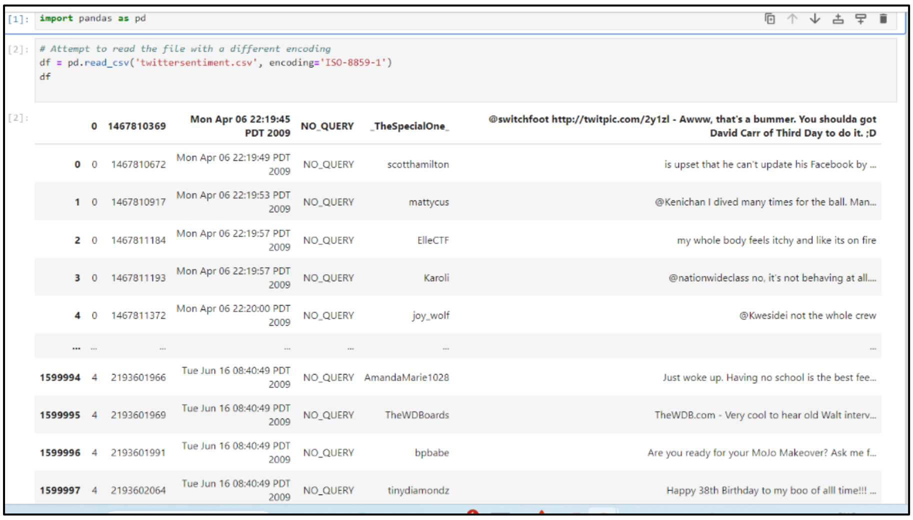
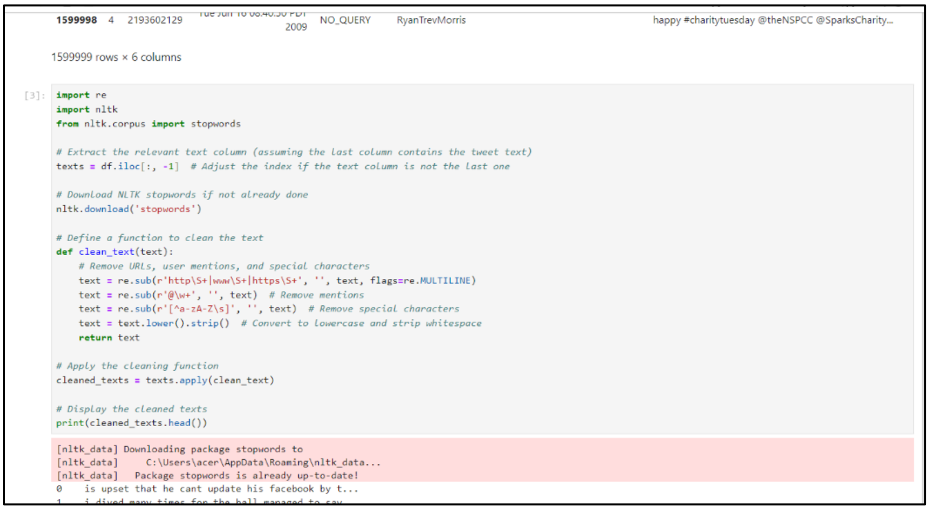


[1]:
```
import pandas as pd
```

[2]:
```
# Attempt to read the file with a different encoding
df = pd.read_csv('twittersentiment.csv', encoding='ISO-8859-1')
df
```

[2]:

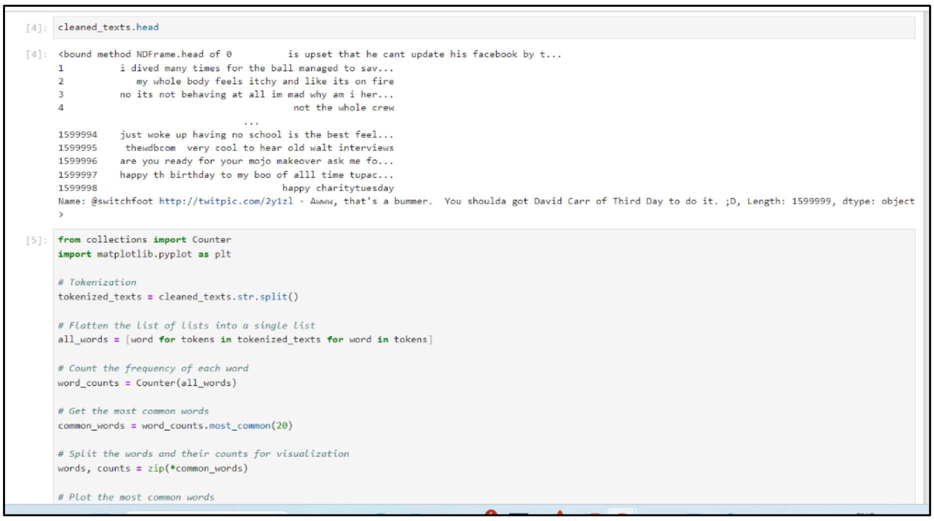
	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D
0	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...
1	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...
2	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
3	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....
4	0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	@Kwesidei not the whole crew
...	...	...	...	...	...	...
1599994	4	2193601966	Tue Jun 16 08:40:49 PDT 2009	NO_QUERY	AmandaMarie1028	Just woke up. Having no school is the best fee...
1599995	4	2193601969	Tue Jun 16 08:40:49 PDT 2009	NO_QUERY	TheWDBoards	TheWDB.com - Very cool to hear old Walt interv...
1599996	4	2193601991	Tue Jun 16 08:40:49 PDT 2009	NO_QUERY	bpbabe	Are you ready for your MoJo Makeover? Ask me f...
1599997	4	2193602064	Tue Jun 16 08:40:49 PDT 2009	NO_QUERY	tinydiamondz	Happy 38th Birthday to my boo of alll time!!! ...
1599998	4	2193602129	Tue Jun 16 08:40:50 PDT 2009	NO_QUERY	RyanTrevMorris	happy #charitytuesday @theNSPCC @SparksCharity...



