Indian Institute of Information Technology Surat

Lab Report on
Machine Learning (CS 601) Practical

Submitted by

[RAHUL KUMAR SINGH] (UI21CS44)

Course Faculty

Dr. Pradeep Kumar Roy

Dr. Rajesh K. Ahir

Department of Computer Science and Engineering
Indian Institute of Information Technology Surat
Gujarat-394190, India

Jan-2024

Lab No: 4

Aim:

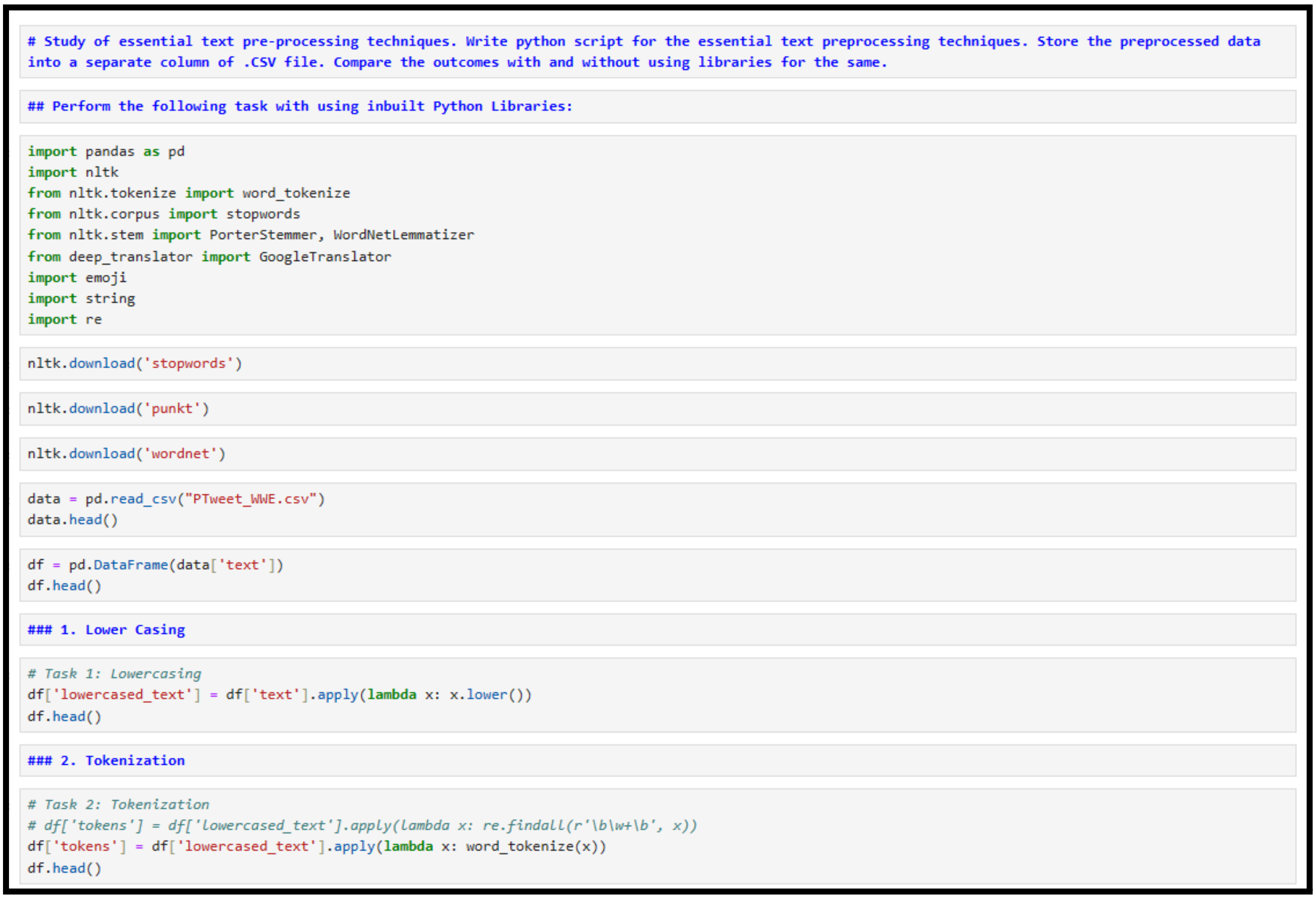
Study of essential text pre-processing techniques. Write python script for the essential text preprocessing techniques. Store the preprocessed data into a separate column of .CSV file. Compare the outcomes with and without using libraries for the same.

Description:

Perform the following task with using inbuilt Python Libraries:

- Lower Casing: Converts text into lower case text. It Helps ensure uniformity in text analysis and processing, as it treats uppercase and lowercase forms of words as the same.
- Tokenization: Break the text into individual words or tokens. It Facilitates analysis at the word level, making it easier to extract meaningful information and perform various natural language processing tasks.
- Punctuation Mark Removal: Eliminate punctuation marks from the text. Enhances the accuracy of text analysis by removing non-alphanumeric characters that don't contribute to the core meaning of the text.
- Stop Word Removal: Exclude common words (stop words) like "and," "the," and "is" that don't carry significant meaning. Improves the efficiency of text processing and analysis by focusing on content-bearing words.
- Stemming: Reduce words to their root or base form by removing suffixes. Aims to group variations of a word together, simplifying analysis and information retrieval. For example, "running" becomes "run."
- Lemmatization: Similar to Stemming but considers the word's context to reduce it to its base or dictionary form (lemma). Results in more accurate representation of the base form of a word, addressing potential ambiguities introduced by stemming.
- Translation: Convert text from one language to another. Facilitates cross-language communication and analysis, enabling understanding of content in different linguistic contexts.
- Emoji to Text: Translate emojis (emotion icons) into their corresponding textual representation. Helps in extracting meaning from textual data that includes emojis, making it easier for analysis and understanding sentiment.

Source Code:



```
# Study of essential text pre-processing techniques. Write python script for the essential text preprocessing techniques. Store the preprocessed data
into a separate column of .CSV file. Compare the outcomes with and without using libraries for the same.
```

```
## Perform the following task with using inbuilt Python Libraries:
```

```python
import pandas as pd
import nltk
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer, WordNetLemmatizer
from deep_translator import GoogleTranslator
import emoji
import string
import re
```

```python
nltk.download('stopwords')
```

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

```python
nltk.download('wordnet')
```

```python
data = pd.read_csv("PTweet_WWE.csv")
data.head()
```

