

# eda\_version2

March 11, 2024

## 1 Predicting tweet sentiment

Dataset from <https://www.kaggle.com/datasets/bhavikjikadara/tweets-dataset>

### Context

This is the sentiment140 dataset. It contains 1,600,000 tweets extracted using the Twitter API. The tweets have been annotated (0 = negative, 4 = positive) and can be used to detect sentiment.

### Content

It contains the following 6 fields:

- target: the polarity of the tweet (0 = negative and 4 = positive)
- ids: The id of the tweet ( 2087)
- date: the date of the tweet (Sat May 16 23:58:44 UTC 2009)
- flag: The query (lyx). If there is no query, then this value is NO\_QUERY.
- user: the user that tweeted.
- text: the text of the tweet.

### 1.1 1. Exploratory Data Analysis

```
[ ]: # import libraries
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split

# text processing libraries
import re
import contractions

from collections import Counter
# import string
import nltk
# import warnings
# %matplotlib inline
# warnings.filterwarnings("ignore")
from nltk.stem.porter import PorterStemmer
```

```
from wordcloud import WordCloud
```

Pandas and Numpy have been used for data manipulation and numerical calculations  
Matplotlib and Seaborn have been used for data visualizations

```
[ ]: # import data
tweets = pd.read_csv("../data//tweets.csv", encoding="latin-1")
```

```
[ ]: # check the shape of the dataframe
print("Shape of the dataframe:", tweets.shape)
```

Shape of the dataframe: (1048575, 6)

```
[ ]: # display the first few rows of the dataframe
tweets.head()
```

```
[ ]: Target      ID      Date      flag      User \
0      0  1467810672  Mon Apr 06 22:19:49 PDT 2009  NO_QUERY  scotthamilton
1      0  1467810917  Mon Apr 06 22:19:53 PDT 2009  NO_QUERY      mattycus
2      0  1467811184  Mon Apr 06 22:19:57 PDT 2009  NO_QUERY      ElleCTF
3      0  1467811193  Mon Apr 06 22:19:57 PDT 2009  NO_QUERY      Karoli
4      0  1467811372  Mon Apr 06 22:20:00 PDT 2009  NO_QUERY      joy_wolf
```

Text

```
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
```

```
[ ]: # display the last few rows of the dataframe
tweets.tail()
```

```
[ ]: Target      ID      Date      flag \
1048570      4  1960186342  Fri May 29 07:33:44 PDT 2009  NO_QUERY
1048571      4  1960186409  Fri May 29 07:33:43 PDT 2009  NO_QUERY
1048572      4  1960186429  Fri May 29 07:33:44 PDT 2009  NO_QUERY
1048573      4  1960186445  Fri May 29 07:33:44 PDT 2009  NO_QUERY
1048574      4  1960186607  Fri May 29 07:33:45 PDT 2009  NO_QUERY
```

User

Text

```
1048570  Madelinedugganx  My GrandMa is making Dinenr with my Mum
1048571  OffRoad_Dude  Mid-morning snack time... A bowl of cheese noo...
1048572  Falchion  @ShaDeLa same here say it like from the Termi...
1048573  jonasobsessedx  @DestinyHope92 im great thaanks wbuu?
1048574  sugababez  cant wait til her date this weekend
```

```
[ ]: # display information about data
tweets.info()
```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1048575 entries, 0 to 1048574
Data columns (total 6 columns):
 #   Column  Non-Null Count  Dtype
---  -
 0   Target  1048575 non-null  int64
 1   ID       1048575 non-null  int64
 2   Date     1048575 non-null  object
 3   flag     1048575 non-null  object
 4   User     1048575 non-null  object
 5   Text     1048575 non-null  object
dtypes: int64(2), object(4)
memory usage: 48.0+ MB

```

```

[ ]: # check for duplication
     tweets.nunique()

```

```

[ ]: Target      2
     ID      1048044
     Date      662453
     flag        1
     User     511366
     Text     1036135
     dtype: int64

```

```

[ ]: # check for missing values
     tweets.isnull().sum()

```

```

[ ]: Target      0
     ID          0
     Date        0
     flag        0
     User        0
     Text        0
     dtype: int64

```

```

[ ]: # summary statistics of numerical columns
     tweets.describe()

```

```

[ ]:
count      Target      ID
mean      1.048575e+06  1.976166e+09
std        1.701120e+00  2.300580e+08
min         0.000000e+00  1.467811e+09
25%         0.000000e+00  1.824526e+09
50%         0.000000e+00  1.990869e+09
75%         0.000000e+00  2.198903e+09
max         4.000000e+00  2.329206e+09

```

### Data reduction

Some columns or variables can be dropped if they do not add value to our analysis

In our dataset, columns ID, Date, flag, User don't have any predictive power to predict the dependent variable

```
[ ]: data = tweets.drop(['ID', 'Date', 'flag', 'User'], axis = 'columns')
data
```

```
[ ]:      Target      Text
0         0  is upset that he can't update his Facebook by ...
1         0  @Kenichan I dived many times for the ball. Man...
2         0    my whole body feels itchy and like its on fire
3         0  @nationwideclass no, it's not behaving at all...
4         0                @Kwesidei not the whole crew
...      ...
1048570    4      My GrandMa is making Dinenr with my Mum
1048571    4  Mid-morning snack time... A bowl of cheese noo...
1048572    4  @ShaDeLa same here say it like from the Termi...
1048573    4                @DestinyHope92 im great thaanks  wbuu?
1048574    4      cant wait til her date this weekend
```

[1048575 rows x 2 columns]

### Data cleaning

Some names of the variables are not relevant and not easy to understand

Some data may have data entry errors, and some variables may need data type conversion.

We need to fix this issue in the data

```
[ ]: # adjusting target values
data['Target'] = data['Target'].replace(4, 1)
data
```