1599999 rows × 6 columns

[3]:
```
import re
import nltk
from nltk.corpus import stopwords

# Extract the relevant text column (assuming the last column contains the tweet text)
texts = df.iloc[:, -1]  # Adjust the index if the text column is not the last one

# Download NLTK stopwords if not already done
nltk.download('stopwords')

# Define a function to clean the text
def clean_text(text):
    # Remove URLs, user mentions, and special characters
    text = re.sub(r'http\S+|www\S+|https\S+', '', text, flags=re.MULTILINE)
    text = re.sub(r'@\w+', '', text)  # Remove mentions
    text = re.sub(r'[^a-zA-Z\s]', '', text)  # Remove special characters
    text = text.lower().strip()  # Convert to lowercase and strip whitespace
    return text

# Apply the cleaning function
cleaned_texts = texts.apply(clean_text)

# Display the cleaned texts
print(cleaned_texts.head())
```

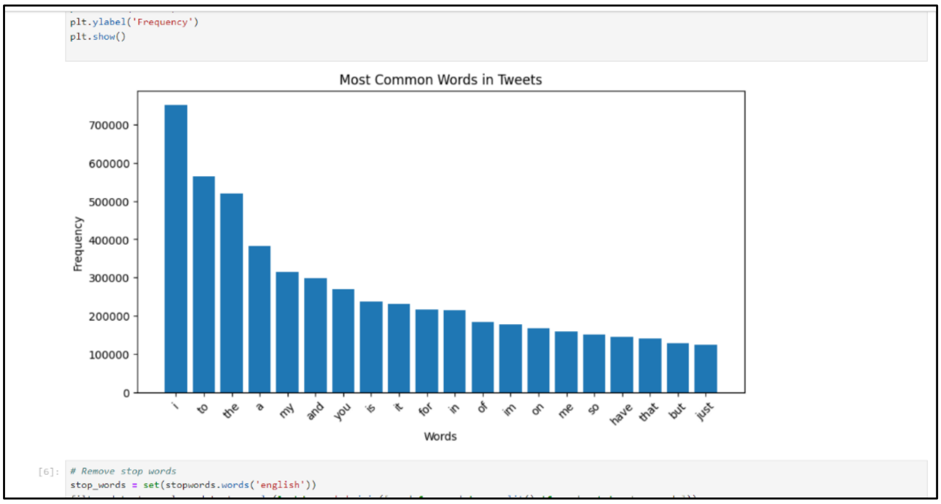
```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\acer\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
0    is upset that he cant update his facebook by t...
1    i dived many times for the ball managed to sav
```



[4]:
```
cleaned_texts.head
```

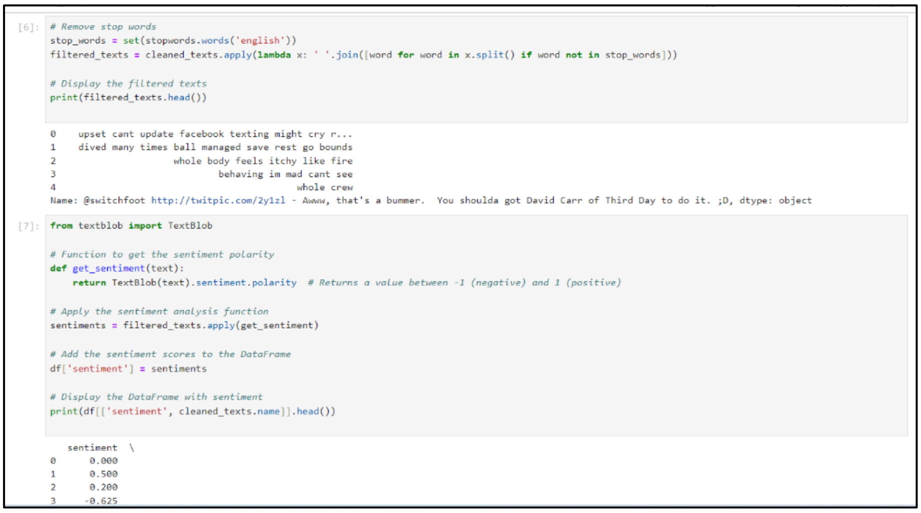
[4]:
```
<bound method NDFrame.head of 0          is upset that he cant update his facebook by t...
1          i dived many times for the ball managed to sav...
2            my whole body feels itchy and like its on fire
3          no its not behaving at all im mad why am i her...
4                                     not the whole crew
                                ...
1599994    just woke up having no school is the best feel...
1599995     thewdbcom  very cool to hear old walt interviews
1599996    are you ready for your mojo makeover ask me fo...
1599997    happy th birthday to my boo of alll time tupac...
1599998                                happy charitytuesday
Name: @switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer.  You shoulda got David Carr of Third Day to do it. ;D, Length: 1599999, dtype: object
>
```

