remove stopwords, Remove symbols, Remove URLs

[57]:
```
# Data Cleansing: Remove stopwords, remove symbols, remove URLs
# Importing the libraries
import re
from nltk.corpus import stopwords

stop_words = set(stopwords.words('english'))
```

[58]:
```
# Defining a function to clean the text
def clean_Text(text):
    # Remove URLs
    text = re.sub(r'http\S+', '', text)
    # Remove symbols and numbers
    text = re.sub(r'[^\w\s]', '', text)
    # Remove stopwords
    text = " ".join([word for word in text.split() if word.lower() not in␣
 ↪stop_words])

    # Remove excess whitespaces
    text = ' '.join(text.split())
```

```
    # Replace abbreviations (you can add more if needed)
    text = re.sub(r"won't", "will not", text)
    text = re.sub(r"can't", "cannot", text)

    # Fix contractions
    text = re.sub(r"n't", " not", text)
    text = re.sub(r"'re", " are", text)
    text = re.sub(r"'s", " is", text)
    text = re.sub(r"'d", " would", text)
    text = re.sub(r"'ll", " will", text)
    text = re.sub(r"'t", " not", text)
    text = re.sub(r"'ve", " have", text)
    return text
```

[59]:
```
df.rename(columns={'Lyrics': 'lyrics'}, inplace=True)
df.rename(columns={'Genre': 'genre'}, inplace=True)
```

[60]:
```
# Applying the clean Text function to the Text column
df['lyrics'] = df['lyrics'].apply(clean_Text)

# Displaying the first 5 rows of the dataset
df.head()
```

[60]:

	Artist	Song	genre	Language	lyrics
0	12 stones	world so cold	Rock	en	start pain followed hate fueled endless questi…
1	12 stones	broken	Rock	en	freedom alone alone patiently waiting phone ho…
2	12 stones	3 leaf loser	Rock	en	biting hand feed lying voice inside reach beg …
3	12 stones	anthem for the underdog	Rock	en	say know ca nt imagine wait across line though…
4	12 stones	adrenaline	Rock	en	heart beating faster ca nt control feeling any…

8 4. Split the dataset into train and test sets

[63]:
```
X_train, X_test, y_train, y_test = train_test_split(df['lyrics'], df['genre'],␣
↪test_size=0.2, random_state=42)
```

9 5. Model training

[64]:
```
# Set 1 Parameters
batch_size_1 = 4
max_sequence_length_1 = 50
embedding_dim_1 = 50
max_words_1 = 10000
lstm_units_1 = 32

tokenizer = Tokenizer(num_words=max_words_1)
tokenizer.fit_on_texts(X_train)
X_train_seq = tokenizer.texts_to_sequences(X_train)
X_test_seq = tokenizer.texts_to_sequences(X_test)

X_train_pad = pad_sequences(X_train_seq, maxlen=max_sequence_length_1)
X_test_pad = pad_sequences(X_test_seq, maxlen=max_sequence_length_1)

# Encode genre labels
label_encoder = LabelEncoder()
y_train_encoded = label_encoder.fit_transform(y_train)
y_test_encoded = label_encoder.transform(y_test)
```