```
[ ]:      Target      Text
0         0  is upset that he can't update his Facebook by ...
1         0  @Kenichan I dived many times for the ball. Man...
2         0    my whole body feels itchy and like its on fire
3         0  @nationwideclass no, it's not behaving at all...
4         0                @Kwesidei not the whole crew
...      ...
1048570    1      My GrandMa is making Dinenr with my Mum
1048571    1  Mid-morning snack time... A bowl of cheese noo...
1048572    1  @ShaDeLa same here say it like from the Termi...
1048573    1                @DestinyHope92 im great thaanks  wbuu?
1048574    1      cant wait til her date this weekend
```

[1048575 rows x 2 columns]

```
[ ]: # removing unnecessary user tags
data['Text'] = data['Text'].replace(r"@\\w+", "", regex=True)
data
```

```
[ ]:      Target      Text
0         0  is upset that he can't update his Facebook by ...
```

```

1          0   I dived many times for the ball. Managed to s...
2          0   my whole body feels itchy and like its on fire
3          0   no, it's not behaving at all. i'm mad. why am...
4          0   not the whole crew
...
1048570     1   My GrandMa is making Dinenr with my Mum
1048571     1   Mid-morning snack time... A bowl of cheese noo...
1048572     1   same here say it like from the Terminator m...
1048573     1   im great thaanks wbuu?
1048574     1   cant wait til her date this weekend

```

[1048575 rows x 2 columns]

```

[ ]: # resolving contractions (and slang)
data['Text'] = data['Text'].apply(lambda x: contractions.fix(x))
data

```

```

[ ]:      Target      Text
0          0   is upset that he cannot update his Facebook by...
1          0   I dived many times for the ball. Managed to s...
2          0   my whole body feels itchy and like its on fire
3          0   no, it is not behaving at all. i am mad. why ...
4          0   not the whole crew
...
1048570     1   My GrandMa is making Dinenr with my Mum
1048571     1   Mid-morning snack time... A bowl of cheese noo...
1048572     1   same here say it like from the Terminator m...
1048573     1   i am great thaanks wbuu?
1048574     1   cannot wait til her date this weekend

```

[1048575 rows x 2 columns]

```

[ ]: # removing punctuation marks
data['Text'] = data['Text'].apply(lambda x: re.sub(r'[\w\s]', '', x))
data

```

```

[ ]:      Target      Text
0          0   is upset that he cannot update his Facebook by...
1          0   I dived many times for the ball Managed to sa...
2          0   my whole body feels itchy and like its on fire
3          0   no it is not behaving at all i am mad why am ...
4          0   not the whole crew
...
1048570     1   My GrandMa is making Dinenr with my Mum
1048571     1   Midmorning snack time A bowl of cheese noodles...
1048572     1   same here say it like from the Terminator m...
1048573     1   i am great thaanks wbuu
1048574     1   cannot wait til her date this weekend

```

[1048575 rows x 2 columns]

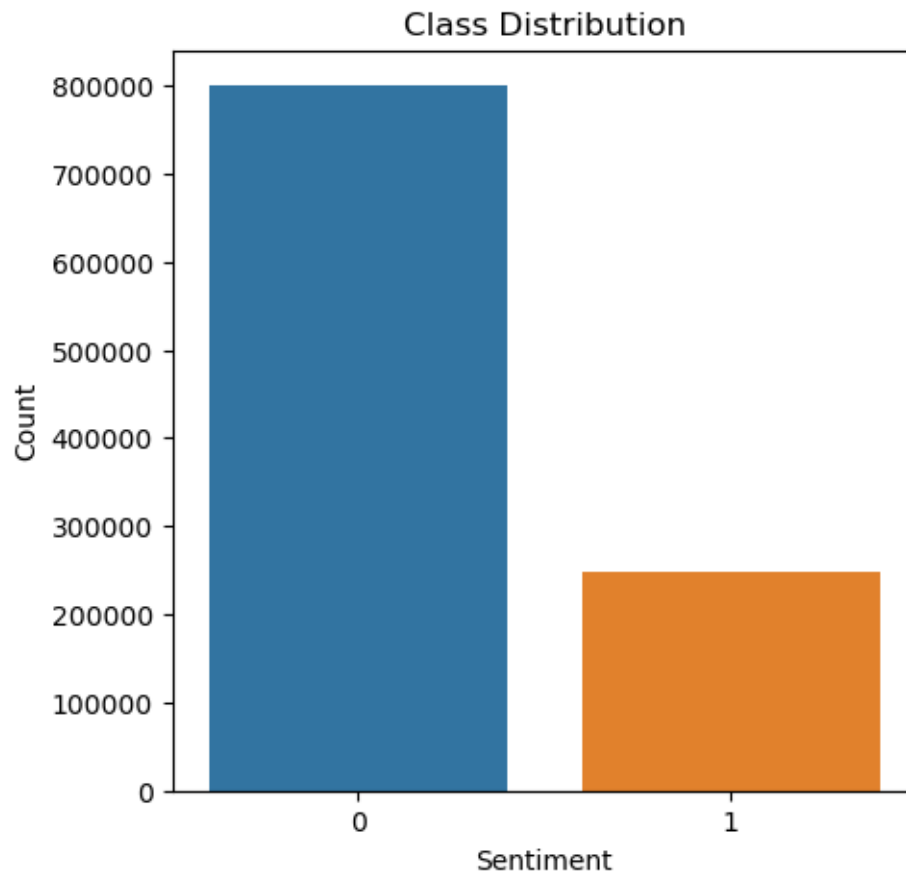
```
[ ]: # lowercasing letters in the text
data['Text'] = data['Text'].str.lower()
data
```