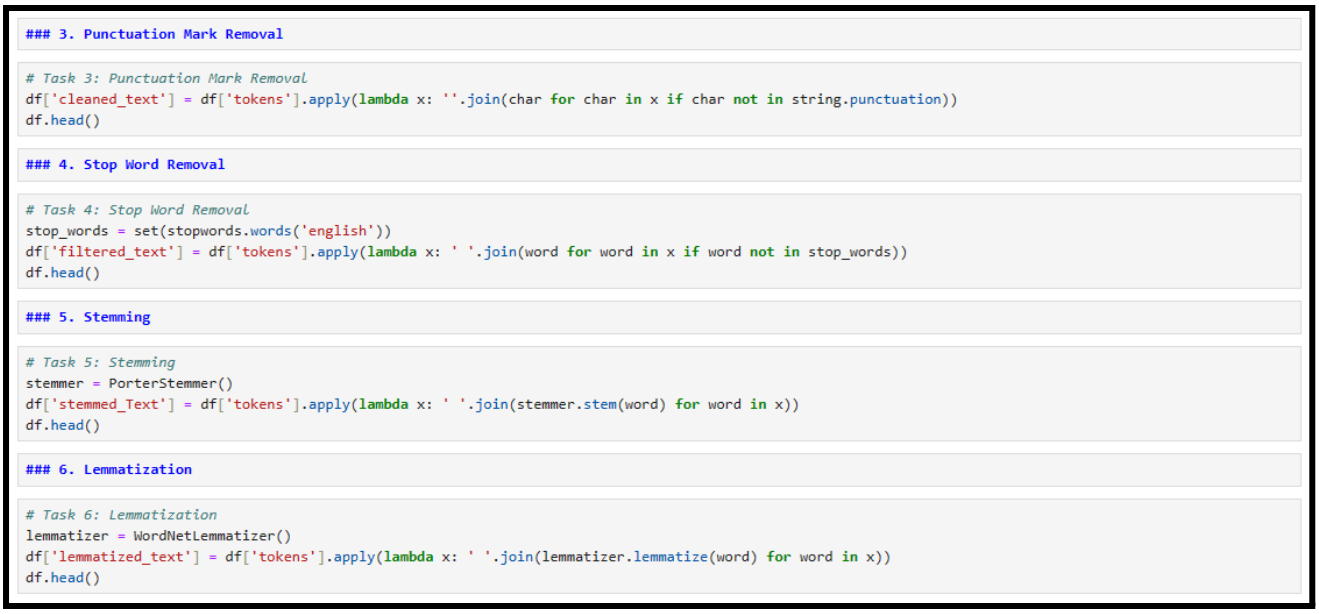
```python
df = pd.DataFrame(data['text'])
df.head()
```

```
### 1. Lower Casing
```

```python
# Task 1: Lowercasing
df['lowercased_text'] = df['text'].apply(lambda x: x.lower())
df.head()
```

```
### 2. Tokenization
```

```python
# Task 2: Tokenization
# df['tokens'] = df['lowercased_text'].apply(lambda x: re.findall(r'\b\w+\b', x))
df['tokens'] = df['lowercased_text'].apply(lambda x: word_tokenize(x))
df.head()
```



```
### 3. Punctuation Mark Removal
```

```python
# Task 3: Punctuation Mark Removal
df['cleaned_text'] = df['tokens'].apply(lambda x: ''.join(char for char in x if char not in string.punctuation))
df.head()
```

```
### 4. Stop Word Removal
```

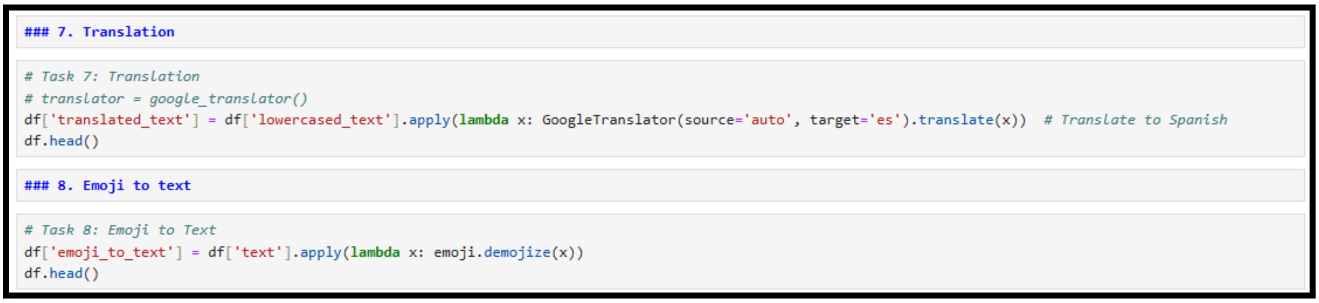
```python
# Task 4: Stop Word Removal
stop_words = set(stopwords.words('english'))
df['filtered_text'] = df['tokens'].apply(lambda x: ' '.join(word for word in x if word not in stop_words))
df.head()
```

```
### 5. Stemming
```

```python
# Task 5: Stemming
stemmer = PorterStemmer()
df['stemmed_Text'] = df['tokens'].apply(lambda x: ' '.join(stemmer.stem(word) for word in x))
df.head()
```

```
### 6. Lemmatization
```

```python
# Task 6: Lemmatization
lemmatizer = WordNetLemmatizer()
df['lemmatized_text'] = df['tokens'].apply(lambda x: ' '.join(lemmatizer.lemmatize(word) for word in x))
df.head()
```

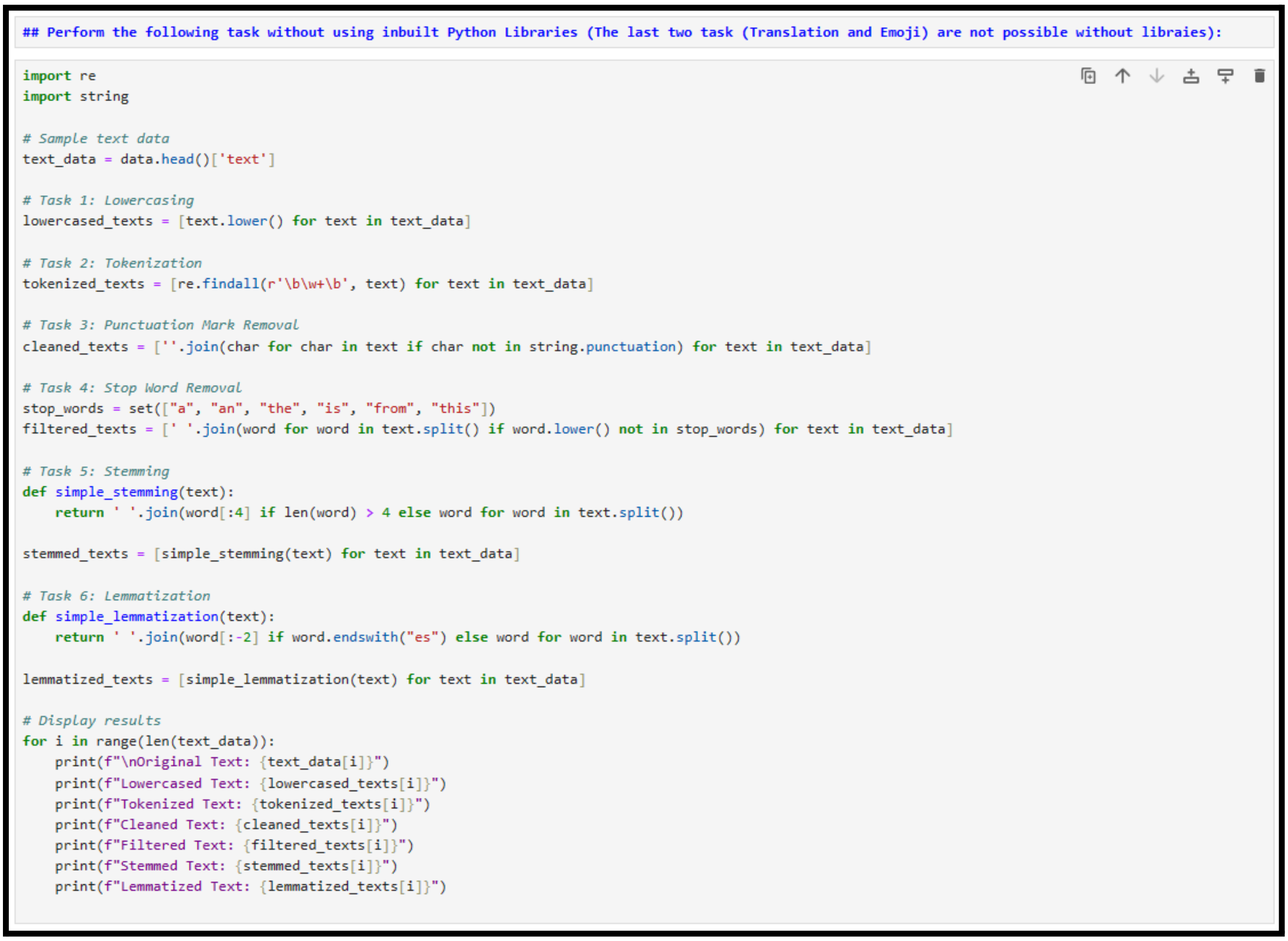


```
### 7. Translation
```

```
# Task 7: Translation
# translator = google_translator()
df['translated_text'] = df['lowercased_text'].apply(lambda x: GoogleTranslator(source='auto', target='es').translate(x))  # Translate to Spanish
df.head()
```

```
### 8. Emoji to text
```

```
# Task 8: Emoji to Text
df['emoji_to_text'] = df['text'].apply(lambda x: emoji.demojize(x))
df.head()
```

