

STEP 1: Load Dataset & Preprocess Tweets

```
!pip install nltk pandas
```

```
Requirement already satisfied: nltk in /usr/local/lib/python3.12/dist-package
Requirement already satisfied: pandas in /usr/local/lib/python3.12/dist-packa
Requirement already satisfied: click in /usr/local/lib/python3.12/dist-packag
Requirement already satisfied: joblib in /usr/local/lib/python3.12/dist-packa
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.12/d
Requirement already satisfied: tqdm in /usr/local/lib/python3.12/dist-package
Requirement already satisfied: numpy>=1.26.0 in /usr/local/lib/python3.12/dis
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/pytho
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.12/dist
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.12/di
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.12/dist-pac
```

STEP 1: Check Dataset is Loaded Properly

```
import pandas as pd

df = pd.read_csv("/content/Twitter_Data.csv")
print(df.shape)
```

```
(162980, 2)
```

STEP 2: Check Column Names

```
print(df.columns)
```

```
Index(['clean_text', 'category'], dtype='object')
```

STEP 3: Extract Tweets Correctly

```
tweets = df['clean_text'].astype(str)
print(tweets.head())
```

```
0    when modi promised "minimum government maximum...
1    talk all the nonsense and continue all the dra...
2    what did just say vote for modi  welcome bjp t...
3    asking his supporters prefix chowkidar their n...
4    answer who among these the most powerful world...
Name: clean_text, dtype: object
```

STEP 4: Preprocessing

```
import re
```

```
def preprocess_tweet(tweet):
    tweet = re.sub(r"http\S+", "", tweet)
    tweet = re.sub(r"@w+", "", tweet)
    tweet = re.sub(r"#w+", "", tweet)
    tweet = re.sub(r"^[a-zA-Z\s]", "", tweet)
    tweet = tweet.lower().strip()
    return tweet

cleaned_tweets = tweets.apply(preprocess_tweet)

print(cleaned_tweets.head())
```

```
0    when modi promised minimum government maximum ...
1    talk all the nonsense and continue all the dra...
2    what did just say vote for modi welcome bjp t...
3    asking his supporters prefix chowkidar their n...
4    answer who among these the most powerful world...
Name: clean_text, dtype: object
```

STEP 5: Tokenization

```
import nltk
nltk.download('punkt')

from nltk.tokenize import word_tokenize

tokenized_tweets = cleaned_tweets.apply(word_tokenize)

print(tokenized_tweets.head())
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data] Package punkt is already up-to-date!
0    [when, modi, promised, minimum, government, ma...
1    [talk, all, the, nonsense, and, continue, all,...
2    [what, did, just, say, vote, for, modi, welcom...
3    [asking, his, supporters, prefix, chowkidar, t...
4    [answer, who, among, these, the, most, powerfu...
Name: clean_text, dtype: object
```

STEP 6: POS Tagging

```
nltk.download('averaged_perceptron_tagger')
from nltk import pos_tag

tagged_tweets = tokenized_tweets.apply(pos_tag)

print(tagged_tweets.head())
```

```
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data] /root/nltk_data...
[nltk_data] Package averaged_perceptron_tagger is already up-to-
[nltk_data] date!
0    [(when, WRB), (modi, NN), (promised, VBD), (mi...
1    [(talk, NN), (all, PDT), (the, DT), (nonsense,...
```

```
2    [(what, WP), (did, VBD), (just, RB), (say, VB)...
3    [(asking, VBG), (his, PRP$), (supporters, NNS)...
4    [(answer, NN), (who, WP), (among, IN), (these,...
Name: clean_text, dtype: object
```