```
[ ]:
      Target      Text
0          0  is upset that he cannot update his facebook by...
1          0  i dived many times for the ball managed to sa...
2          0   my whole body feels itchy and like its on fire
3          0  no it is not behaving at all i am mad why am ...
4          0                                not the whole crew
...      ...
1048570      1          my grandma is making dinenr with my mum
1048571      1  midmorning snack time a bowl of cheese noodles...
1048572      1  same here say it like from the terminator m...
1048573      1                                i am great thaanks wbuu
1048574      1          cannot wait til her date this weekend
```

[1048575 rows x 2 columns]

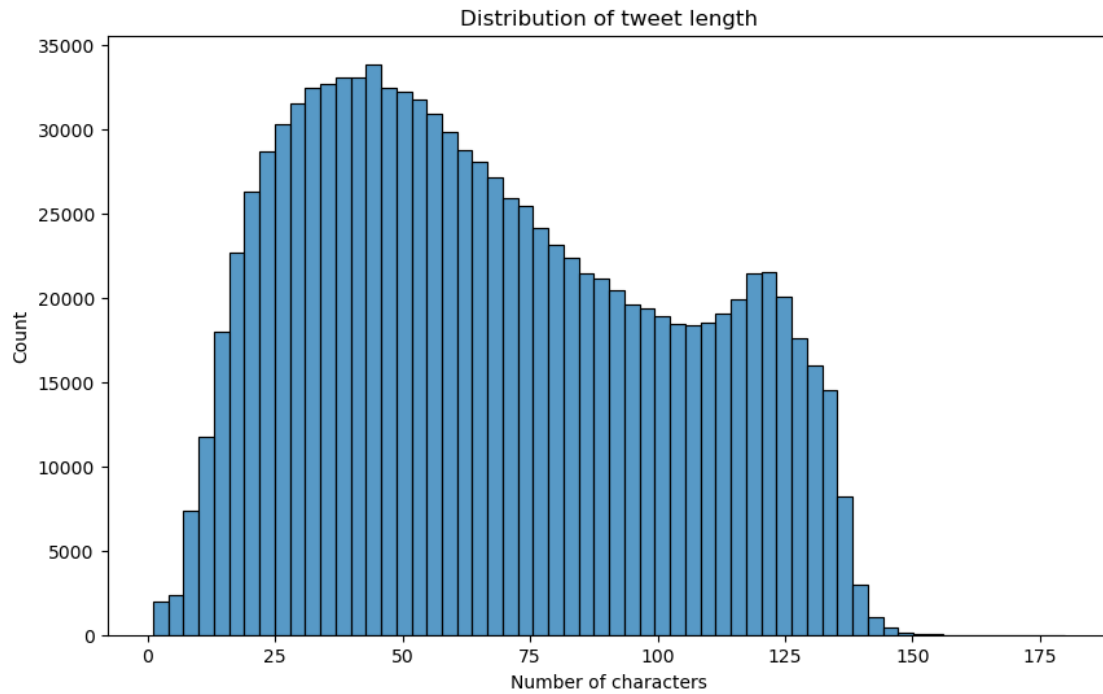
### Visualization

```
[ ]: # visualize class distribution
plt.figure(figsize=(5, 5))
sns.countplot(x = 'Target' , data = data)
plt.title('Class Distribution')
plt.xlabel('Sentiment')
plt.ylabel('Count')
plt.show()
```



```
[ ]: # explore tweet length
data['characters'] = data['Text'].apply(lambda x: len(x))

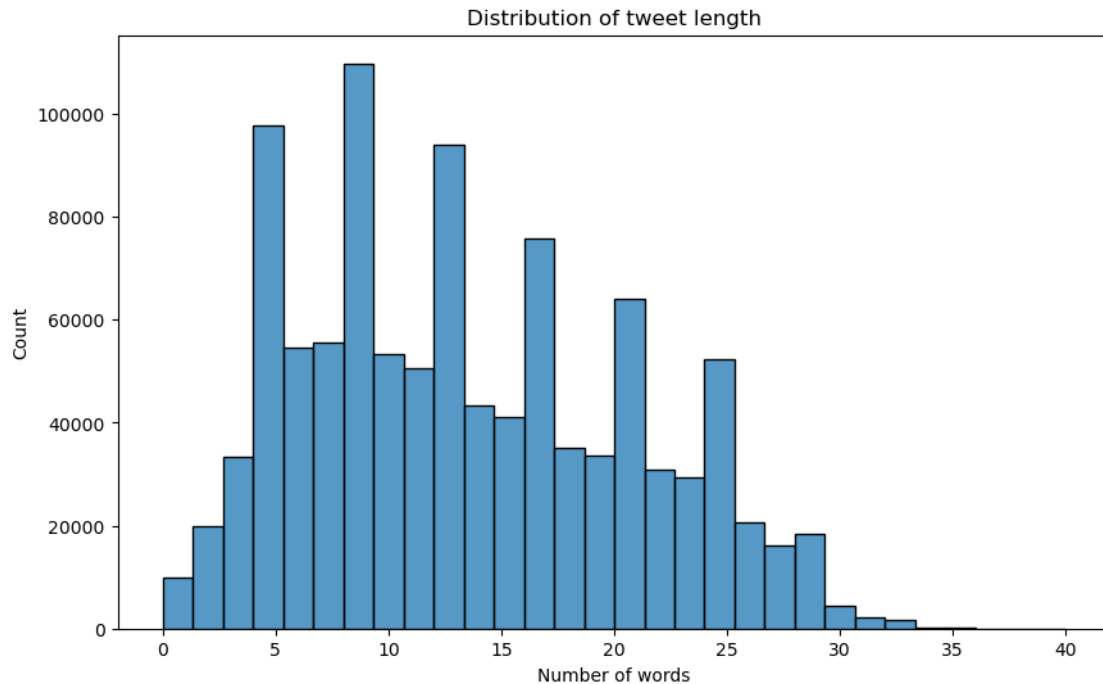
# visualize tweet length distribution
plt.figure(figsize = (10, 6))
sns.histplot(data['characters'], bins = 60)
plt.title('Distribution of tweet length')
plt.xlabel('Number of characters')
plt.ylabel('Count')
plt.show()
```



```
[ ]: # explore tweet length
data['words'] = data['Text'].apply(lambda x: len(x.split()))

# visualize tweet length distribution
plt.figure(figsize = (10, 6))
sns.histplot(data['words'], bins = 30)
plt.title('Distribution of tweet length')
plt.xlabel('Number of words')
plt.ylabel('Count')
plt.show()
```





```
[ ]: # combine all the text into a single string
all_text = ' '.join(data['Text'])