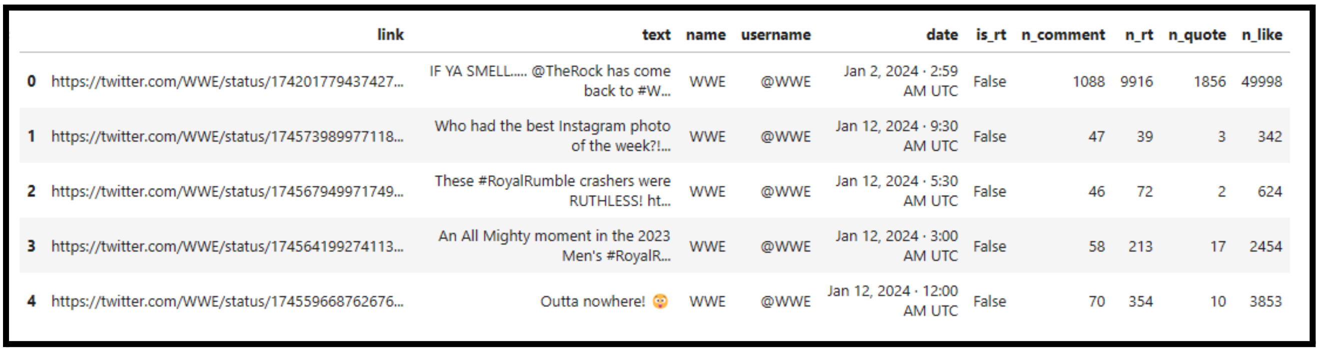


```
## Perform the following task without using inbuilt Python Libraries (The last two task (Translation and Emoji) are not possible without libraies):
```

```
import re
import string

# Sample text data
text_data = data.head()['text']

# Task 1: Lowercasing
lowercased_texts = [text.lower() for text in text_data]

# Task 2: Tokenization
tokenized_texts = [re.findall(r'\b\w+\b', text) for text in text_data]

# Task 3: Punctuation Mark Removal
cleaned_texts = [''.join(char for char in text if char not in string.punctuation) for text in text_data]

# Task 4: Stop Word Removal
stop_words = set(["a", "an", "the", "is", "from", "this"])
filtered_texts = [' '.join(word for word in text.split() if word.lower() not in stop_words) for text in text_data]

# Task 5: Stemming
def simple_stemming(text):
    return ' '.join(word[:4] if len(word) > 4 else word for word in text.split())

stemmed_texts = [simple_stemming(text) for text in text_data]

# Task 6: Lemmatization
def simple_lemmatization(text):
    return ' '.join(word[:-2] if word.endswith("es") else word for word in text.split())

lemmatized_texts = [simple_lemmatization(text) for text in text_data]

# Display results
for i in range(len(text_data)):
    print(f"\nOriginal Text: {text_data[i]}")
    print(f"Lowercased Text: {lowercased_texts[i]}")
    print(f"Tokenized Text: {tokenized_texts[i]}")
    print(f"Cleaned Text: {cleaned_texts[i]}")
    print(f"Filtered Text: {filtered_texts[i]}")
    print(f"Stemmed Text: {stemmed_texts[i]}")
    print(f"Lemmatized Text: {lemmatized_texts[i]}")
```

Output:

Twitter Data:

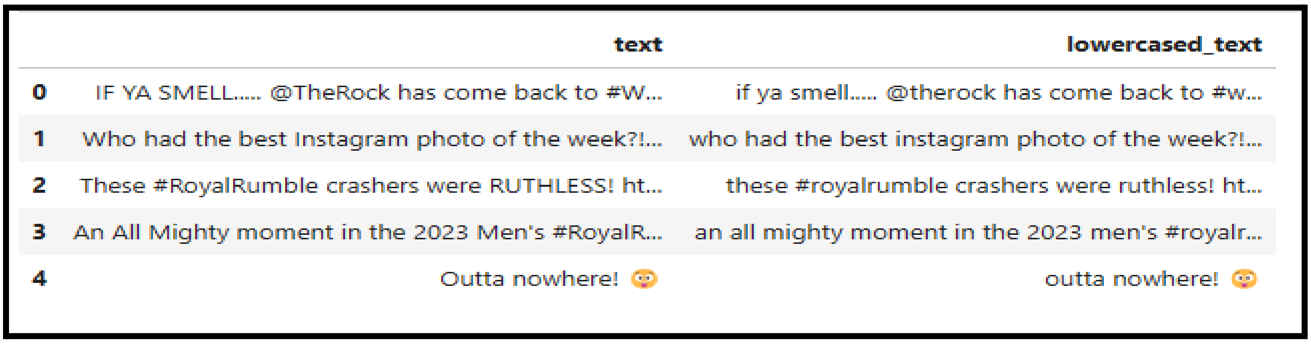


	link	text	name	username	date	is_rt	n_comment	n_rt	n_quote	n_like
0	https://twitter.com/WWE/status/174201779437427...	IF YA SMELL..... @TheRock has come back to #W...	WWE	@WWE	Jan 2, 2024 · 2:59 AM UTC	False	1088	9916	1856	49998
1	https://twitter.com/WWE/status/174573989977118...	Who had the best Instagram photo of the week?!...	WWE	@WWE	Jan 12, 2024 · 9:30 AM UTC	False	47	39	3	342
2	https://twitter.com/WWE/status/174567949971749...	These #RoyalRumble crashers were RUTHLESS! ht...	WWE	@WWE	Jan 12, 2024 · 5:30 AM UTC	False	46	72	2	624
3	https://twitter.com/WWE/status/174564199274113...	An All Mighty moment in the 2023 Men's #RoyalR...	WWE	@WWE	Jan 12, 2024 · 3:00 AM UTC	False	58	213	17	2454
4	https://twitter.com/WWE/status/174559668762676...	Outta nowhere! 😜	WWE	@WWE	Jan 12, 2024 · 12:00 AM UTC	False	70	354	10	3853