[65]:
```
# Build and train the LSTM model for Set 1
model_1 = Sequential()
model_1.add(Embedding(input_dim=max_words_1, output_dim=embedding_dim_1,␣
 ↪input_length=max_sequence_length_1))
model_1.add(LSTM(lstm_units_1))
model_1.add(Dense(5, activation='softmax'))  # 5 genres

model_1.compile(loss='sparse_categorical_crossentropy', optimizer='adam',␣
 ↪metrics=['accuracy'])
model_1.fit(X_train_pad, y_train_encoded, epochs=30, batch_size=batch_size_1,␣
 ↪validation_split=0.2)

# Evaluate the model for Set 1
y_pred_1 = model_1.predict(X_test_pad)
y_pred_classes_1 = [label_encoder.classes_[i] for i in y_pred_1.argmax(axis=-1)]
```

```
Epoch 1/30
23643/23643 [==============================] - 165s 7ms/step - loss: 0.6223 -
accuracy: 0.7924 - val_loss: 0.5554 - val_accuracy: 0.8140
Epoch 2/30
23643/23643 [==============================] - 141s 6ms/step - loss: 0.5164 -
accuracy: 0.8277 - val_loss: 0.5468 - val_accuracy: 0.8182
Epoch 3/30
23643/23643 [==============================] - 146s 6ms/step - loss: 0.4650 -
accuracy: 0.8463 - val_loss: 0.5541 - val_accuracy: 0.8181
```

```
Epoch 4/30
23643/23643 [==============================] - 140s 6ms/step - loss: 0.4185 -
accuracy: 0.8611 - val_loss: 0.5876 - val_accuracy: 0.8175
Epoch 5/30
23643/23643 [==============================] - 140s 6ms/step - loss: 0.3747 -
accuracy: 0.8758 - val_loss: 0.5985 - val_accuracy: 0.8147
Epoch 6/30
23643/23643 [==============================] - 147s 6ms/step - loss: 0.3340 -
accuracy: 0.8898 - val_loss: 0.6279 - val_accuracy: 0.8105
Epoch 7/30
23643/23643 [==============================] - 140s 6ms/step - loss: 0.2952 -
accuracy: 0.9034 - val_loss: 0.6584 - val_accuracy: 0.8082
Epoch 8/30
23643/23643 [==============================] - 147s 6ms/step - loss: 0.2624 -
accuracy: 0.9134 - val_loss: 0.7111 - val_accuracy: 0.8004
Epoch 9/30
23643/23643 [==============================] - 146s 6ms/step - loss: 0.2344 -
accuracy: 0.9239 - val_loss: 0.7636 - val_accuracy: 0.8038
Epoch 10/30
23643/23643 [==============================] - 141s 6ms/step - loss: 0.2102 -
accuracy: 0.9310 - val_loss: 0.8082 - val_accuracy: 0.7874
Epoch 11/30
23643/23643 [==============================] - 146s 6ms/step - loss: 0.1900 -
accuracy: 0.9378 - val_loss: 0.8482 - val_accuracy: 0.7974
Epoch 12/30
23643/23643 [==============================] - 141s 6ms/step - loss: 0.1745 -
accuracy: 0.9422 - val_loss: 0.9212 - val_accuracy: 0.7866
Epoch 13/30
23643/23643 [==============================] - 140s 6ms/step - loss: 0.1616 -
accuracy: 0.9467 - val_loss: 0.9678 - val_accuracy: 0.7883
Epoch 14/30
23643/23643 [==============================] - 146s 6ms/step - loss: 0.1501 -
accuracy: 0.9500 - val_loss: 0.9751 - val_accuracy: 0.7809
Epoch 15/30
23643/23643 [==============================] - 140s 6ms/step - loss: 0.1409 -
accuracy: 0.9538 - val_loss: 1.0404 - val_accuracy: 0.7785
Epoch 16/30
23643/23643 [==============================] - 146s 6ms/step - loss: 0.1320 -
accuracy: 0.9555 - val_loss: 1.0564 - val_accuracy: 0.7853
Epoch 17/30
23643/23643 [==============================] - 148s 6ms/step - loss: 0.1262 -
accuracy: 0.9579 - val_loss: 1.1136 - val_accuracy: 0.7708
Epoch 18/30
23643/23643 [==============================] - 147s 6ms/step - loss: 0.1216 -
accuracy: 0.9591 - val_loss: 1.1310 - val_accuracy: 0.7849
Epoch 19/30
23643/23643 [==============================] - 140s 6ms/step - loss: 0.1178 -
accuracy: 0.9611 - val_loss: 1.0978 - val_accuracy: 0.7762
```

```
Epoch 20/30
23643/23643 [==============================] - 146s 6ms/step - loss: 0.1108 -
accuracy: 0.9633 - val_loss: 1.1766 - val_accuracy: 0.7793
Epoch 21/30
23643/23643 [==============================] - 141s 6ms/step - loss: 0.1102 -
accuracy: 0.9628 - val_loss: 1.1689 - val_accuracy: 0.7747
Epoch 22/30
23643/23643 [==============================] - 144s 6ms/step - loss: 0.1072 -
accuracy: 0.9643 - val_loss: 1.1605 - val_accuracy: 0.7818
Epoch 23/30
23643/23643 [==============================] - 141s 6ms/step - loss: 0.1042 -
accuracy: 0.9652 - val_loss: 1.1842 - val_accuracy: 0.7767
Epoch 24/30
23643/23643 [==============================] - 142s 6ms/step - loss: 0.0999 -
accuracy: 0.9664 - val_loss: 1.2065 - val_accuracy: 0.7798
Epoch 25/30
23643/23643 [==============================] - 148s 6ms/step - loss: 0.0999 -
accuracy: 0.9672 - val_loss: 1.1780 - val_accuracy: 0.7858
Epoch 26/30
23643/23643 [==============================] - 146s 6ms/step - loss: 0.0952 -
accuracy: 0.9681 - val_loss: 1.2213 - val_accuracy: 0.7826
Epoch 27/30
23643/23643 [==============================] - 146s 6ms/step - loss: 0.0940 -
accuracy: 0.9687 - val_loss: 1.2718 - val_accuracy: 0.7669
Epoch 28/30
23643/23643 [==============================] - 140s 6ms/step - loss: 0.0919 -
accuracy: 0.9693 - val_loss: 1.2749 - val_accuracy: 0.7785
Epoch 29/30
23643/23643 [==============================] - 146s 6ms/step - loss: 0.0918 -
accuracy: 0.9698 - val_loss: 1.2867 - val_accuracy: 0.7746
Epoch 30/30
23643/23643 [==============================] - 147s 6ms/step - loss: 0.0925 -
accuracy: 0.9695 - val_loss: 1.2603 - val_accuracy: 0.7784
924/924 [==============================] - 3s 2ms/step
```