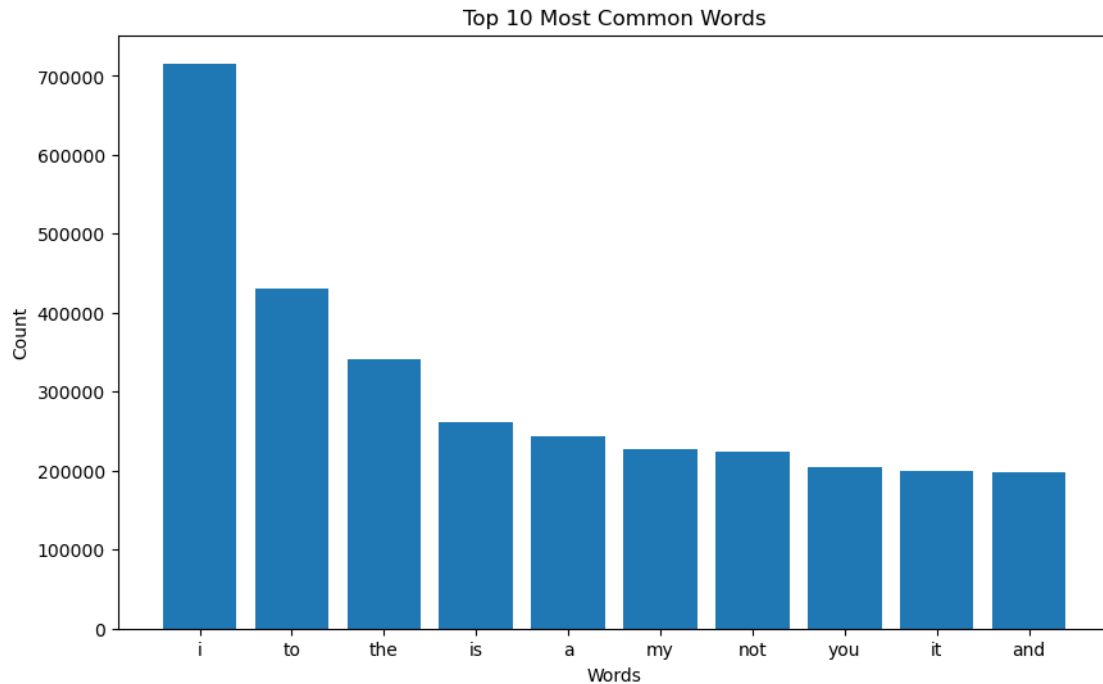
# split the text into individual words
words = all_text.split()

# count the frequency of each word
word_counts = Counter(words)

# get the top 10 most common words
top_10_words = word_counts.most_common(10)

# extract the words and their counts
top_10_words, top_10_counts = zip(*top_10_words)

# plot the bar chart
plt.figure(figsize=(10, 6))
plt.bar(top_10_words, top_10_counts)
plt.title('Top 10 Most Common Words')
plt.xlabel('Words')
plt.ylabel('Count')
plt.show()
```

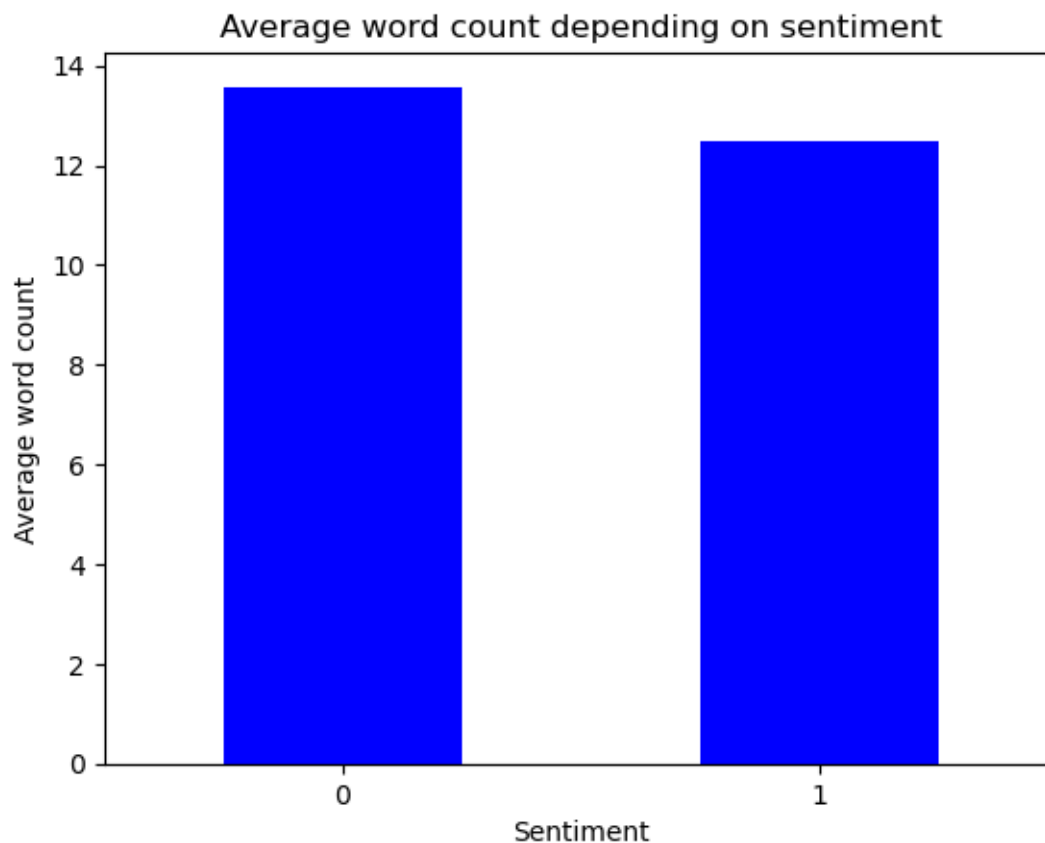


```
[ ]: # average word count depending on sentiment

df = data.groupby("Target").words.agg("mean")

df.plot(kind = 'bar', color = 'blue')

plt.title('Average word count depending on sentiment')
plt.xlabel('Sentiment')
plt.ylabel('Average word count')
plt.xticks(rotation = 0)
plt.show()
```