[5]:
```
from collections import Counter
import matplotlib.pyplot as plt

# Tokenization
tokenized_texts = cleaned_texts.str.split()

# Flatten the list of lists into a single list
all_words = [word for tokens in tokenized_texts for word in tokens]

# Count the frequency of each word
word_counts = Counter(all_words)

# Get the most common words
common_words = word_counts.most_common(20)

# Split the words and their counts for visualization
words, counts = zip(*common_words)

# Plot the most common words
```



```
plt.ylabel('Frequency')
plt.show()
```

```
[6]: # Remove stop words
     stop_words = set(stopwords.words('english'))
```

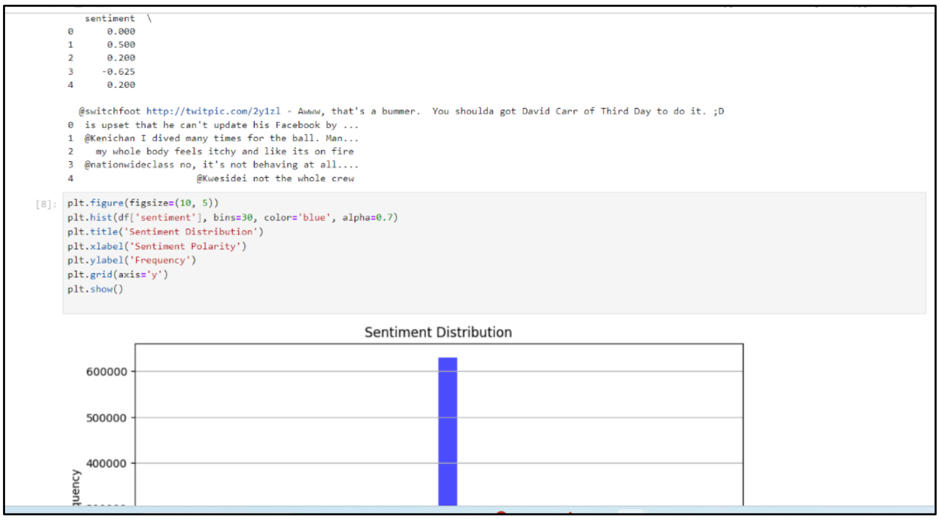


```
[6]: # Remove stop words
     stop_words = set(stopwords.words('english'))
     filtered_texts = cleaned_texts.apply(lambda x: ' '.join([word for word in x.split() if word not in stop_words]))

     # Display the filtered texts
     print(filtered_texts.head())
```

```
0    upset cant update facebook texting might cry r...
1    dived many times ball managed save rest go bounds
2                       whole body feels itchy like fire
3                              behaving im mad cant see
4                                             whole crew
Name: @switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer.  You shoulda got David Carr of Third Day to do it. ;D, dtype: object
```

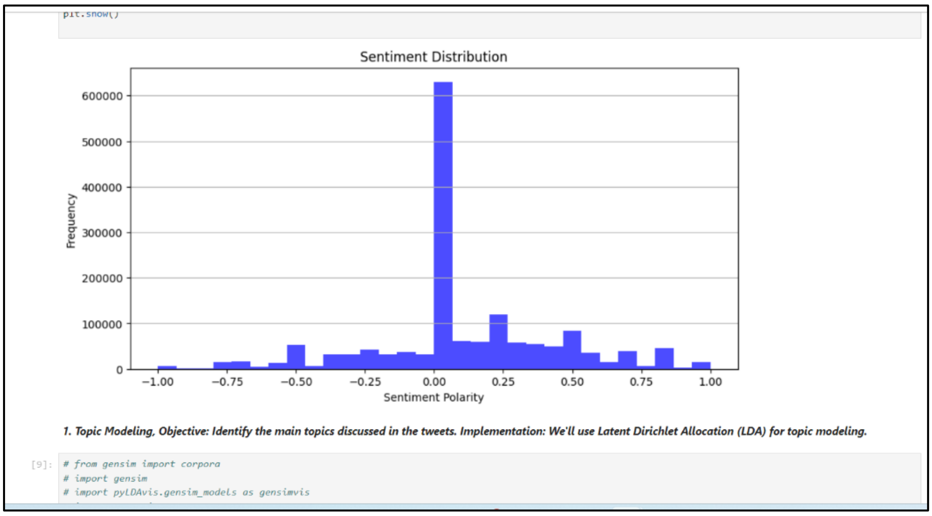
```
[7]: from textblob import TextBlob

     # Function to get the sentiment polarity
     def get_sentiment(text):
         return TextBlob(text).sentiment.polarity  # Returns a value between -1 (negative) and 1 (positive)

     # Apply the sentiment analysis function
     sentiments = filtered_texts.apply(get_sentiment)

     # Add the sentiment scores to the DataFrame
     df['sentiment'] = sentiments

     # Display the DataFrame with sentiment
     print(df[['sentiment', cleaned_texts.name]].head())
```

```
   sentiment  \
0      0.000
1      0.500
2      0.200
3     -0.625
```



```
   sentiment  \
0      0.000
1      0.500
2      0.200
3     -0.625
4      0.200

  @switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer.  You shoulda got David Carr of Third Day to do it. ;D
0  is upset that he can't update his Facebook by ...
1  @Kenichan I dived many times for the ball. Man...
2    my whole body feels itchy and like its on fire
3  @nationwideclass no, it's not behaving at all....
4                      @Kwesidei not the whole crew
```