Figure 4.0 Twitter Data

Perform the following task with using inbuilt Python Libraries:

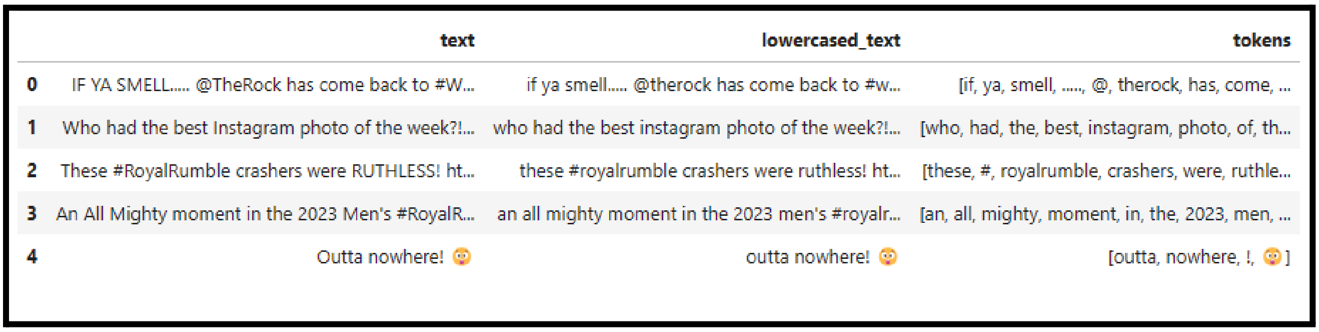
1. Lower Casing



	text	lowercased_text
0	IF YA SMELL..... @TheRock has come back to #W...	if ya smell..... @therock has come back to #w...
1	Who had the best Instagram photo of the week?!...	who had the best instagram photo of the week?!...
2	These #RoyalRumble crashers were RUTHLESS! ht...	these #royalrumble crashers were ruthless! ht...
3	An All Mighty moment in the 2023 Men's #RoyalR...	an all mighty moment in the 2023 men's #royalr...
4	Outta nowhere! 😳	outta nowhere! 😳

Figure 4.1.1 Lower Casing

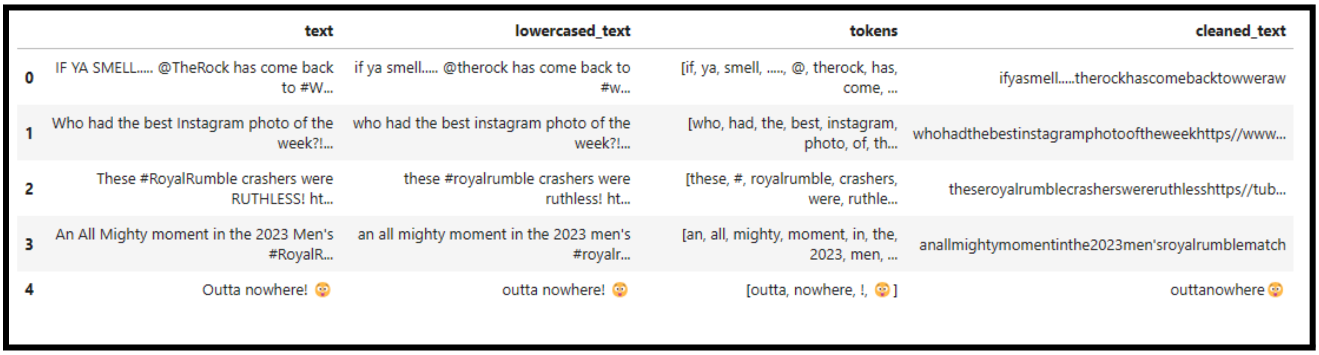
2. Tokenization



	text	lowercased_text	tokens
0	IF YA SMELL..... @TheRock has come back to #W...	if ya smell..... @therock has come back to #w...	[if, ya, smell, ..., @, therock, has, come, ...
1	Who had the best Instagram photo of the week?!...	who had the best instagram photo of the week?!...	[who, had, the, best, instagram, photo, of, th...
2	These #RoyalRumble crashers were RUTHLESS! ht...	these #royalrumble crashers were ruthless! ht...	[these, #, royalrumble, crashers, were, ruthle...
3	An All Mighty moment in the 2023 Men's #RoyalR...	an all mighty moment in the 2023 men's #royalr...	[an, all, mighty, moment, in, the, 2023, men, ...
4	Outta nowhere! 😳	outta nowhere! 😳	[outta, nowhere, !, 😳]

Figure 4.1.2 Tokenization

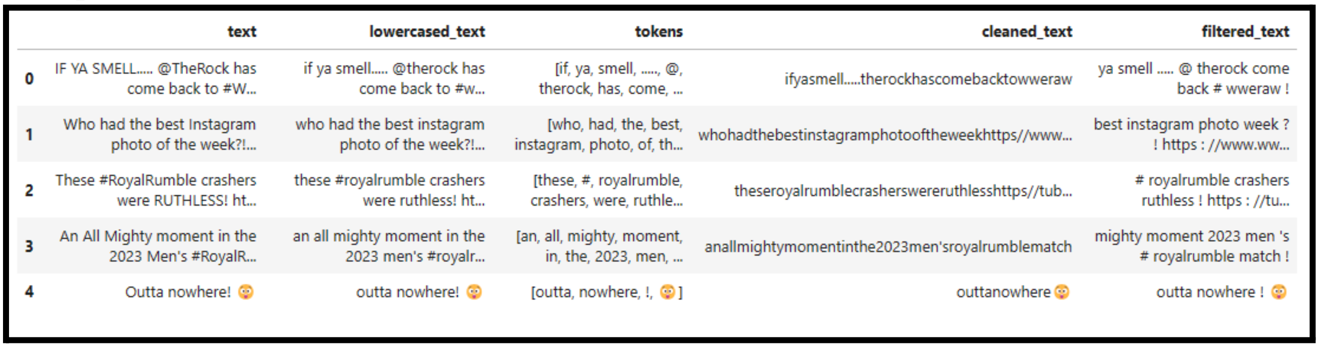
3. Punctuation Mark Removal



	text	lowercased_text	tokens	cleaned_text
0	IF YA SMELL..... @TheRock has come back to #W...	if ya smell..... @therock has come back to #w...	[if, ya, smell, ..., @, therock, has, come, ...	ifyasmell.....therockhascomebacktowweraw
1	Who had the best Instagram photo of the week?!...	who had the best instagram photo of the week?!...	[who, had, the, best, instagram, photo, of, th...	whohadthebestinstagramphotooftheweekhttps//www...
2	These #RoyalRumble crashers were RUTHLESS! ht...	these #royalrumble crashers were ruthless! ht...	[these, #, royalrumble, crashers, were, ruthle...	theseroyalrumblecrasherswereruthlesshttps//tub...
3	An All Mighty moment in the 2023 Men's #RoyalR...	an all mighty moment in the 2023 men's #royalr...	[an, all, mighty, moment, in, the, 2023, men, ...	anallmightymomentinthe2023men'sroyalrumblematch
4	Outta nowhere! 😳	outta nowhere! 😳	[outta, nowhere, !, 😳]	outtanowhere😳