9.1 Classification Result

[66]:
```
print("Results for Set 1:")
print(classification_report(y_test, y_pred_classes_1))
```

```
Results for Set 1:
              precision    recall  f1-score   support

     Country       0.04      0.03      0.03       375
     Hip-Hop       0.40      0.37      0.38       434
        Jazz       0.66      0.61      0.63      2718
       Metal       0.53      0.51      0.52      3838
        Rock       0.85      0.87      0.86     22190
```

```
    accuracy                           0.78     29555
   macro avg       0.50      0.48      0.49     29555
weighted avg       0.78      0.78      0.78     29555
```

[67]:
```
# Set 2 Parameters
batch_size_2 = 8
max_sequence_length_2 = 30
embedding_dim_2 = 30
max_words_2 = 25000
lstm_units_2 = 32

tokenizer = Tokenizer(num_words=max_words_2)
tokenizer.fit_on_texts(X_train)
X_train_seq = tokenizer.texts_to_sequences(X_train)
X_test_seq = tokenizer.texts_to_sequences(X_test)

X_train_pad = pad_sequences(X_train_seq, maxlen=max_sequence_length_2)
X_test_pad = pad_sequences(X_test_seq, maxlen=max_sequence_length_2)

# Encode genre labels
label_encoder = LabelEncoder()
y_train_encoded = label_encoder.fit_transform(y_train)
y_test_encoded = label_encoder.transform(y_test)
```