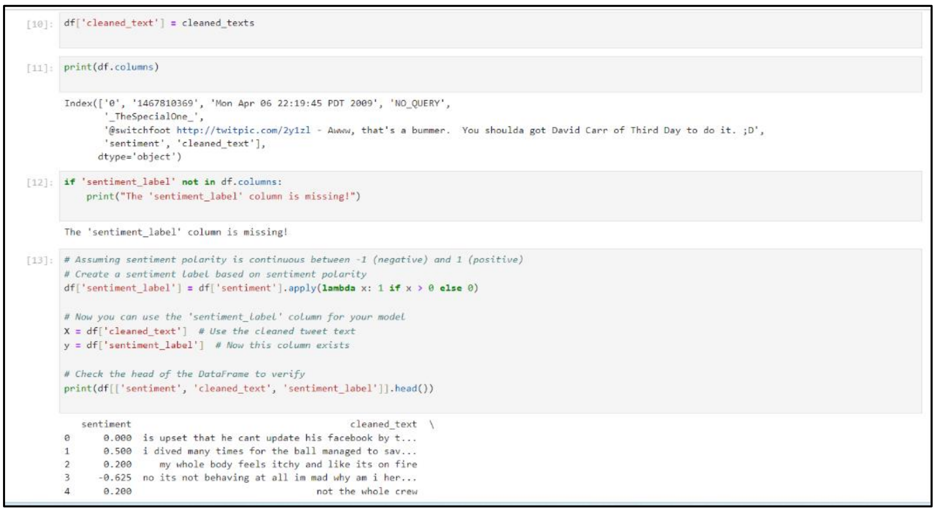
```
[8]: plt.figure(figsize=(10, 5))
     plt.hist(df['sentiment'], bins=30, color='blue', alpha=0.7)
     plt.title('Sentiment Distribution')
     plt.xlabel('Sentiment Polarity')
     plt.ylabel('Frequency')
     plt.grid(axis='y')
     plt.show()
```



```
plt.show()
```

1. Topic Modeling, Objective: Identify the main topics discussed in the tweets. Implementation: We'll use Latent Dirichlet Allocation (LDA) for topic modeling.

```
[9]: # from gensim import corpora
     # import gensim
     # import pyLDAvis.gensim_models as gensimvis
```



```
[10]: df['cleaned_text'] = cleaned_texts
```

```
[11]: print(df.columns)
```

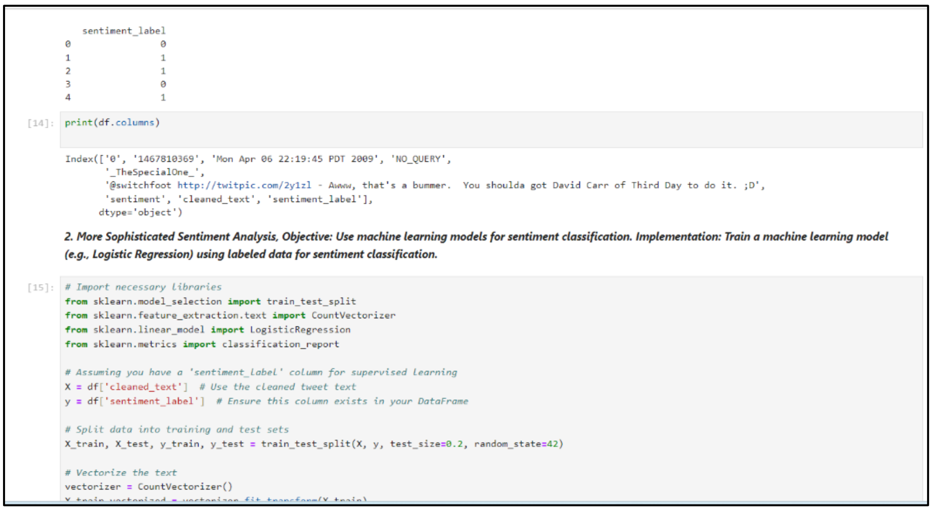
```
Index(['0', '1467810369', 'Mon Apr 06 22:19:45 PDT 2009', 'NO_QUERY',
       '_TheSpecialOne_',
       '@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer.  You shoulda got David Carr of Third Day to do it. ;D',
       'sentiment', 'cleaned_text'],
      dtype='object')
```

```
[12]: if 'sentiment_label' not in df.columns:
          print("The 'sentiment_label' column is missing!")
```

```
The 'sentiment_label' column is missing!
```

```
[13]: # Assuming sentiment polarity is continuous between -1 (negative) and 1 (positive)
      # Create a sentiment label based on sentiment polarity
      df['sentiment_label'] = df['sentiment'].apply(lambda x: 1 if x > 0 else 0)

      # Now you can use the 'sentiment_label' column for your model
      X = df['cleaned_text']  # Use the cleaned tweet text
      y = df['sentiment_label']  # Now this column exists

      # Check the head of the DataFrame to verify
      print(df[['sentiment', 'cleaned_text', 'sentiment_label']].head())
```

```
   sentiment                                       cleaned_text  \
0      0.000  is upset that he cant update his facebook by t...
1      0.500  i dived many times for the ball managed to sav...
2      0.200     my whole body feels itchy and like its on fire
3     -0.625  no its not behaving at all im mad why am i her...
4      0.200                                   not the whole crew
```