Figure 4.1.3 Punctuation Mark Removal

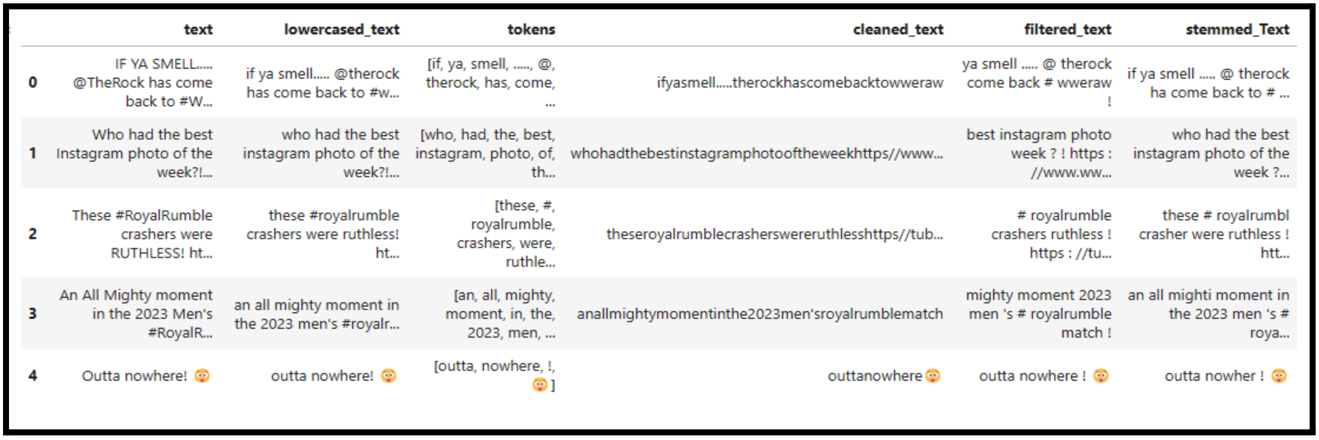
4. Stop Word Removal



	text	lowercased_text	tokens	cleaned_text	filtered_text
0	IF YA SMELL..... @TheRock has come back to #W...	if ya smell..... @therock has come back to #w...	[if, ya, smell, ..., @, therock, has, come, ...	ifyasmell.....therockhascomebacktowweraw	ya smell ..... @ therock come back # wweraw !
1	Who had the best Instagram photo of the week?!...	who had the best instagram photo of the week?!...	[who, had, the, best, instagram, photo, of, th...	whohadthebestinstagramphotooftheweekhttps//www...	best instagram photo week ? ! https : //www.ww...
2	These #RoyalRumble crashers were RUTHLESS! ht...	these #royalrumble crashers were ruthless! ht...	[these, #, royalrumble, crashers, were, ruthle...	theseroyalrumblecrasherswereruthlesshttps//tub...	# royalrumble crashers ruthless ! https : //tu...
3	An All Mighty moment in the 2023 Men's #RoyalR...	an all mighty moment in the 2023 men's #royalr...	[an, all, mighty, moment, in, the, 2023, men, ...	anallmightymomentinthe2023men'sroyalrumblematch	mighty moment 2023 men 's # royalrumble match !
4	Outta nowhere! 😳	outta nowhere! 😳	[outta, nowhere, !, 😳]	outtanowhere😳	outta nowhere ! 😳

Figure 4.1.4 Stop Word Removal

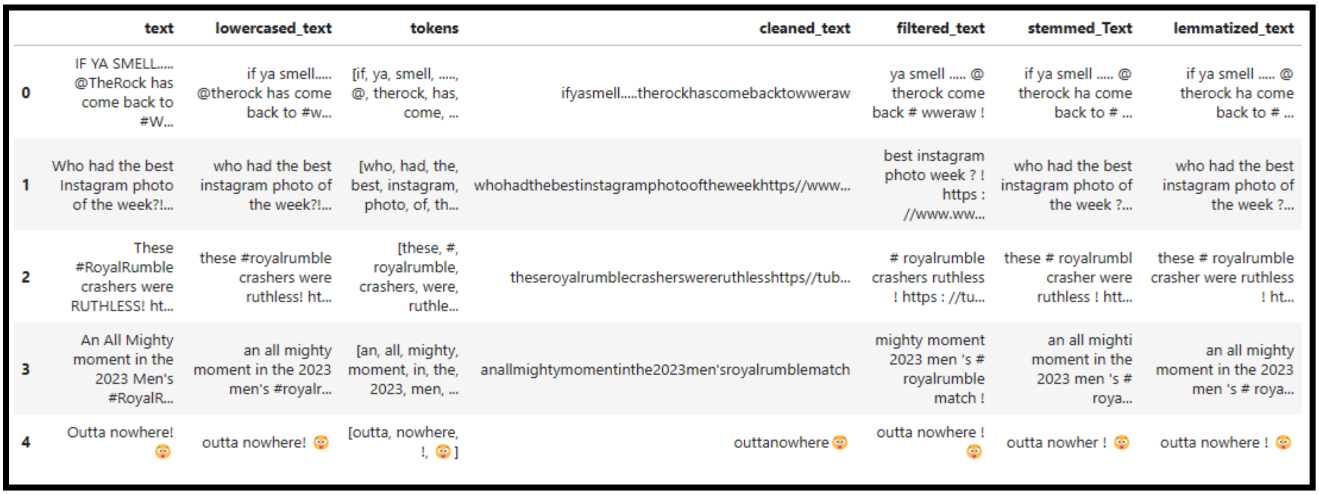
5. Stemming



	text	lowercased_text	tokens	cleaned_text	filtered_text	stemmed_Text
0	IF YA SMELL..... @TheRock has come back to #W...	if ya smell..... @therock has come back to #w...	[if, ya, smell, ..., @, therock, has, come, ...	ifyasmell.....therockhascomebacktowweraw	ya smell ..... @ therock come back # wweraw !	if ya smell ..... @ therock ha come back to # ...
1	Who had the best Instagram photo of the week?!...	who had the best instagram photo of the week?!...	[who, had, the, best, instagram, photo, of, th...	whohadthebestinstagramphotooftheweekhttps//www...	best instagram photo week ? ! https : //www.ww...	who had the best instagram photo of the week ?...
2	These #RoyalRumble crashers were RUTHLESS! ht...	these #royalrumble crashers were ruthless! ht...	[these, #, royalrumble, crashers, were, ruthle...	theseroyalrumblecrasherswereruthlesshttps//tub...	# royalrumble crashers ruthless ! https : //tu...	these # royalrumbl crasher were ruthless ! htt...
3	An All Mighty moment in the 2023 Men's #RoyalR...	an all mighty moment in the 2023 men's #royalr...	[an, all, mighty, moment, in, the, 2023, men, ...	anallmightymomentinthe2023men'sroyalrumblematch	mighty moment 2023 men 's # royalrumble match !	an all mighti moment in the 2023 men 's # roya...
4	Outta nowhere! 😳	outta nowhere! 😳	[outta, nowhere, !, 😳]	outtanowhere😳	outta nowhere ! 😳	outta nowher ! 😳