[68]:
```
# Build and train the LSTM model for Set 2
model_2 = Sequential()
model_2.add(Embedding(input_dim=max_words_2, output_dim=embedding_dim_2,␣
 ↪input_length=max_sequence_length_2))
model_2.add(LSTM(lstm_units_2, return_sequences=True))  # Two layers of LSTM
model_2.add(LSTM(lstm_units_2))
model_2.add(Dense(5, activation='softmax'))  # 5 genres

model_2.compile(loss='sparse_categorical_crossentropy', optimizer='adam',␣
 ↪metrics=['accuracy'])
model_2.fit(X_train_pad, y_train_encoded, epochs=25, batch_size=batch_size_2,␣
 ↪validation_split=0.2)

# Evaluate the model for Set 2
y_pred_2 = model_2.predict(X_test_pad)
y_pred_classes_2 = [label_encoder.classes_[i] for i in y_pred_2.argmax(axis=-1)]
```

```
Epoch 1/25
11822/11822 [==============================] - 121s 10ms/step - loss: 0.6605 -
accuracy: 0.7817 - val_loss: 0.5933 - val_accuracy: 0.8030
Epoch 2/25
```

```
11822/11822 [==============================] - 97s 8ms/step - loss: 0.5480 -
accuracy: 0.8194 - val_loss: 0.5728 - val_accuracy: 0.8101
Epoch 3/25
11822/11822 [==============================] - 97s 8ms/step - loss: 0.4909 -
accuracy: 0.8394 - val_loss: 0.5807 - val_accuracy: 0.8142
Epoch 4/25
11822/11822 [==============================] - 96s 8ms/step - loss: 0.4448 -
accuracy: 0.8554 - val_loss: 0.5929 - val_accuracy: 0.8115
Epoch 5/25
11822/11822 [==============================] - 98s 8ms/step - loss: 0.4010 -
accuracy: 0.8702 - val_loss: 0.6360 - val_accuracy: 0.8048
Epoch 6/25
11822/11822 [==============================] - 97s 8ms/step - loss: 0.3590 -
accuracy: 0.8840 - val_loss: 0.6555 - val_accuracy: 0.8072
Epoch 7/25
11822/11822 [==============================] - 97s 8ms/step - loss: 0.3212 -
accuracy: 0.8966 - val_loss: 0.7048 - val_accuracy: 0.7993
Epoch 8/25
11822/11822 [==============================] - 97s 8ms/step - loss: 0.2870 -
accuracy: 0.9071 - val_loss: 0.7490 - val_accuracy: 0.7950
Epoch 9/25
11822/11822 [==============================] - 97s 8ms/step - loss: 0.2572 -
accuracy: 0.9171 - val_loss: 0.8193 - val_accuracy: 0.7817
Epoch 10/25
11822/11822 [==============================] - 97s 8ms/step - loss: 0.2296 -
accuracy: 0.9261 - val_loss: 0.8726 - val_accuracy: 0.7729
Epoch 11/25
11822/11822 [==============================] - 97s 8ms/step - loss: 0.2078 -
accuracy: 0.9318 - val_loss: 0.9253 - val_accuracy: 0.7812
Epoch 12/25
11822/11822 [==============================] - 96s 8ms/step - loss: 0.1880 -
accuracy: 0.9385 - val_loss: 0.9486 - val_accuracy: 0.7851
Epoch 13/25
11822/11822 [==============================] - 96s 8ms/step - loss: 0.1728 -
accuracy: 0.9434 - val_loss: 1.0166 - val_accuracy: 0.7643
Epoch 14/25
11822/11822 [==============================] - 96s 8ms/step - loss: 0.1588 -
accuracy: 0.9479 - val_loss: 1.0805 - val_accuracy: 0.7588
Epoch 15/25
11822/11822 [==============================] - 96s 8ms/step - loss: 0.1469 -
accuracy: 0.9520 - val_loss: 1.0864 - val_accuracy: 0.7572
Epoch 16/25
11822/11822 [==============================] - 98s 8ms/step - loss: 0.1353 -
accuracy: 0.9550 - val_loss: 1.1761 - val_accuracy: 0.7658
Epoch 17/25
11822/11822 [==============================] - 98s 8ms/step - loss: 0.1272 -
accuracy: 0.9574 - val_loss: 1.1812 - val_accuracy: 0.7640
Epoch 18/25
```

```
11822/11822 [==============================] - 99s 8ms/step - loss: 0.1195 -
accuracy: 0.9601 - val_loss: 1.1693 - val_accuracy: 0.7562
Epoch 19/25
11822/11822 [==============================] - 100s 8ms/step - loss: 0.1111 -
accuracy: 0.9620 - val_loss: 1.2564 - val_accuracy: 0.7551
Epoch 20/25
11822/11822 [==============================] - 100s 8ms/step - loss: 0.1065 -
accuracy: 0.9638 - val_loss: 1.2924 - val_accuracy: 0.7634
Epoch 21/25
11822/11822 [==============================] - 99s 8ms/step - loss: 0.1024 -
accuracy: 0.9651 - val_loss: 1.2585 - val_accuracy: 0.7641
Epoch 22/25
11822/11822 [==============================] - 99s 8ms/step - loss: 0.0970 -
accuracy: 0.9673 - val_loss: 1.3178 - val_accuracy: 0.7566
Epoch 23/25
11822/11822 [==============================] - 99s 8ms/step - loss: 0.0932 -
accuracy: 0.9683 - val_loss: 1.3247 - val_accuracy: 0.7556
Epoch 24/25
11822/11822 [==============================] - 99s 8ms/step - loss: 0.0899 -
accuracy: 0.9690 - val_loss: 1.3403 - val_accuracy: 0.7588
Epoch 25/25
11822/11822 [==============================] - 98s 8ms/step - loss: 0.0872 -
accuracy: 0.9700 - val_loss: 1.3393 - val_accuracy: 0.7567
924/924 [==============================] - 4s 3ms/step
```