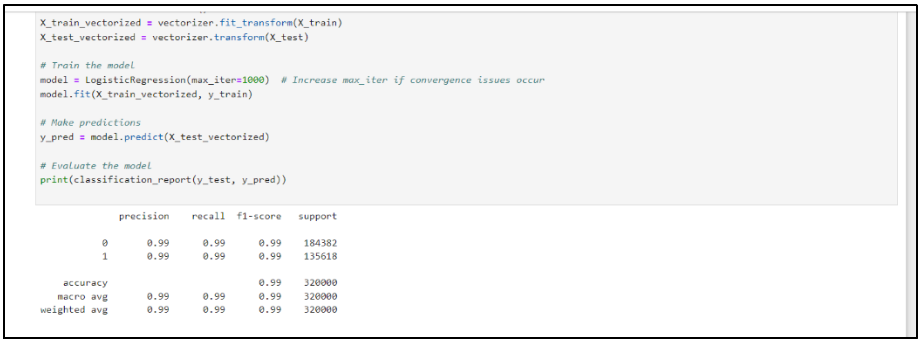
```
   sentiment_label
0                0
1                1
2                1
3                0
4                1
```

```
[14]: print(df.columns)
```

```
Index(['0', '1467810369', 'Mon Apr 06 22:19:45 PDT 2009', 'NO_QUERY',
       '_TheSpecialOne_',
       '@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer.  You shoulda got David Carr of Third Day to do it. ;D',
       'sentiment', 'cleaned_text', 'sentiment_label'],
      dtype='object')
```

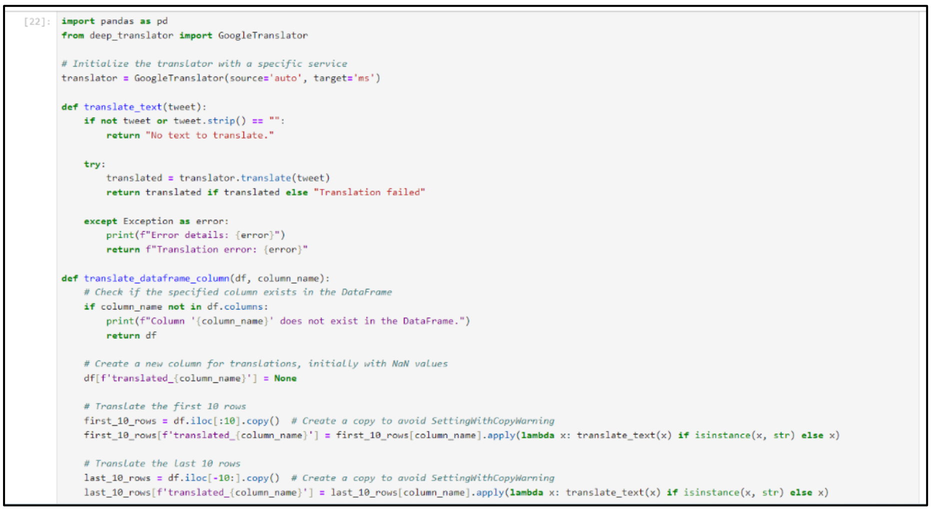
2. More Sophisticated Sentiment Analysis, Objective: Use machine learning models for sentiment classification. Implementation: Train a machine learning model (e.g., Logistic Regression) using labeled data for sentiment classification.

```
[15]: # Import necessary libraries
      from sklearn.model_selection import train_test_split
      from sklearn.feature_extraction.text import CountVectorizer
      from sklearn.linear_model import LogisticRegression
      from sklearn.metrics import classification_report

      # Assuming you have a 'sentiment_label' column for supervised learning
      X = df['cleaned_text']  # Use the cleaned tweet text
      y = df['sentiment_label']  # Ensure this column exists in your DataFrame

      # Split data into training and test sets
      X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

      # Vectorize the text
      vectorizer = CountVectorizer()
      X_train_vectorized = vectorizer.fit_transform(X_train)
```