Figure 4.1.5 Stemming

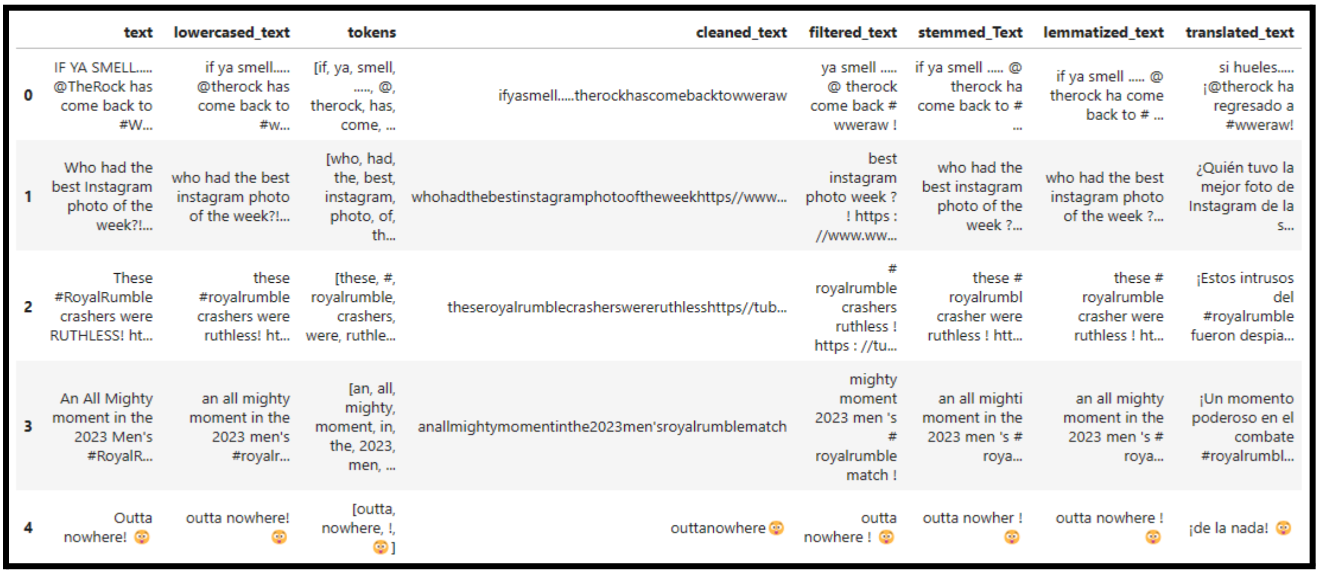
6. Lemmatization



	text	lowercased_text	tokens	cleaned_text	filtered_text	stemmed_Text	lemmatized_text
0	IF YA SMELL..... @TheRock has come back to #W...	if ya smell..... @therock has come back to #w...	[if, ya, smell, ..., @, therock, has, come, ...	ifyasmell.....therockhascomebacktowweraw	ya smell ..... @ therock come back # wweraw !	if ya smell ..... @ therock ha come back to # ...	if ya smell ..... @ therock ha come back to # ...
1	Who had the best Instagram photo of the week?!...	who had the best instagram photo of the week?!...	[who, had, the, best, instagram, photo, of, th...	whohadthebestinstagramphotooftheweekhttps//www...	best instagram photo week ? ! https : //www.ww...	who had the best instagram photo of the week ?...	who had the best instagram photo of the week ?...
2	These #RoyalRumble crashers were RUTHLESS! ht...	these #royalrumble crashers were ruthless! ht...	[these, #, royalrumble, crashers, were, ruthle...	theseroyalrumblecrasherswereruthlesshttps//tub...	# royalrumble crashers ruthless ! https : //tu...	these # royalrumbl crasher were ruthless ! htt...	these # royalrumble crasher were ruthless ! ht...
3	An All Mighty moment in the 2023 Men's #RoyalR...	an all mighty moment in the 2023 men's #royalr...	[an, all, mighty, moment, in, the, 2023, men, ...	anallmightymomentinthe2023men'sroyalrumblematch	mighty moment 2023 men 's # royalrumble match !	an all mighti moment in the 2023 men 's # roya...	an all mighty moment in the 2023 men 's # roya...
4	Outta nowhere! 😳	outta nowhere! 😳	[outta, nowhere, !, 😳]	outtanowhere😳	outta nowhere ! 😳	outta nowher ! 😳	outta nowhere ! 😳