9.2 Classification Result

[69]:
```
print("\nResults for Set 2:")
print(classification_report(y_test, y_pred_classes_2))
```

```
Results for Set 2:
              precision    recall  f1-score   support

     Country       0.04      0.05      0.05       375
     Hip-Hop       0.35      0.27      0.31       434
        Jazz       0.56      0.62      0.59      2718
       Metal       0.49      0.47      0.48      3838
        Rock       0.85      0.84      0.85     22190

    accuracy                           0.76     29555
   macro avg       0.46      0.45      0.45     29555
weighted avg       0.76      0.76      0.76     29555
```

[70]:
```
import matplotlib.pyplot as plt
from sklearn.metrics import confusion_matrix
import numpy as np
```

```
# Function to plot the multiclass confusion matrix
def plot_multiclass_confusion_matrix(y_true, y_pred, classes, title):
    cm = confusion_matrix(y_true, y_pred)
    plt.figure(figsize=(8, 6))
    plt.imshow(cm, interpolation='nearest', cmap=plt.cm.Blues)
    plt.title(title)
    plt.colorbar()
    tick_marks = np.arange(len(classes))
    plt.xticks(tick_marks, classes, rotation=45)
    plt.yticks(tick_marks, classes)
    plt.xlabel('Predicted label')
    plt.ylabel('True label')

    for i in range(len(classes)):
        for j in range(len(classes)):
            plt.text(j, i, format(cm[i, j], 'd'), horizontalalignment="center",
                     color="white" if cm[i, j] > cm.max() / 2 else "black")

    plt.show()
```