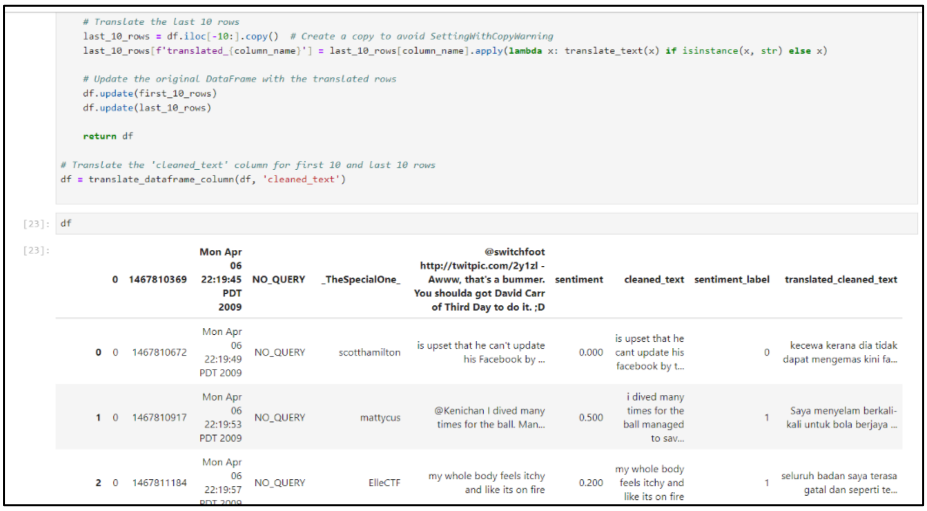


```
X_train_vectorized = vectorizer.fit_transform(X_train)
X_test_vectorized = vectorizer.transform(X_test)

# Train the model
model = LogisticRegression(max_iter=1000)  # Increase max_iter if convergence issues occur
model.fit(X_train_vectorized, y_train)

# Make predictions
y_pred = model.predict(X_test_vectorized)

# Evaluate the model
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.99      0.99      0.99    184382
           1       0.99      0.99      0.99    135618

    accuracy                           0.99    320000
   macro avg       0.99      0.99      0.99    320000
weighted avg       0.99      0.99      0.99    320000
```



```
[22]: import pandas as pd
from deep_translator import GoogleTranslator

# Initialize the translator with a specific service
translator = GoogleTranslator(source='auto', target='ms')

def translate_text(tweet):
    if not tweet or tweet.strip() == "":
        return "No text to translate."

    try:
        translated = translator.translate(tweet)
        return translated if translated else "Translation failed"

    except Exception as error:
        print(f"Error details: {error}")
        return f"Translation error: {error}"

def translate_dataframe_column(df, column_name):
    # Check if the specified column exists in the DataFrame
    if column_name not in df.columns:
        print(f"Column '{column_name}' does not exist in the DataFrame.")
        return df

    # Create a new column for translations, initially with NaN values
    df[f'translated_{column_name}'] = None

    # Translate the first 10 rows
    first_10_rows = df.iloc[:10].copy()  # Create a copy to avoid SettingWithCopyWarning
    first_10_rows[f'translated_{column_name}'] = first_10_rows[column_name].apply(lambda x: translate_text(x) if isinstance(x, str) else x)

    # Translate the last 10 rows
    last_10_rows = df.iloc[-10:].copy()  # Create a copy to avoid SettingWithCopyWarning
    last_10_rows[f'translated_{column_name}'] = last_10_rows[column_name].apply(lambda x: translate_text(x) if isinstance(x, str) else x)
```



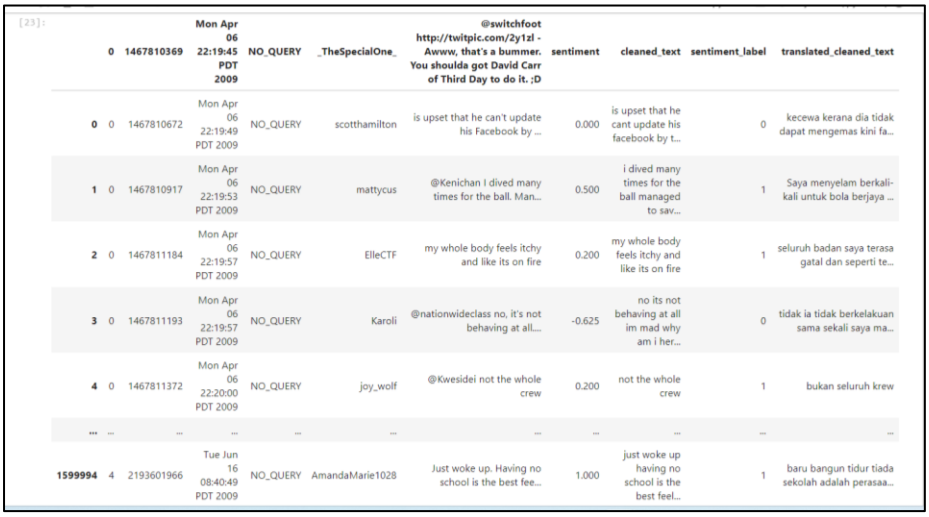
```
    # Translate the last 10 rows
    last_10_rows = df.iloc[-10:].copy()  # Create a copy to avoid SettingWithCopyWarning
    last_10_rows[f'translated_{column_name}'] = last_10_rows[column_name].apply(lambda x: translate_text(x) if isinstance(x, str) else x)
```

```
    # Update the original DataFrame with the translated rows
    df.update(first_10_rows)
    df.update(last_10_rows)

    return df

# Translate the 'cleaned_text' column for first 10 and last 10 rows
df = translate_dataframe_column(df, 'cleaned_text')
```

```
[23]: df
```