Figure 4.1.6 Lemmatization

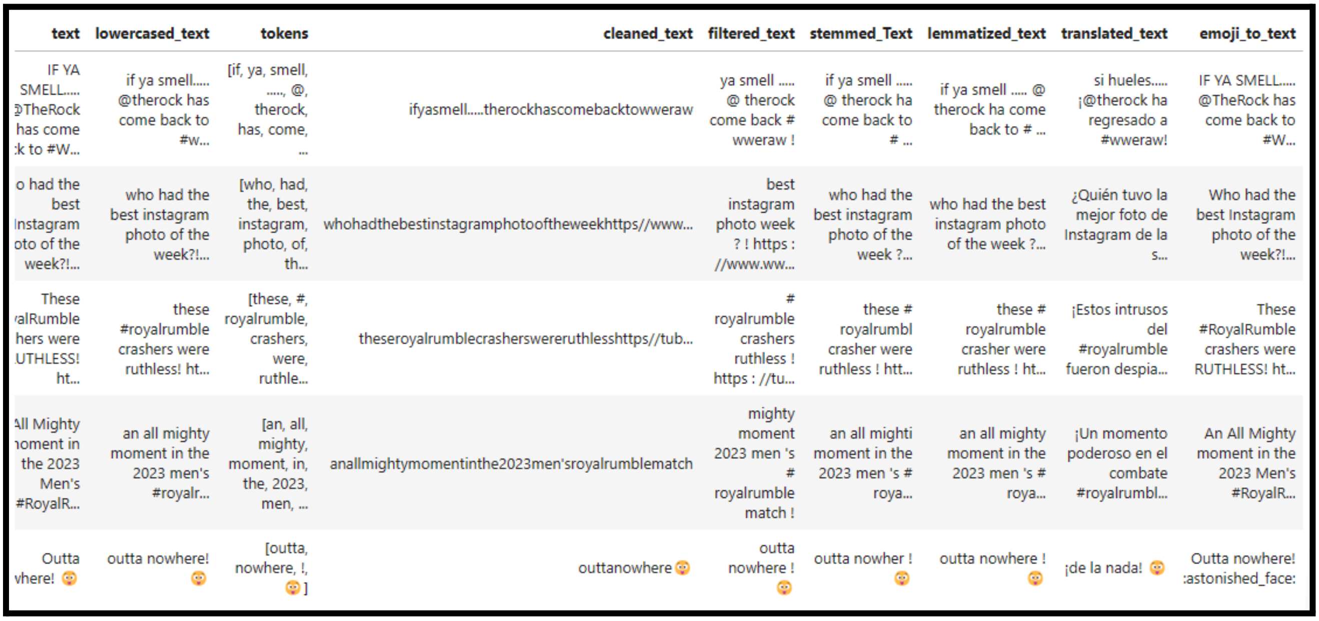
7. Translation



	text	lowercased_text	tokens	cleaned_text	filtered_text	stemmed_Text	lemmatized_text	translated_text
0	IF YA SMELL..... @TheRock has come back to #W...	if ya smell..... @therock has come back to #w...	[if, ya, smell, ..., @, therock, has, come, ...	ifyasmell.....therockhascomebacktowweraw	ya smell ..... @ therock come back # wweraw !	if ya smell ..... @ therock ha come back to # ...	if ya smell ..... @ therock ha come back to # ...	si hueles..... ¡@therock ha regresado a #wweraw!
1	Who had the best Instagram photo of the week?!...	who had the best instagram photo of the week?!...	[who, had, the, best, instagram, photo, of, th...	whohadthebestinstagramphotooftheweekhttps//www...	best instagram photo week ? ! https : //www.ww...	who had the best instagram photo of the week ?...	who had the best instagram photo of the week ?...	¿Quién tuvo la mejor foto de Instagram de la s...
2	These #RoyalRumble crashers were RUTHLESS! ht...	these #royalrumble crashers were ruthless! ht...	[these, #, royalrumble, crashers, were, ruthle...	theseroyalrumblecrasherswereruthlesshttps//tub...	# royalrumble crashers ruthless ! https : //tu...	these # royalrumbl crasher were ruthless ! htt...	these # royalrumble crasher were ruthless ! ht...	¡Estos intrusos del #royalrumble fueron despia...
3	An All Mighty moment in the 2023 Men's #RoyalR...	an all mighty moment in the 2023 men's #royalr...	[an, all, mighty, moment, in, the, 2023, men, ...	anallmightymomentinthe2023men'sroyalrumblematch	mighty moment 2023 men 's # royalrumble match !	an all mighti moment in the 2023 men 's # roya...	an all mighty moment in the 2023 men 's # roya...	¡Un momento poderoso en el combate #royalrumbl...
4	Outta nowhere! 😳	outta nowhere! 😳	[outta, nowhere, !, 😳]	outtanowhere😳	outta nowhere ! 😳	outta nowher ! 😳	outta nowhere ! 😳	¡de la nada! 😳