```
[ ]: # is # in tweet?
```

```
data['has_hashtag'] = tweets['Text'].str.contains(r'#\w+')
data
```

```
[ ]:
      Target      Text \
0      0  is upset that he cannot update his facebook by...
1      0  i dived many times for the ball managed to sa...
2      0  my whole body feels itchy and like its on fire
3      0  no it is not behaving at all i am mad why am ...
4      0  not the whole crew
...      ...
1048570  1  my grandma is making dinenr with my mum
1048571  1  midmorning snack time a bowl of cheese noodles...
1048572  1  same here say it like from the terminator m...
1048573  1  i am great thaanks wbuu
1048574  1  cannot wait til her date this weekend

      characters  words  has_hashtag
0      107      21      False
1      78      17      False
```

2	47	10	False
3	93	22	False
4	20	4	False
...	...	...	...
1048570	40	8	False
1048571	51	9	False
1048572	81	16	False
1048573	25	5	False
1048574	38	7	False

[1048575 rows x 5 columns]

```
[ ]: # is hashtag present in negatives tweets?
```

```
data[data['Target'] == 0]['has_hashtag'].value_counts().apply(lambda x: x /
↳ len(data[data['Target'] == 0]) * 100)
```

```
[ ]: False    98.098373
      True     1.901627
      Name: has_hashtag, dtype: float64
```

```
[ ]: # is hashtag present in positives tweets?
```

```
data[data['Target'] == 1]['has_hashtag'].value_counts().apply(lambda x: x /
↳ len(data[data['Target'] == 1]) * 100)
```

```
[ ]: False    97.43861
      True     2.56139
      Name: has_hashtag, dtype: float64
```

```
[ ]: # is "not" in tweet?
```

```
data['has_not'] = data['Text'].str.contains('not')
data
```

```
[ ]:      Target      Text \
0          0  is upset that he cannot update his facebook by...
1          0  i dived many times for the ball managed to sa...
2          0  my whole body feels itchy and like its on fire
3          0  no it is not behaving at all i am mad why am ...
4          0                not the whole crew
...      ...
1048570     1      my grandma is making dinenr with my mum
1048571     1  midmorning snack time a bowl of cheese noodles...
1048572     1  same here say it like from the terminator m...
1048573     1                i am great thaanks wbuu
1048574     1      cannot wait til her date this weekend

      characters  words  has_hashtag  has_not
0              107    21         False     True
```

1	78	17	False	False
2	47	10	False	False
3	93	22	False	True
4	20	4	False	True
...	...	...	...	...
1048570	40	8	False	False
1048571	51	9	False	False
1048572	81	16	False	False
1048573	25	5	False	False
1048574	38	7	False	True

[1048575 rows x 6 columns]

```
[ ]: # is "not" present in negatives tweets?
```

```
data[data['Target'] == 0]['has_not'].value_counts().apply(lambda x: x /
↳ len(data[data['Target'] == 0]) * 100)
```

```
[ ]: False    70.632338
      True     29.367662
      Name: has_not, dtype: float64
```

```
[ ]: # is "not" present in positives tweets?
```

```
data[data['Target'] == 1]['has_not'].value_counts().apply(lambda x: x /
↳ len(data[data['Target'] == 1]) * 100)
```

```
[ ]: False    86.185714
      True     13.814286
      Name: has_not, dtype: float64
```

```
[ ]: # extract hour from the Date column
```

```
data['Hour'] = pd.to_datetime(tweets['Date']).dt.hour
data
```

```
c:\Users\flang\anaconda3\lib\site-packages\dateutil\parser\_parser.py:1207:
UnknownTimezoneWarning: tzname PDT identified but not understood. Pass
`tzinfos` argument in order to correctly return a timezone-aware datetime. In a
future version, this will raise an exception.
  warnings.warn("tzname {tzname} identified but not understood. "
```

```
[ ]: Target      Text \
0      0  is upset that he cannot update his facebook by...
1      0  i dived many times for the ball managed to sa...
2      0  my whole body feels itchy and like its on fire
3      0  no it is not behaving at all i am mad why am ...
4      0  not the whole crew
...      ...
1048570  1  my grandma is making dinenr with my mum
```

```

1048571      1  midmorning snack time a bowl of cheese noodles...
1048572      1   same here  say it like from the terminator m...
1048573      1                                     i am great thaanks  wbuu
1048574      1                                cannot wait til her date this weekend

```

	characters	words	has_hashtag	has_not	Hour
0	107	21	False	True	22
1	78	17	False	False	22
2	47	10	False	False	22
3	93	22	False	True	22
4	20	4	False	True	22
...	...	...	...	...	...
1048570	40	8	False	False	7
1048571	51	9	False	False	7
1048572	81	16	False	False	7
1048573	25	5	False	False	7
1048574	38	7	False	True	7

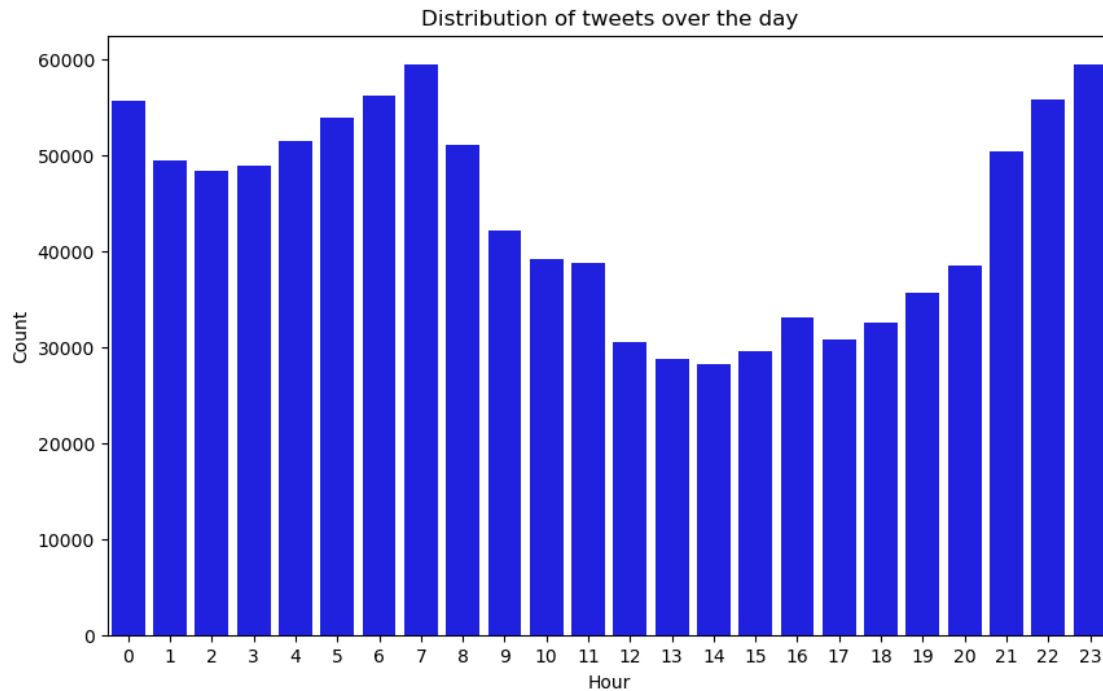
[1048575 rows x 7 columns]