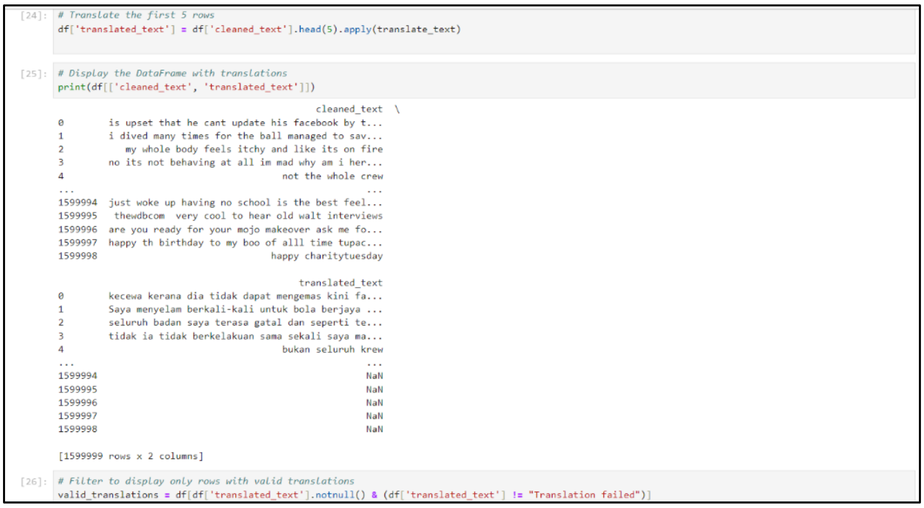
[23]:

	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D	sentiment	cleaned_text	sentiment_label	translated_cleaned_text
0	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...	0.000	is upset that he cant update his facebook by t...	0	kecewa kerana dia tidak dapat mengemas kini fa...
1	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...	0.500	i dived many times for the ball managed to sav...	1	Saya menyelam berkali-kali untuk bola berjaya ...
2	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire	0.200	my whole body feels itchy and like its on fire	1	seluruh badan saya terasa gatal dan seperti te...



[23]:

	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. You shoulda got David Carr of Third Day to do it. ;D	sentiment	cleaned_text	sentiment_label	translated_cleaned_text
0	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...	0.000	is upset that he cant update his facebook by t...	0	kecewa kerana dia tidak dapat mengemas kini fa...
1	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...	0.500	i dived many times for the ball managed to sav...	1	Saya menyelam berkali-kali untuk bola berjaya ...
2	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire	0.200	my whole body feels itchy and like its on fire	1	seluruh badan saya terasa gatal dan seperti te...
3	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....	-0.625	no its not behaving at all im mad why am i her...	0	tidak ia tidak berkelakuan sama sekali saya ma...
4	0	1467811372	Mon Apr 06 22:20:00 PDT 2009	NO_QUERY	joy_wolf	@Kwesidei not the whole crew	0.200	not the whole crew	1	bukan seluruh krew
...	...	...	...	...	...	...	...	...	...	...
1599994	4	2193601966	Tue Jun 16 08:40:49 PDT 2009	NO_QUERY	AmandaMarie1028	Just woke up. Having no school is the best fee...	1.000	just woke up having no school is the best feel...	1	baru bangun tidur tiada sekolah adalah perasaa...

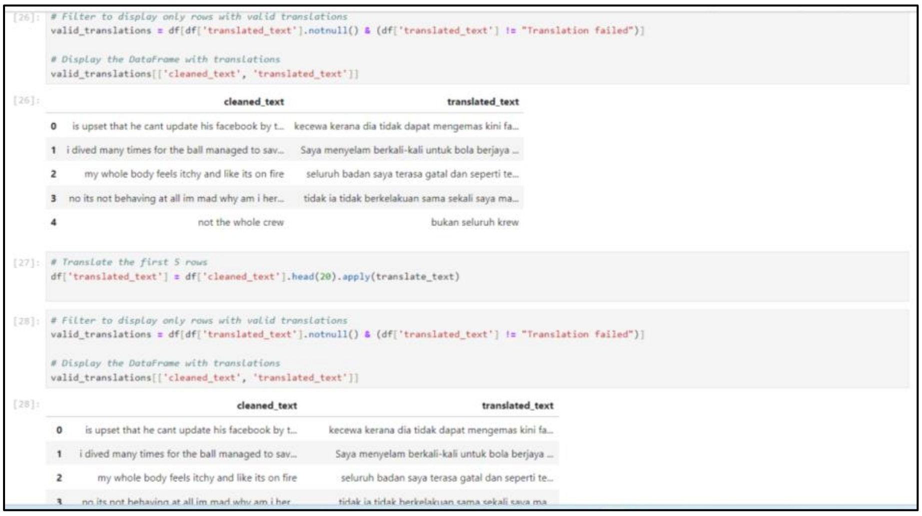


```
[24]: # Translate the first 5 rows
      df['translated_text'] = df['cleaned_text'].head(5).apply(translate_text)
```

```
[25]: # Display the DataFrame with translations
      print(df[['cleaned_text', 'translated_text']])
```

```
                                              cleaned_text  \
0        is upset that he cant update his facebook by t...
1        i dived many times for the ball managed to sav...
2           my whole body feels itchy and like its on fire
3        no its not behaving at all im mad why am i her...
4                                       not the whole crew
...                                                    ...
1599994  just woke up having no school is the best feel...
1599995   thewdbcom  very cool to hear old walt interviews
1599996  are you ready for your mojo makeover ask me fo...
1599997  happy th birthday to my boo of alll time tupac...
1599998                              happy charitytuesday

                                           translated_text
0        kecewa kerana dia tidak dapat mengemas kini fa...
1        Saya menyelam berkali-kali untuk bola berjaya ...
2        seluruh badan saya terasa gatal dan seperti te...
3        tidak ia tidak berkelakuan sama sekali saya ma...
4                                       bukan seluruh krew
...                                                    ...
1599994                                                NaN
1599995                                                NaN
1599996                                                NaN
1599997                                                NaN
1599998                                                NaN

[1599999 rows x 2 columns]
```

```
[26]: # Filter to display only rows with valid translations
      valid_translations = df[df['translated_text'].notnull() & (df['translated_text'] != "Translation failed")]
```



```
[26]: # Filter to display only rows with valid translations
      valid_translations = df[df['translated_text'].notnull() & (df['translated_text'] != "Translation failed")]

      # Display the DataFrame with translations
      valid_translations[['cleaned_text', 'translated_text']]
```