9.3 Confusion Matrix

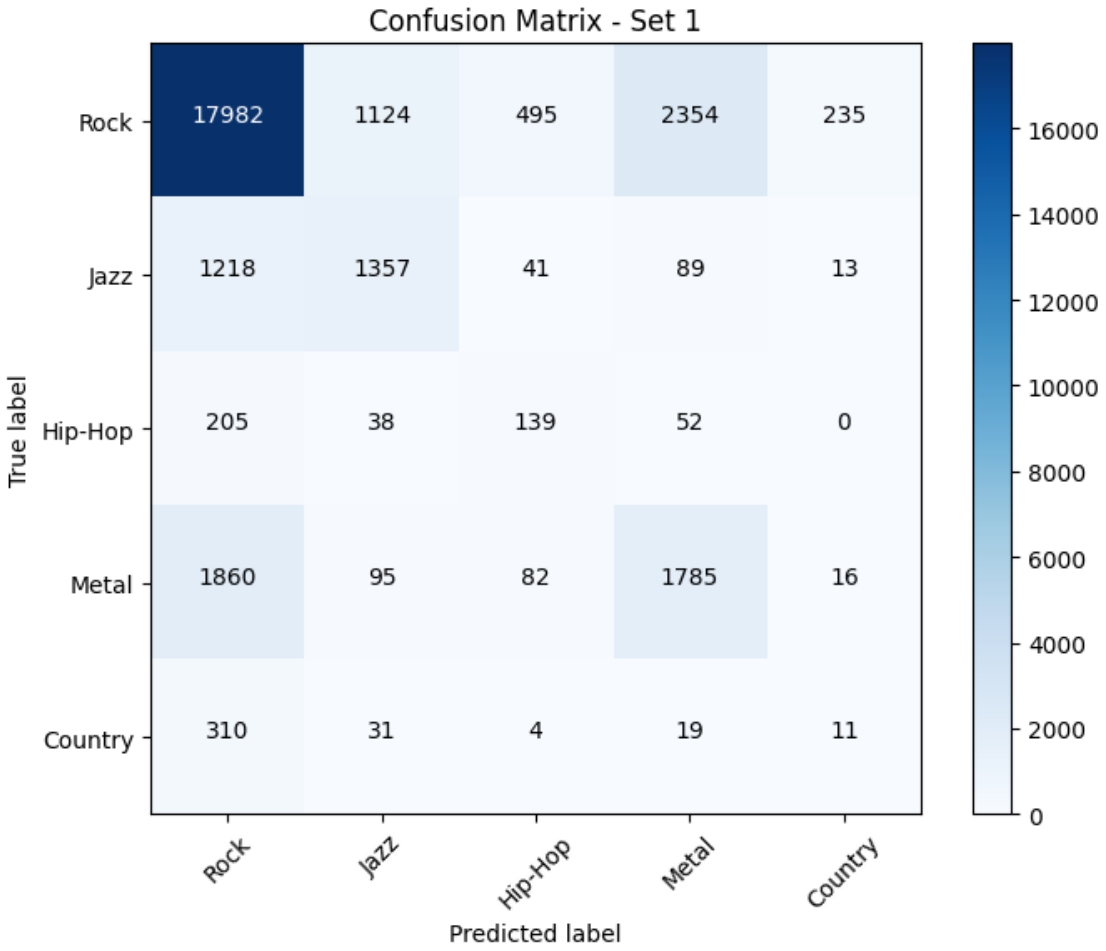
[71]:
```
# Define genre classes
genre_classes = ['Rock', 'Jazz', 'Hip-Hop', 'Metal', 'Country']

# Evaluate the model for Set 1
y_pred_1 = model_1.predict(X_test_pad)
y_pred_classes_1 = [np.argmax(pred) for pred in y_pred_1]

# Map the encoded genre labels back to genre names
y_true_genre = [genre_classes[label] for label in y_test_encoded]
y_pred_genre = [genre_classes[label] for label in y_pred_classes_1]

# Plot the multiclass confusion matrix for Set 1
plot_multiclass_confusion_matrix(y_true_genre, y_pred_genre, genre_classes,␣
 ↪"Confusion Matrix - Set 1")
```

```
924/924 [==============================] - 3s 2ms/step
```