```

[:]: # visualize the distribution of tweets over the day

plt.figure(figsize=(10, 6))
sns.countplot(x = 'Hour', data = data, color = 'blue')
plt.title('Distribution of tweets over the day')
plt.xlabel('Hour')
plt.ylabel('Count')
plt.show()

```

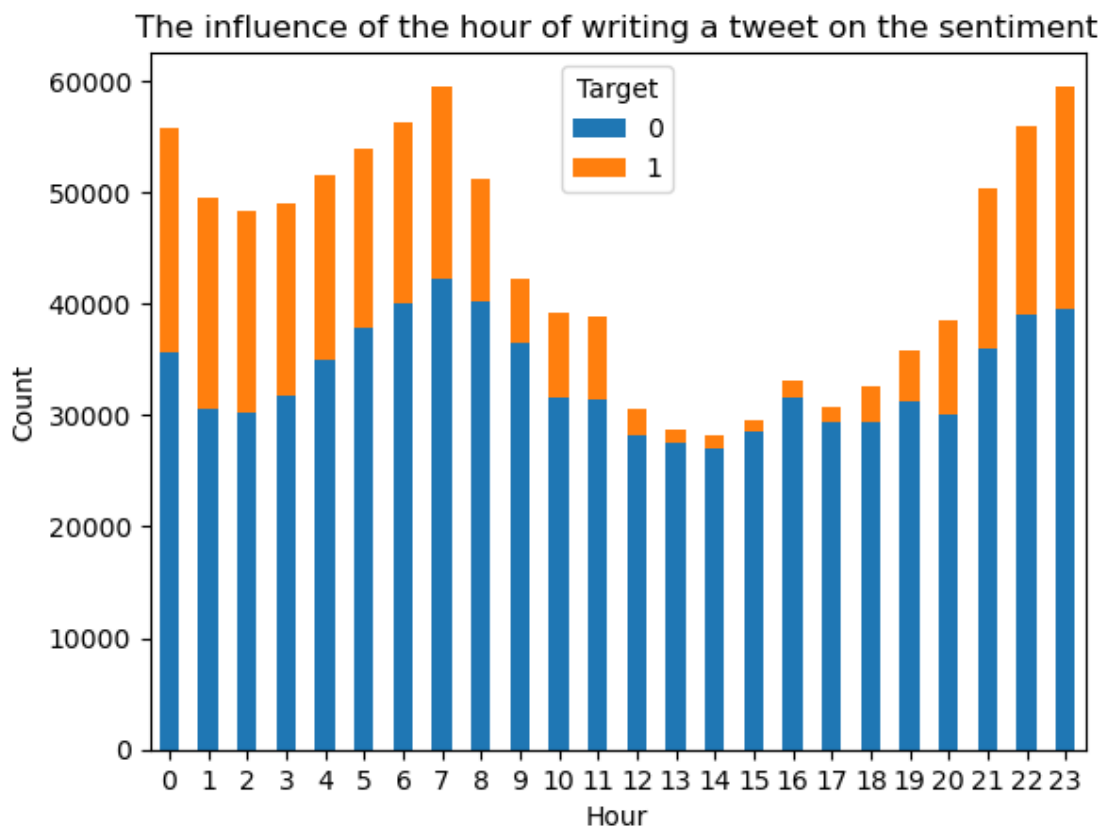


```
[ ]: # visualize the the influence of the hour of writing a tweet on the Target
      ↳variable
hourly_target_counts = data.groupby('Hour')['Target'].value_counts().
      ↳unstack(fill_value=0)
plt.figure(figsize=(15, 6))
hourly_target_counts.plot(kind='bar', stacked=True)

plt.title('The influence of the hour of writing a tweet on the sentiment')
plt.xlabel('Hour')
plt.ylabel('Count')
plt.xticks(rotation = 0)

plt.show()
```

<Figure size 1500x600 with 0 Axes>



```
[ ]: # deleting words which have less characters than 3

data['clean_text'] = data["Text"].apply(lambda x: " ".join([w for w in x.
    ↳split() if len(w)>=3]))
data
```

```
[ ]: Target      Text \
0      0  is upset that he cannot update his facebook by...
1      0  i dived many times for the ball managed to sa...
2      0    my whole body feels itchy and like its on fire
3      0  no it is not behaving at all i am mad why am ...
4      0                                not the whole crew
...    ...
1048570 1      my grandma is making dinenr with my mum
1048571 1  midmorning snack time a bowl of cheese noodles...
1048572 1  same here say it like from the terminator m...
1048573 1                                i am great thaanks wbuu
1048574 1      cannot wait til her date this weekend
```

```
characters  words  has_hashtag  has_not  Hour \
0          107    21          False    True    22
```



1	78	17	False	False	22
2	47	10	False	False	22
3	93	22	False	True	22
4	20	4	False	True	22
...	...	...	...	...	...
1048570	40	8	False	False	7
1048571	51	9	False	False	7
1048572	81	16	False	False	7
1048573	25	5	False	False	7
1048574	38	7	False	True	7

```

                                clean_text
0      upset that cannot update his facebook texting ...
1      dived many times for the ball managed save the...
2              whole body feels itchy and like its fire
3      not behaving all mad why here because cannot s...
4                                not the whole crew
...
1048570                                grandma making dinenr with mum
1048571      midmorning snack time bowl cheese noodles yum
1048572 same here say like from the terminator movies...
1048573                                great thaanks wbuu
1048574      cannot wait til her date this weekend

```

[1048575 rows x 8 columns]

```
[ ]: # individual words considered as tokens
```

```

tokenized_tweet = data['clean_text'].apply(lambda x: x.split())
tokenized_tweet

```

```

[ ]: 0      [upset, that, cannot, update, his, facebook, t...
1      [dived, many, times, for, the, ball, managed, ...
2      [whole, body, feels, itchy, and, like, its, fire]
3      [not, behaving, all, mad, why, here, because, ...
4      [not, the, whole, crew]

...