[26]:

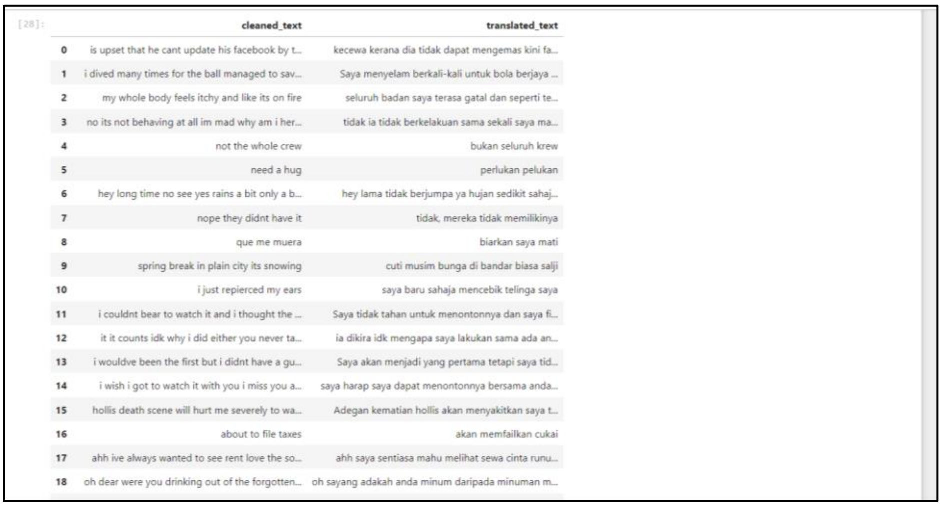
	cleaned_text	translated_text
0	is upset that he cant update his facebook by t...	kecewa kerana dia tidak dapat mengemas kini fa...
1	i dived many times for the ball managed to sav...	Saya menyelam berkali-kali untuk bola berjaya ...
2	my whole body feels itchy and like its on fire	seluruh badan saya terasa gatal dan seperti te...
3	no its not behaving at all im mad why am i her...	tidak ia tidak berkelakuan sama sekali saya ma...
4	not the whole crew	bukan seluruh krew

```
[27]: # Translate the first 5 rows
      df['translated_text'] = df['cleaned_text'].head(20).apply(translate_text)
```

```
[28]: # Filter to display only rows with valid translations
      valid_translations = df[df['translated_text'].notnull() & (df['translated_text'] != "Translation failed")]

      # Display the DataFrame with translations
      valid_translations[['cleaned_text', 'translated_text']]
```

[28]:

	cleaned_text	translated_text
0	is upset that he cant update his facebook by t...	kecewa kerana dia tidak dapat mengemas kini fa...
1	i dived many times for the ball managed to sav...	Saya menyelam berkali-kali untuk bola berjaya ...
2	my whole body feels itchy and like its on fire	seluruh badan saya terasa gatal dan seperti te...
3	no its not behaving at all im mad why am i her...	tidak ia tidak berkelakuan sama sekali saya ma...



[28]:

	cleaned_text	translated_text
0	is upset that he cant update his facebook by t...	kecewa kerana dia tidak dapat mengemas kini fa...
1	i dived many times for the ball managed to sav...	Saya menyelam berkali-kali untuk bola berjaya ...
2	my whole body feels itchy and like its on fire	seluruh badan saya terasa gatal dan seperti te...
3	no its not behaving at all im mad why am i her...	tidak ia tidak berkelakuan sama sekali saya ma...
4	not the whole crew	bukan seluruh krew
5	need a hug	perlukan pelukan
6	hey long time no see yes rains a bit only a b...	hey lama tidak berjumpa ya hujan sedikit sahaj...
7	nope they didnt have it	tidak, mereka tidak memilikinya
8	que me muera	biarkan saya mati
9	spring break in plain city its snowing	cuti musim bunga di bandar biasa salji
10	i just repierced my ears	saya baru sahaja mencebik telinga saya
11	i couldnt bear to watch it and i thought the ...	Saya tidak tahan untuk menontonnya dan saya fi...
12	it it counts idk why i did either you never ta...	ia dikira idk mengapa saya lakukan sama ada an...
13	i wouldve been the first but i didnt have a gu...	Saya akan menjadi yang pertama tetapi saya tid...
14	i wish i got to watch it with you i miss you a...	saya harap saya dapat menontonnya bersama anda...
15	hollis death scene will hurt me severely to wa...	Adegan kematian hollis akan menyakitkan saya t...
16	about to file taxes	akan memfailkan cukai
17	ahh ive always wanted to see rent love the so...	ahh saya sentiasa mahu melihat sewa cinta runu...
18	oh dear were you drinking out of the forgotten...	oh sayang adakah anda minum daripada minuman m...