Figure 4.1.7 Translation

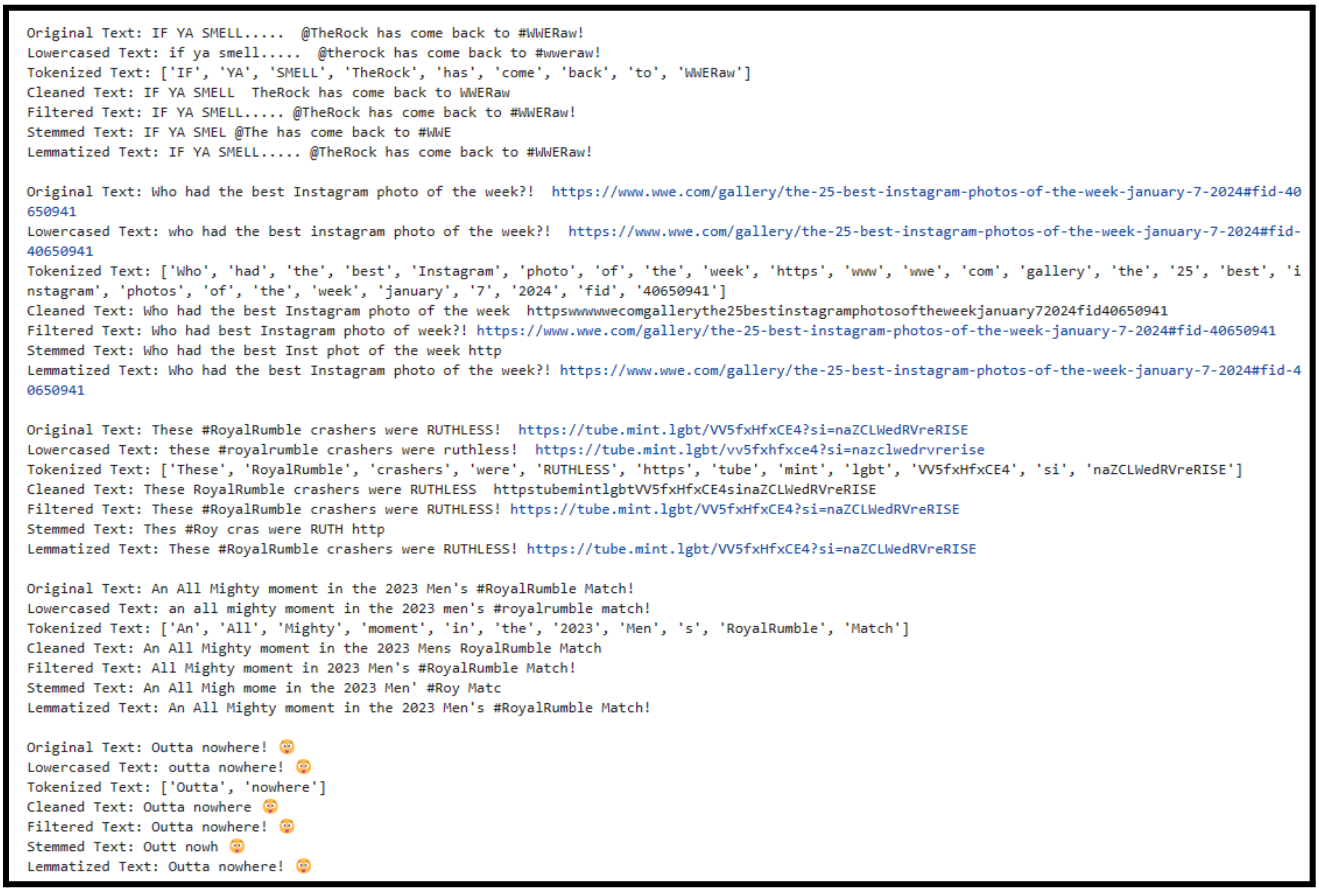
8. Emoji to text



text	lowercased_text	tokens	cleaned_text	filtered_text	stemmed_Text	lemmatized_text	translated_text	emoji_to_text
IF YA SMELL..... ɔTheRock has come :k to #W...	if ya smell..... @therock has come back to #w...	[if, ya, smell, ..., @, therock, has, come, ...	ifyasmell.....therockhascomebacktowweraw	ya smell ..... @ therock come back # wweraw !	if ya smell ..... @ therock ha come back to # ...	if ya smell ..... @ therock ha come back to # ...	si hueles..... ¡@therock ha regresado a #wweraw!	IF YA SMELL..... @TheRock has come back to #W...
o had the best Instagram oto of the week?!...	who had the best instagram photo of the week?!...	[who, had, the, best, instagram, photo, of, th...	whohadthebestinstagramphotooftheweekhttps//www...	best instagram photo week ? ! https : //www.ww...	who had the best instagram photo of the week ?...	who had the best instagram photo of the week ?...	¿Quién tuvo la mejor foto de Instagram de la s...	Who had the best Instagram photo of the week?!...
These ɪalRumble hers were UTHLESS! ht...	these #royalrumble crashers were ruthless! ht...	[these, #, royalrumble, crashers, were, ruthle...	theseroyalrumblecrasherswereruthlesshttps//tub...	# royalrumble crashers ruthless ! https : //tu...	these # royalrumbl crasher were ruthless ! htt...	these # royalrumble crasher were ruthless ! ht...	¡Estos intrusos del #royalrumble fueron despia...	These #RoyalRumble crashers were RUTHLESS! ht...
ᐱll Mighty ɩoment in the 2023 Men's #RoyalR...	an all mighty moment in the 2023 men's #royalr...	[an, all, mighty, moment, in, the, 2023, men, ...	anallmightymomentinthe2023men'sroyalrumblematch	mighty moment 2023 men 's # royalrumble match !	an all mighti moment in the 2023 men 's # roya...	an all mighty moment in the 2023 men 's # roya...	¡Un momento poderoso en el combate #royalrumbl...	An All Mighty moment in the 2023 Men's #RoyalR...
Outta vhere! 😲	outta nowhere! 😲	[outta, nowhere, !, 😲]	outtanowhere😲	outta nowhere ! 😲	outta nowher ! 😲	outta nowhere ! 😲	¡de la nada! 😲	Outta nowhere! :astonished_face:

Figure 4.1.8 Emoji To Text

Perform the following task without using inbuilt Python Libraries (Wont't work for last two tasks):



```
Original Text: IF YA SMELL.....  @TheRock has come back to #WWERaw!
Lowercased Text: if ya smell.....  @therock has come back to #wweraw!
Tokenized Text: ['IF', 'YA', 'SMELL', 'TheRock', 'has', 'come', 'back', 'to', 'WWERaw']
Cleaned Text: IF YA SMELL  TheRock has come back to WWERaw
Filtered Text: IF YA SMELL..... @TheRock has come back to #WWERaw!
Stemmed Text: IF YA SMEL @The has come back to #WWE
Lemmatized Text: IF YA SMELL..... @TheRock has come back to #WWERaw!

Original Text: Who had the best Instagram photo of the week?!  https://www.wwe.com/gallery/the-25-best-instagram-photos-of-the-week-january-7-2024#fid-40
650941
Lowercased Text: who had the best instagram photo of the week?!  https://www.wwe.com/gallery/the-25-best-instagram-photos-of-the-week-january-7-2024#fid-
40650941
Tokenized Text: ['Who', 'had', 'the', 'best', 'Instagram', 'photo', 'of', 'the', 'week', 'https', 'www', 'wwe', 'com', 'gallery', 'the', '25', 'best', 'i
nstagram', 'photos', 'of', 'the', 'week', 'january', '7', '2024', 'fid', '40650941']
Cleaned Text: Who had the best Instagram photo of the week  httpswwwwwecomgallerythe25bestinstagramphotosoftheweekjanuary72024fid40650941
Filtered Text: Who had best Instagram photo of week?! https://www.wwe.com/gallery/the-25-best-instagram-photos-of-the-week-january-7-2024#fid-40650941
Stemmed Text: Who had the best Inst phot of the week http
Lemmatized Text: Who had the best Instagram photo of the week?! https://www.wwe.com/gallery/the-25-best-instagram-photos-of-the-week-january-7-2024#fid-4
0650941

Original Text: These #RoyalRumble crashers were RUTHLESS!  https://tube.mint.lgbt/VV5fxHfxCE4?si=naZCLWedRVreRISE
Lowercased Text: these #royalrumble crashers were ruthless!  https://tube.mint.lgbt/vv5fxhfxce4?si=nazclwedrvrerise
Tokenized Text: ['These', 'RoyalRumble', 'crashers', 'were', 'RUTHLESS', 'https', 'tube', 'mint', 'lgbt', 'VV5fxHfxCE4', 'si', 'naZCLWedRVreRISE']
Cleaned Text: These RoyalRumble crashers were RUTHLESS  httpstubemintlgbtVV5fxHfxCE4sinaZCLWedRVreRISE
Filtered Text: These #RoyalRumble crashers were RUTHLESS! https://tube.mint.lgbt/VV5fxHfxCE4?si=naZCLWedRVreRISE
Stemmed Text: Thes #Roy cras were RUTH http
Lemmatized Text: These #RoyalRumble crashers were RUTHLESS! https://tube.mint.lgbt/VV5fxHfxCE4?si=naZCLWedRVreRISE

Original Text: An All Mighty moment in the 2023 Men's #RoyalRumble Match!
Lowercased Text: an all mighty moment in the 2023 men's #royalrumble match!
Tokenized Text: ['An', 'All', 'Mighty', 'moment', 'in', 'the', '2023', 'Men', 's', 'RoyalRumble', 'Match']
Cleaned Text: An All Mighty moment in the 2023 Mens RoyalRumble Match
Filtered Text: All Mighty moment in 2023 Men's #RoyalRumble Match!
Stemmed Text: An All Migh mome in the 2023 Men' #Roy Matc
Lemmatized Text: An All Mighty moment in the 2023 Men's #RoyalRumble Match!

Original Text: Outta nowhere! 😲
Lowercased Text: outta nowhere! 😲
Tokenized Text: ['Outta', 'nowhere']
Cleaned Text: Outta nowhere 😲
Filtered Text: Outta nowhere! 😲
Stemmed Text: Outt nowh 😲
Lemmatized Text: Outta nowhere! 😲
```

Figure 4.2 Without Library

Conclusion:

- Lowercasing ensures uniformity, treating uppercase and lowercase forms equally, preventing discrepancies in analysis.
- Tokenization breaks down text into meaningful units, enabling granular analysis at the word level and facilitating various natural language processing tasks.
- Punctuation mark removal eliminates non-alphanumeric characters, reducing noise and focusing on the core meaning of the text.
- Stop word removal improves efficiency by excluding common words, allowing a focus on content-bearing words and enhancing the relevance of analysis.
- Stemming and lemmatization contribute to word form normalization, reducing words to their base form for better consistency and information retrieval.
- Translation enables the understanding of text in different languages, fostering cross-language communication and analysis.
- Emoji-to-text conversion aids in extracting emotional context from textual data, contributing to sentiment analysis and understanding user expressions.