1048570      [grandma, making, dinenr, with, mum]
1048571      [midmorning, snack, time, bowl, cheese, noodle...
1048572      [same, here, say, like, from, the, terminator...
1048573      [great, thaanks, wbuu]
1048574      [cannot, wait, til, her, date, this, weekend]
Name: clean_text, Length: 1048575, dtype: object

```

```
[ ]: # stem the words
```

```
stemmer = PorterStemmer()
```

```
tokenized_tweet = tokenized_tweet.apply(lambda s: [stemmer.stem(word) for word_
→in s]) # stemming
tokenized_tweet
```

```
[]: 0      [upset, that, cannot, updat, hi, facebook, tex...
1      [dive, mani, time, for, the, ball, manag, save...
2      [whole, bodi, feel, itchi, and, like, it, fire]
3      [not, behav, all, mad, whi, here, becaus, cann...
4      [not, the, whole, crew]
```

```
...
1048570      [grandma, make, dinenr, with, mum]
1048571      [midmorn, snack, time, bowl, chees, noodl, yum]
1048572      [same, here, say, like, from, the, termini, mo...
1048573      [great, thaank, wbuu]
1048574      [cannot, wait, til, her, date, thi, weekend]
Name: clean_text, Length: 1048575, dtype: object
```

```
[]: # combining to sentences
for i in range(len(tokenized_tweet)):
    tokenized_tweet[i] = ' '.join(tokenized_tweet[i])
data['clean_text'] = tokenized_tweet
data
```

```
[]: Target      Text \
0      0  is upset that he cannot update his facebook by...
1      0  i dived many times for the ball managed to sa...
2      0  my whole body feels itchy and like its on fire
3      0  no it is not behaving at all i am mad why am ...
4      0  not the whole crew
```

```
...
1048570      1  my grandma is making dinenr with my mum
1048571      1  midmorning snack time a bowl of cheese noodles...
1048572      1  same here say it like from the terminator m...
1048573      1  i am great thaanks wbuu
1048574      1  cannot wait til her date this weekend
```

```
characters  words  has_hashtag  has_not  Hour \
0           107    21         False    True   22
1           78     17         False    False  22
2           47     10         False    False  22
3           93     22         False    True   22
4           20      4         False    True   22
...
1048570      40      8         False    False   7
1048571      51      9         False    False   7
1048572      81     16         False    False   7
1048573      25      5         False    False   7
1048574      38      7         False    True    7
```

```

                                clean_text
0      upset that cannot updat hi facebook text and m...
1      dive mani time for the ball manag save the res...
2      whole bodi feel itchi and like it fire
3      not behav all mad whi here becaus cannot see y...
4      not the whole crew
...
1048570      grandma make dinenr with mum
1048571      midmorn snack time bowl chees noodl yum
1048572 same here say like from the termini movi come ...
1048573      great thaank wbuu
1048574      cannot wait til her date thi weekend

```

```
[1048575 rows x 8 columns]
```

```

[ ]: all_words = ' '.join([text for text in data['clean_text']])
all_words_pos = ' '.join([text for text in data['clean_text'][data['Target'] == 1]])
all_words_neg = ' '.join([text for text in data['clean_text'][data['Target'] == 0]])
wordcloud = WordCloud(width=800, height=500, random_state=42,
    max_font_size=100).generate(all_words)
wordcloud_pos = WordCloud(width=800, height=500, random_state=42,
    max_font_size=100).generate(all_words_pos)
wordcloud_neg = WordCloud(width=800, height=500, random_state=42,
    max_font_size=100).generate(all_words_neg)

# plot the graph

fig, ax = plt.subplots(1, 3, figsize=(15, 10))
ax[0].imshow(wordcloud, interpolation="bilinear")
ax[0].set_title('All words')
ax[0].axis('off')
ax[1].imshow(wordcloud_pos, interpolation="bilinear")
ax[1].set_title('Words target 1')
ax[1].axis('off')
ax[2].imshow(wordcloud_neg, interpolation="bilinear")
ax[2].set_title('Words target 0')
ax[2].axis('off')
fig.show()

```

C:\Users\flang\AppData\Local\Temp\ipykernel\_16508\2180637428.py:20: UserWarning: FigureCanvasAgg is non-interactive, and thus cannot be shown

```
fig.show()
```

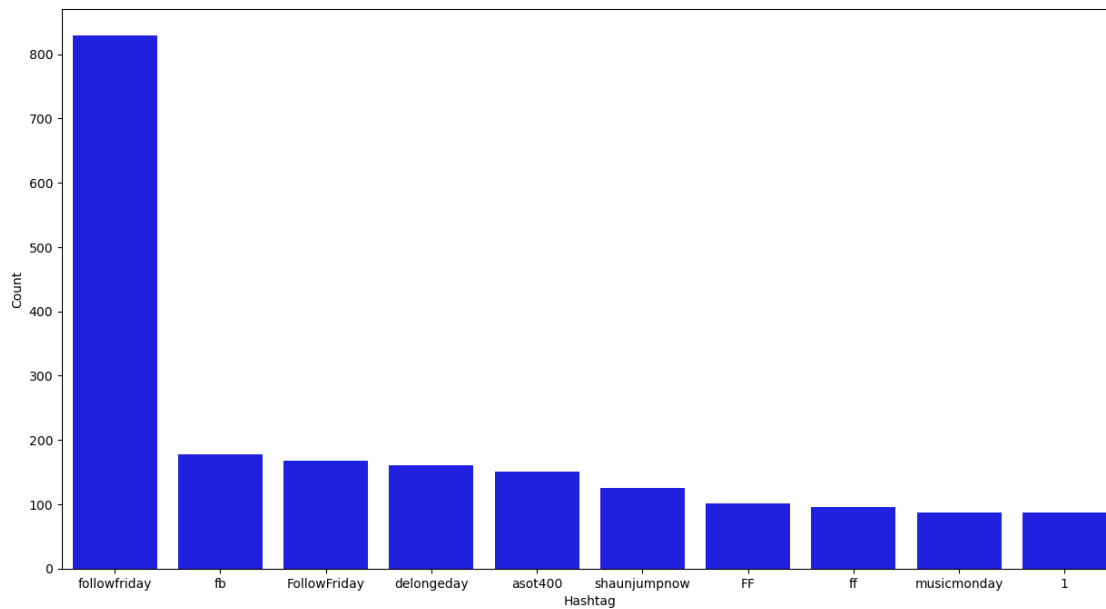


```
1274      com125      1
3217    brownies      1
```

```
[3218 rows x 2 columns]
```

```
[ ]: # selecting top 10 most frequent hashtags positive
d = d.nlargest(columns="Count", n = 10)
plt.figure(figsize=(15,8))
sns.barplot(data=d, x= "Hashtag", y = "Count", color="blue")
```

```
[ ]: <Axes: xlabel='Hashtag', ylabel='Count'>
```



```
[ ]: # converting dictionary to dataframe
freq = nltk.FreqDist(ht_negative)
d = pd.DataFrame({'Hashtag': list(freq.keys()),
                  'Count': list(freq.values())
                  })
d.sort_values(by='Count', ascending=False)
```

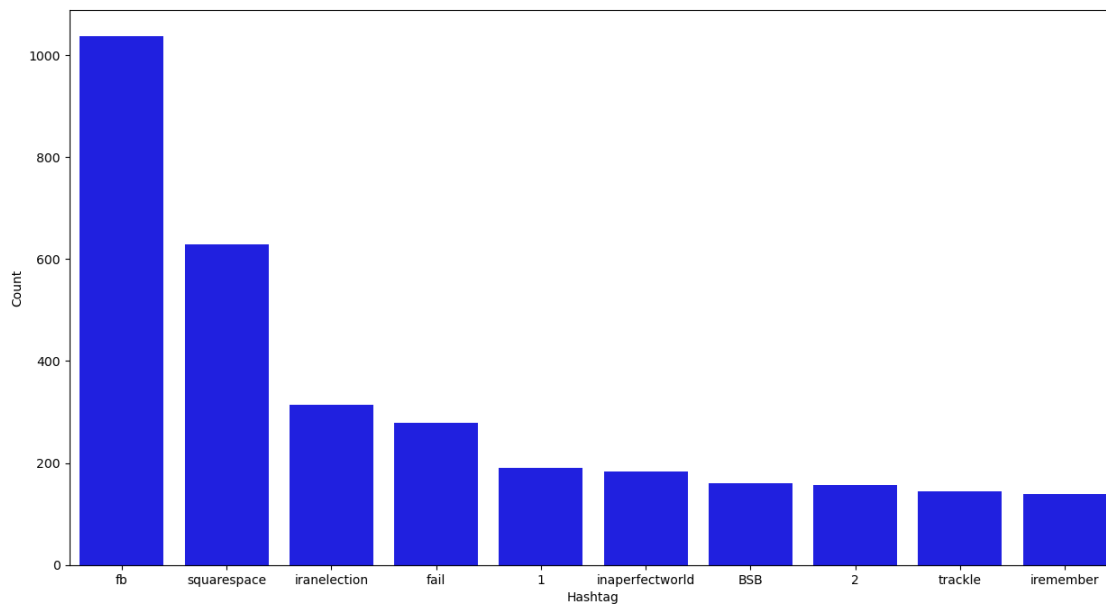
```
[ ]:      Hashtag  Count
2      fb      1037
4246  squarespace  628
4333  iranelection  315
12     fail      278
17     1         191
...     ...     ...
2919   NHL10      1
2918   NBA10      1
2917   Fifa10      1
```

```
2916      WSOP4      1
7089      lbda      1
```

```
[7090 rows x 2 columns]
```

```
[ ]: # selecting top 10 most frequent hashtags negative
d = d.nlargest(columns="Count", n = 10)
plt.figure(figsize=(15,8))
sns.barplot(data=d, x= "Hashtag", y = "Count", color="blue")
```

```
[ ]: <Axes: xlabel='Hashtag', ylabel='Count'>
```



## 1.2 2. Feature engineering

```
[ ]:
```

## 1.3 3. Preparation of training and test sets

```
[ ]: clean_data = data[['Text', 'Target']]
clean_data
```

```
[ ]:
```

	Text	Target
0	is upset that he cannot update his Facebook by...	0
1	I dived many times for the ball. Managed to s...	0
2	my whole body feels itchy and like its on fire	0
3	no, it is not behaving at all. i am mad. why ...	0
4	not the whole crew	0

```

...
1048570          My GrandMa is making Dinenr with my Mum          1
1048571 Mid-morning snack time... A bowl of cheese noo...        1
1048572 same here say it like from the Terminator m...           1
1048573                                i am great thaanks wbuu?      1
1048574          cannot wait til her date this weekend             1

```

```
[1048575 rows x 2 columns]
```

```
[ ]: x_train, x_test, y_train, y_test = train_test_split(
      clean_data[clean_data.columns.values[:-1]], # X
      clean_data[clean_data.columns.values[-1]], # y
      test_size=0.33, random_state=42)
```

```
[ ]: x_train.shape, x_test.shape, y_train.shape, y_test.shape
```

```
[ ]: ((702545, 1), (346030, 1), (702545,), (346030,))
```

```
[ ]:
```

## 1.4 4. Modeling and validation

```
[ ]:
```