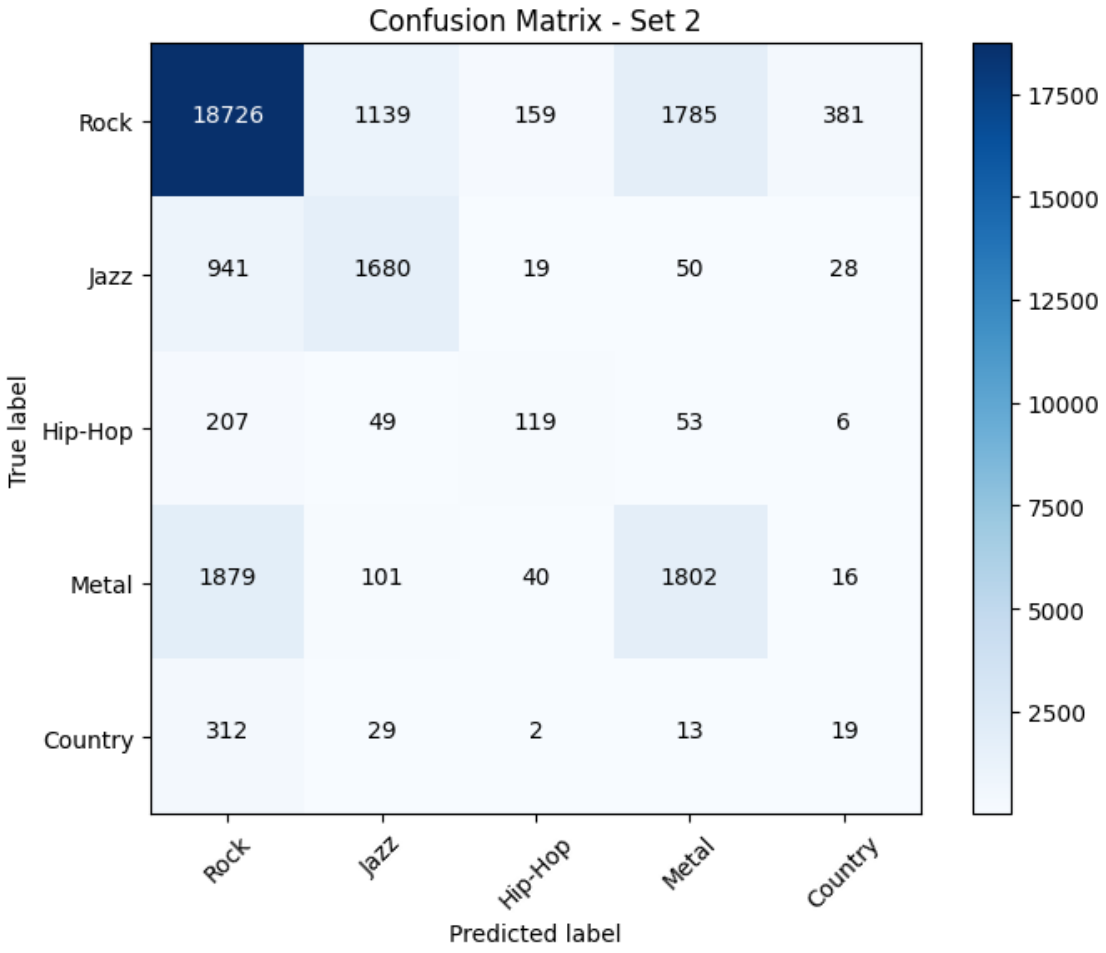
[74]:
```
# Define genre classes
genre_classes = ['Rock', 'Jazz', 'Hip-Hop', 'Metal', 'Country']

# Evaluate the model for Set 1
y_pred_2 = model_2.predict(X_test_pad)
y_pred_classes_2 = [np.argmax(pred) for pred in y_pred_2]

# Map the encoded genre labels back to genre names
y_true_genre = [genre_classes[label] for label in y_test_encoded]
y_pred_genre = [genre_classes[label] for label in y_pred_classes_2]

# Plot the multiclass confusion matrix for Set 1
plot_multiclass_confusion_matrix(y_true_genre, y_pred_genre, genre_classes,␣
↪"Confusion Matrix - Set 2")
```

```
924/924 [==============================] - 3s 3ms/step
```



9.4 Comparison

- Accuracy: Set 1 has a slightly higher accuracy (0.78) compared to Set 2 (0.76).
- Precision: Set 1 generally has higher precision values for most classes compared to Set 2.
- Recall: Set 1 also has higher recall values for most classes.
- F1-Score: Set 1 achieves higher F1-scores for most classes, indicating a better balance between precision and recall.

Based on the provided classification reports, Set 1 with a single layer LSTM and specific parameter settings appears to perform better than Set 2 with two layers of LSTM. However, it's important to note that other factors, such as hyperparameter tuning and dataset size, can also influence model performance. Further optimization may be needed to achieve the best results.

[ ]: