Twitter Sentimental Analysis Report

Data Set Description

- id : The id associated with the tweets in the given dataset.
- tweets : The tweets collected from various sources and having either positive or negative sentiments associated with it.
- label : A tweet with **label '0'** is of **positive sentiment** while a tweet with **label '1'** is of **negative sentiment.**

Importing the necessary packages

Reading the train.csv Pandas file

- In the first line we read the train.csv file using Pandas.
- In the second line as a safe backup we keep a copy of our original train.csv file. **We make a copy of train data so that even if we have to make any changes in this dataset we would not lose the original dataset.**

Overview of the training dataset

	id	label	tweet
0	1	0	@user when a father is dysfunctional and is s...
1	2	0	@user @user thanks for #lyft credit i can't us...
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in ...
4	5	0	factsguide: society now #motivation
5	6	0	[2/2] huge fan fare and big talking before the...
6	7	0	@user camping tomorrow @user @user @user @use...
7	8	0	the next school year is the year for exams.ð□□...
8	9	0	we won!!! love the land!!! #allin #cavs #champ...
9	10	0	@user @user welcome here ! i'm it's so #gr...
10	11	0	â□□ #ireland consumer price index (mom) climb...
11	12	0	we are so selfish. #orlando #standwithorlando ...
12	13	0	i get to see my daddy today!! #80days #getti...
13	14	1	@user #cnn calls #michigan middle school 'buil...
14	15	1	no comment! in #australia #opkillingbay #se...
15	16	0	ouch...junior is angryð□□□#got7 #junior #yugyo...
16	17	0	i am thankful for having a paner. #thankful #p...
17	18	1	retweet if you agree!
18	19	0	its #friday! ð□□□ smiles all around via ig use...
19	20	0	as we all know, essential oils are not made of...
20	21	0	#euro2016 people blaming ha for conceded goal ...

31940	31941	0	@user she's finally here! @user
31941	31942	0	passed first year of uni #yay #love #pass #uni...
31942	31943	0	this week is flying by #humpday - #wednesday...
31943	31944	0	@user modeling photoshoot this friday yay #mo...
31944	31945	0	you're surrounded by people who love you (even...
31945	31946	0	feel like... ð□□□ð□□¶ð□□□ #dog #summer #hot #h...
31946	31947	1	@user omfg i'm offended! i'm a mailbox and i'...
31947	31948	1	@user @user you don't have the balls to hashta...
31948	31949	1	makes you ask yourself, who am i? then am i a...
31949	31950	0	hear one of my new songs! don't go - katie ell...
31950	31951	0	@user you can try to 'tail' us to stop, 'butt...
31951	31952	0	i've just posted a new blog: #secondlife #lone...
31952	31953	0	@user you went too far with @user
31953	31954	0	good morning #instagram #shower #water #berlin...
31954	31955	0	#holiday bull up: you will dominate your bul...
31955	31956	0	less than 2 weeks ð□□□ð□□□ð□□¼ð□□'ð□□□ð□□µ @us...
31956	31957	0	off fishing tomorrow @user carnt wait first ti...
31957	31958	0	ate @user isz that youuu?ð□□□ð□□□ð□□□ð□□□ð□□□ð...
31958	31959	0	to see nina turner on the airwaves trying to...
31959	31960	0	listening to sad songs on a monday morning otw...
31960	31961	1	@user #sikh #temple vandalised in in #calgary,...
31961	31962	0	thank you @user for you follow

31962 rows × 3 columns

As we have 3 attributes present in our dataset and a total of 31962 labeled tweets , ‘1’ standing for tweets with negative sentiment and ‘0’ for tweets with positive sentiments.

Reading the test.csv Pandas file

- In the first line we read the test.csv file using Pandas.
- In the second line as a safe backup we keep a copy of our original test.csv file. **We make a copy of test data so that even if we have to make any changes in this dataset we would not lose the original dataset.**

Overview of the test dataset

	id	tweet
0	31963	#studiolife #aislife #requires #passion #dedic...
1	31964	@user #white #supremacists want everyone to s...
2	31965	safe ways to heal your #acne!! #altwaystohe...
3	31966	is the hp and the cursed child book up for res...
4	31967	3rd #bihday to my amazing, hilarious #nephew...
5	31968	choose to be :) #momtips
6	31969	something inside me dies ð□□¦ð□□¿â□¨ eyes nes...
7	31970	#finished#tattoo#inked#ink#loveitâ□¤ï¸□ #â□¤ï¸...
8	31971	@user @user @user i will never understand why...
9	31972	#delicious #food #lovelife #capetown mannaep...
10	31973	1000dayswasted - narcosis infinite ep.. make m...
11	31974	one of the world's greatest spoing events #l...
12	31975	half way through the website now and #allgoing...
13	31976	good food, good life , #enjoy and ð□□□ð□□□ð□...
14	31977	i'll stand behind this #guncontrolplease #se...
15	31978	i ate,i ate and i ate...ð□□□ð□□□ #jamaisasth...
16	31979	@user got my @user limited edition rain or sh...
17	31980	& #love & #hugs & #kisses too! how...
18	31981	ð□□ð□□□ð□□□ #girls #sun #fave @ london, uni...
19	31982	thought factory: bbc neutrality on right wing ...
20	31983	hey guys tommorow is the last day of my exams ...
21	31984	@user @user @user #levyrroni #recuerdos mem...

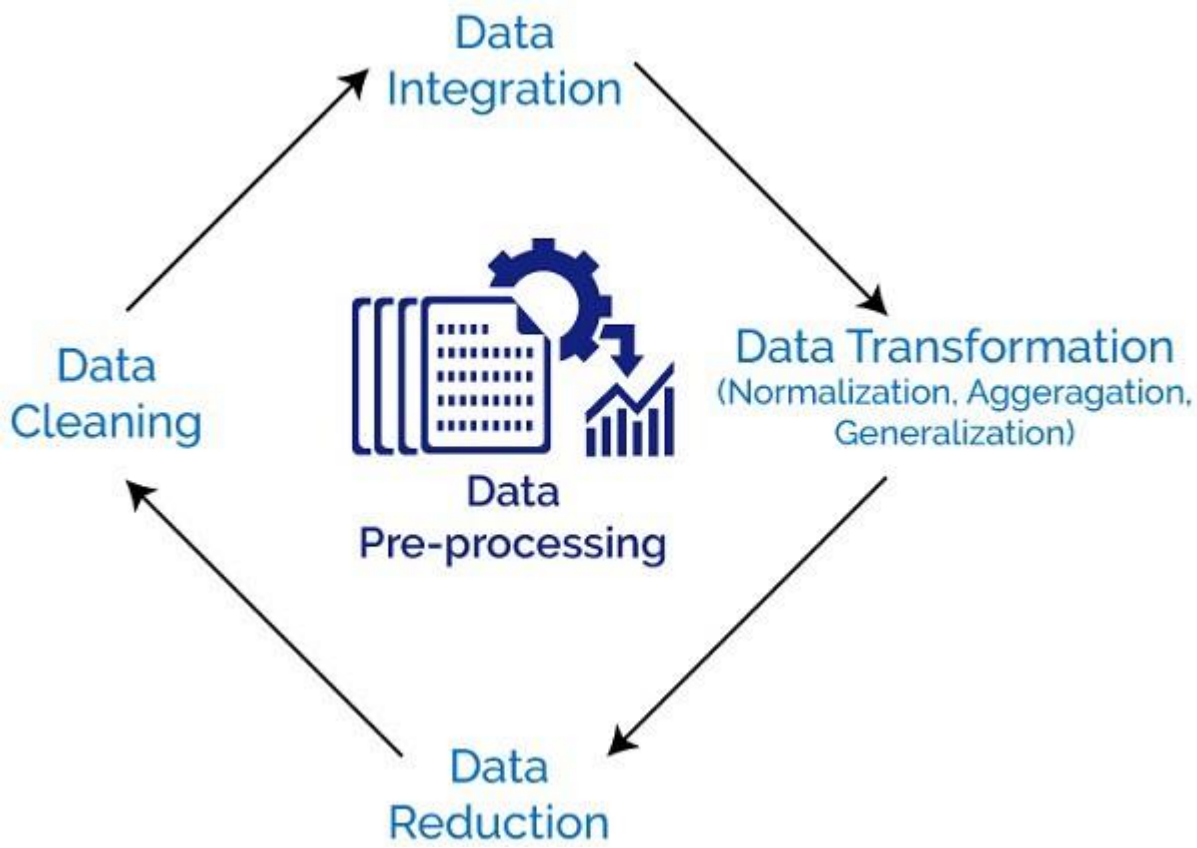
Data Pre-Processing

17175	49138	spread love not hateâ□¤ï¸□ð□□□ð□□□ð□□□ð□□□ #pr...
17176	49139	@user @user are the most racist pay ever!!!!!
17177	49140	i am thankful for children. #thankful #positiv...
17178	49141	liverpool â□¤ï¸□ð□□¬ð□□§ #walk #liverpool #sta...
17179	49142	#bakersfield rooster simulation: i want to c...
17180	49143	por do sol ó¾□□â□¤ï¸□#instagood #beautiful #...
17181	49144	@user hell yeah what a great surprise for your...
17182	49145	when ur the joke ur defensive towards everythi...
17183	49146	#enjoying the #evening #sun in my #bedroom â□¨...
17184	49147	tonight on @user from 9pm gmt you can here a ...
17185	49148	today is a good day for excercise #imready #so...
17186	49149	good night with a tea and music â□□ï¸□ð□□□ð□□□...
17187	49150	loving lifeð□□°ð□□¸â□□ï¸□ð□□□ #createyourfutu...
17188	49151	black professor demonizes, proposes nazi style...
17189	49152	learn how to think positive. #positive #ins...
17190	49153	we love the pretty, happy and fresh you! #teen...
17191	49154	2_damn_tuff-ruff_muff__techno_city-(ng005)-web...
17192	49155	thought factory: left-right polarisation! #tru...
17193	49156	feeling like a mermaid ð□□□ #hairflip #neverre...
17194	49157	#hillary #campaigned today in #ohio((omg)) &am...
17195	49158	happy, at work conference: right mindset leads...
17196	49159	my song "so glad" free download! #shoegaze ...

17197 rows × 2 columns

As we can see we have 2 **attributes** present here that is **'id'** and **'tweets'.** This is the dataset on which we are going to test our Machine Learning models so it is unlabeled.

STEP — 1 :



Steps of data pre-processing

Let's begin with the pre-processing of our dataset.

Combine the train.csv and test.csv files.

Pandas **dataframe.append()** function is used to append rows of other dataframe to the end of the given dataframe, returning a new dataframe object.

Overview of the combined train and test dataset.

Type **combine.head()** in the cell and you get the following result.

	id	label	tweet
0	1	0.0	@user when a father is dysfunctional and is s...
1	2	0.0	@user @user thanks for #lyft credit i can't us...
2	3	0.0	bihday your majesty
3	4	0.0	#model i love u take with u all the time in ...
4	5	0.0	factsguide: society now #motivation

Again type combine.tail() in the cell and you get the following result.

	id	label	tweet
49154	49155	NaN	thought factory: left-right polarisation! #tru...
49155	49156	NaN	feeling like a mermaid ð□□□ #hairflip #neverre...
49156	49157	NaN	#hillary #campaigned today in #ohio((omg)) &am...
49157	49158	NaN	happy, at work conference: right mindset leads...
49158	49159	NaN	my song "so glad" free download! #shoegaze ...

Test.csv appended with the train.csv file

Columns not in the original dataframes are added as new columns and the new cells are populated with NaN value.

STEP — 2

Removing Twitter Handles(@User)

In our analysis we can clearly see that the Twitter handles do not contribute anything significant to solve our problem. So it's better if we remove them in our dataset.

Given below is a user-defined function to remove unwanted text patterns from the tweets. It takes two arguments, one is the original string of text and the other is the pattern of text that we want to remove from the string. The function returns the same input string but without the given pattern. We will use this function to remove the pattern '@user' from all the tweets in our data.

Here NumPy Vectorization **'np.vectorize()'** is used because it is much more faster than the conventional for loops when working on datasets of medium to large sizes.

STEP — 3

	id	label	tweet	Tidy_Tweets
0	1	0.0	@user when a father is dysfunctional and is s...	when a father is dysfunctional and is so sel...
1	2	0.0	@user @user thanks for #lyft credit i can't us...	thanks for #lyft credit i can't use cause th...
2	3	0.0	bihday your majesty	bihday your majesty
3	4	0.0	#model i love u take with u all the time in ...	#model i love u take with u all the time in ...
4	5	0.0	factsguide: society now #motivation	factsguide: society now #motivation

After removing the twitter handles.

Removing Punctuation, Numbers, and Special Characters

Punctuation, numbers and special characters do not help much. It is better to remove them from the text just as we removed the twitter handles. Here we will replace everything except characters and hashtags with spaces.

STEP — 4

	id	label	tweet	Tidy_Tweets
0	1	0.0	@user when a father is dysfunctional and is s...	when a father is dysfunctional and is so sel...
1	2	0.0	@user @user thanks for #lyft credit i can't us...	thanks for #lyft credit i can t use cause th...
2	3	0.0	bihday your majesty	bihday your majesty
3	4	0.0	#model i love u take with u all the time in ...	#model i love u take with u all the time in ...
4	5	0.0	factsguide: society now #motivation	factsguide society now #motivation
5	6	0.0	[2/2] huge fan fare and big talking before the...	huge fan fare and big talking before the...
6	7	0.0	@user camping tomorrow @user @user @user @use...	camping tomorrow danny
7	8	0.0	the next school year is the year for exams.ð□□...	the next school year is the year for exams ...
8	9	0.0	we won!!! love the land!!! #allin #cavs #champ...	we won love the land #allin #cavs #champ...
9	10	0.0	@user @user welcome here ! i'm it's so #gr...	welcome here i m it s so #gr

Removing Short Words

We have to be a little careful here in selecting the length of the words which we want to remove. So, I have decided to remove all the words having length 3 or less. These words are also known as **Stop Words.**

For example, terms like “hmm”, “and”, “oh” are of very little use. It is better to get rid of them.

STEP — 5

	id	label	tweet	Tidy_Tweets
0	1	0.0	@user when a father is dysfunctional and is s...	when father dysfunctional selfish drags kids i...
1	2	0.0	@user @user thanks for #lyft credit i can't us...	thanks #lyft credit cause they offer wheelchai...
2	3	0.0	bihday your majesty	bihday your majesty
3	4	0.0	#model i love u take with u all the time in ...	#model love take with time
4	5	0.0	factsguide: society now #motivation	factsguide society #motivation
5	6	0.0	[2/2] huge fan fare and big talking before the...	huge fare talking before they leave chaos disp...
6	7	0.0	@user camping tomorrow @user @user @user @use...	camping tomorrow danny
7	8	0.0	the next school year is the year for exams.ð□□...	next school year year exams think about that #...
8	9	0.0	we won!!! love the land!!! #allin #cavs #champ...	love land #allin #cavs #champions #cleveland #...
9	10	0.0	@user @user welcome here ! i'm it's so #gr...	welcome here

Tokenization

Now we will tokenize all the cleaned tweets in our dataset. Tokens are individual terms or words, and tokenization is the process of splitting a string of text into tokens.

Here we tokenize our sentences because we will apply Stemming from the "NLTK" package in the next step.

```
0    [when, father, dysfunctional, selfish, drags, ...
1    [thanks, #lyft, credit, cause, they, offer, wh...
2                               [bihday, your, majesty]
3                     [#model, love, take, with, time]
4                   [factsguide, society, #motivation]
Name: Tidy_Tweets, dtype: object
```

STEP — 6

Stemming

Stemming is a rule-based process of stripping the suffixes (“ing”, “ly”, “es”, “s” etc) from a word.

For example — “play”, “player”, “played”, “plays” and “playing” are the different variations of the word — “play”

```
: 0    [when, father, dysfunct, selfish, drag, kid, i...
  1    [thank, #lyft, credit, caus, they, offer, whee...
  2                              [bihday, your, majesti]
  3                     [#model, love, take, with, time]
  4                         [factsguid, societi, #motiv]
  Name: Tidy_Tweets, dtype: object
```

Now let’s stitch these tokens back together

	id	label	tweet	Tidy_Tweets
0	1	0.0	@user when a father is dysfunctional and is s...	when father dysfunct selfish drag kid into dys...
1	2	0.0	@user @user thanks for #lyft credit i can't us...	thank #lyft credit caus they offer wheelchair ...
2	3	0.0	bihday your majesty	bihday your majesti
3	4	0.0	#model i love u take with u all the time in ...	#model love take with time
4	5	0.0	factsguide: society now #motivation	factsguid societi #motiv

Data Visualisation

So finally these are the basic steps to follow when we have to Pre-Process a dataset containing textual data.

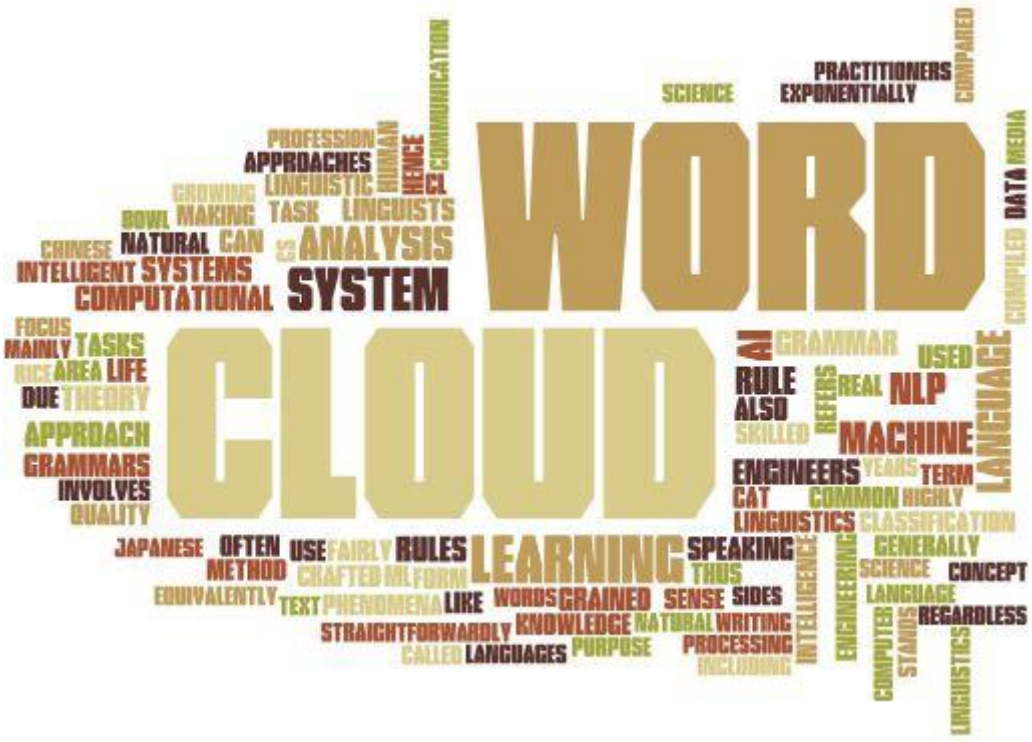
OK, so now we are done with our Data Pre-Processing stages.

Let’s move on to our next step that is **Data Visualisation.**

So Data Visualisation is one of the most important steps in Machine Learning projects because it gives us an approximate idea about the dataset and what it is all about before proceeding to apply different machine learning models.

WordCloud

One of the popular visualisation techniques is **WordCloud.**



A WordCloud is a visualisation wherein the most frequent words appear in large size and the less frequent words appear in smaller sizes.

So, in Python we have a package for generating **WordCloud**.

Let's dive into the code to see how can we generate a **WordCloud.**

Generating WordCloud for tweets with label ‘0’.

Importing packages necessary for generating a WordCloud

Store all the words from the dataset which are non-racist/sexist.

The code to generate the required **WordCloud**.

Each line has been properly commented for a better understanding.

Generating WordCloud for tweets with label ‘1‘.

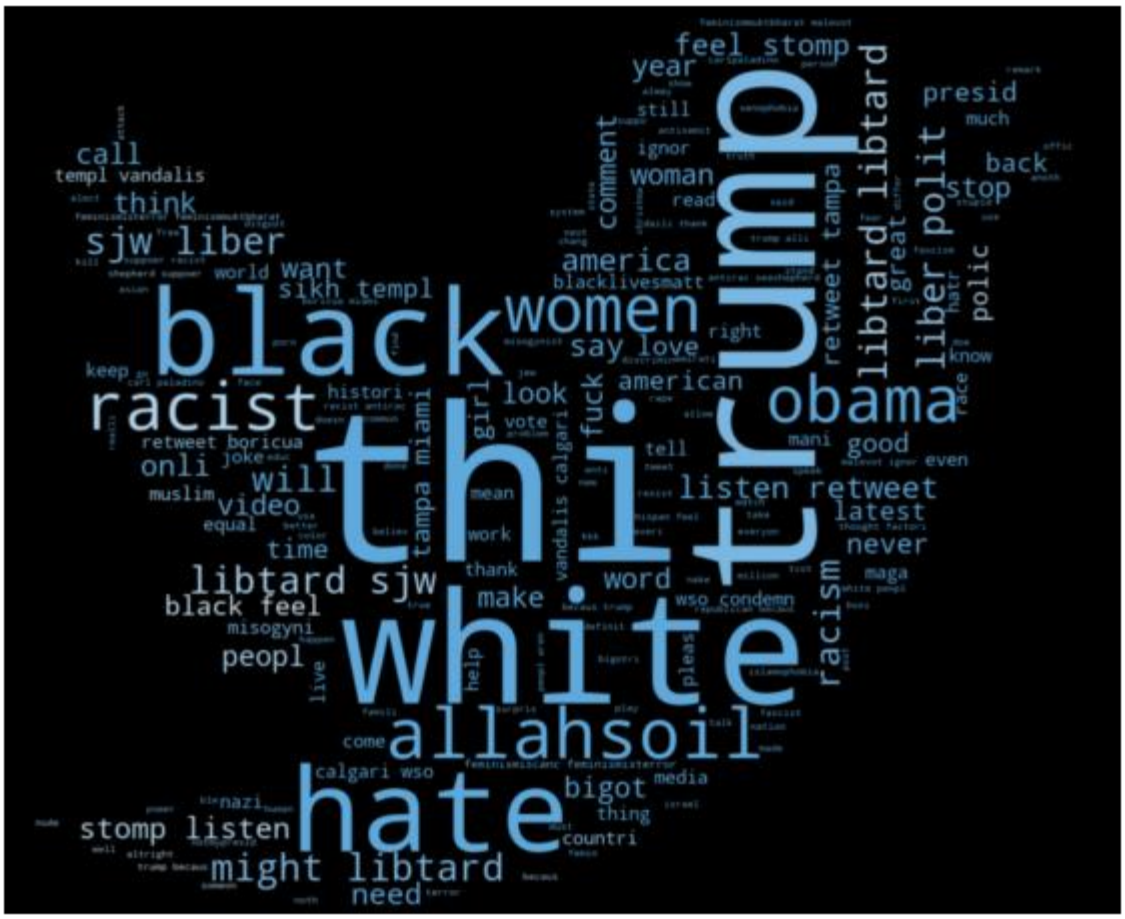
Generated WordCloud

We can see most of the words are positive or neutral. With happy, smile, and love being the most frequent ones. Hence, most of the frequent words are compatible with tweets in positive sentiment.

Store all the words from the dataset which are non-racist/sexist.

The code to generate the required **WordCloud**.

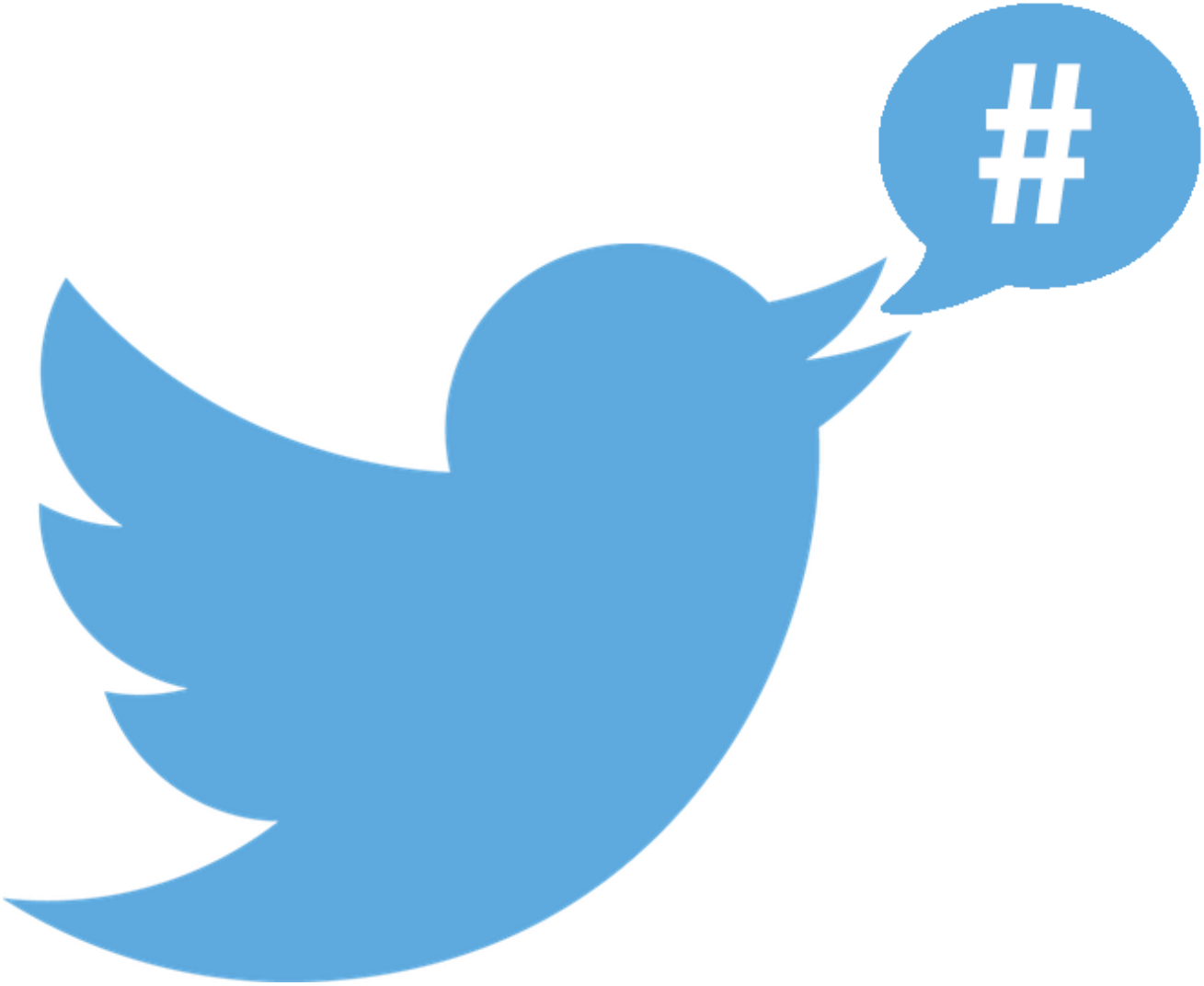
Each line has been properly commented for a better understanding.



Generated WordCloud

We can clearly see, most of the words have negative connotations. So, it seems we have a pretty good text data to work on.

Understanding the impact of Hashtags on tweets sentiment



Hash-tagging on Twitter can have a major impact when it comes to your follower count by using general and non-specific hashtags. If you hashtag

general words, like **#creative,** or events, like **#TIFF,** that are going on, it is more likely that your tweet will reach beyond your follower list.

So we will look how we can extract the hashtags and see which hashtags fall into which category.

Function to extract hashtags from tweets

A nested list of all the hashtags from the positive reviews from the dataset.

OUTPUT :

```
Out[87]: [['run'],
          ['lyft', 'disapoint', 'getthank'],
          [],
          ['model'],
          ['motiv'],
          ['allshowandnogo'],
          [],
          ['school', 'exam', 'hate', 'imagin', 'actorslif', 'revolutionschool', 'girl'],
          ['allin', 'cav', 'champion', 'cleveland', 'clevelandcavali'],
```

```
['got', 'junior', 'yugyoem', 'omg'],
['thank', 'posit'],
['friday', 'cooki'],
[],
['euro'],
['badday', 'coneofsham', 'cat', 'piss', 'funni', 'laugh'],
['wine', 'weekend'],
['tgif', 'gamedev', 'indiedev', 'indiegamedev', 'squad'],
['upsideofflorida', 'shopalyssa', 'love'],
['smile', 'media', 'pressconfer', 'antalya', 'turkey', 'throwback'],
```

Here we unnest the list

OUTPUT :

```
Out[28]: ['run',
          'lyft',
          'disapoint',
          'getthank',
          'model',
          'motiv',
          'allshowandnogo',
          'school',
          'exam',


'love',
'girl',
'snapchat',
'flower',
'instasmil',
'instalov',
'posit',
...]
```

A nested list of all the hashtags from the negative reviews from the dataset

OUTPUT :

```
Out[29]: [['cnn', 'michigan', 'tcot'],
          ['australia',
           'opkillingbay',
           'seashepherd',
           'helpcovedolphin',
           'thecov',
           'helpcovedolphin'],
          [],
          [],
          ['neverump', 'xenophobia'],
          ['love', 'peac'],
          [],
          ['race', 'ident', 'med'],
          ['altright', 'whitesupremaci'],
          ['linguist', 'race', 'power', 'raciolinguist'],
          ['brexit'],

['peopl', 'trump', 'republican'],
[],
['adl', 'hate', 'jewworldord', 'tyranni'],
['korean', 'anaco'],
[],
[],
[],
...]
```

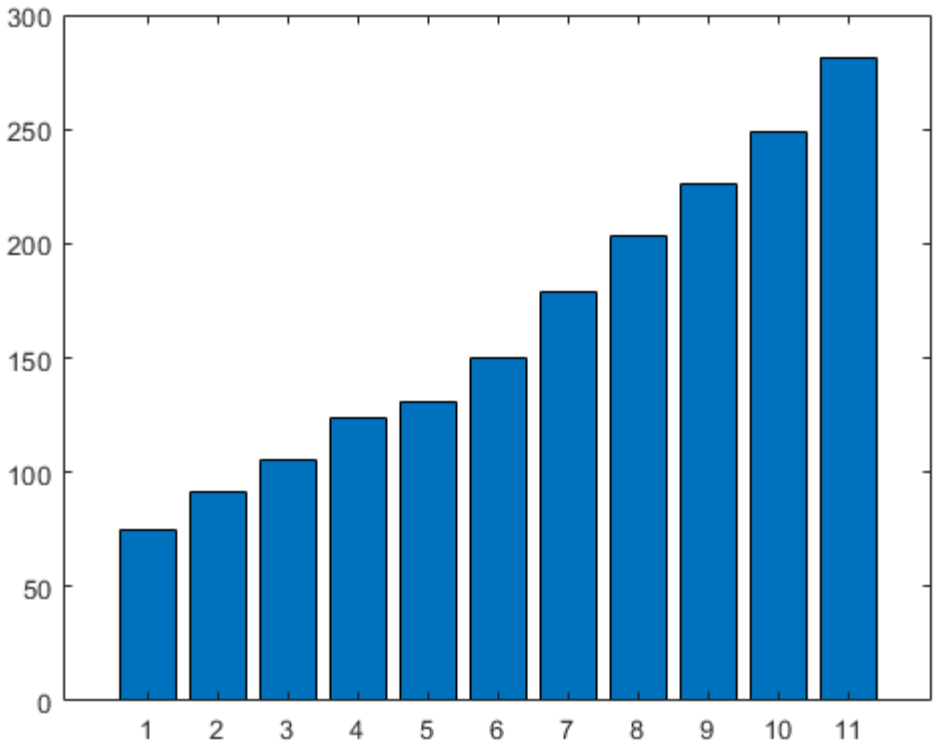
Here we unnest the list

OUTPUT :

```
Out[30]: ['cnn',
          'michigan',
          'tcot',
          'australia',
          'opkillingbay',
          'seashepherd',
          'helpcovedolphin',
          'thecov',
          'helpcovedolphin',
          'neverump',
          'xenophobia',
```

For Positive Tweets in the dataset

```
'lifeguard',
'wow',
'marxist',
'propoganda',
'allahsoil',
'teambt',
'mustread',
'educ',
'stereotyp',
'bustymilf',
...]
```



Counting the frequency of the words having Positive Sentiment

OUTPUT:

FreqDist({'love': 1654, 'posit': 917, 'smile': 676, 'healthi': 573, 'thank': 534, 'fun': 463, 'life': 425, 'affirm': 423, 'summer': 390, 'model': 375, ...})

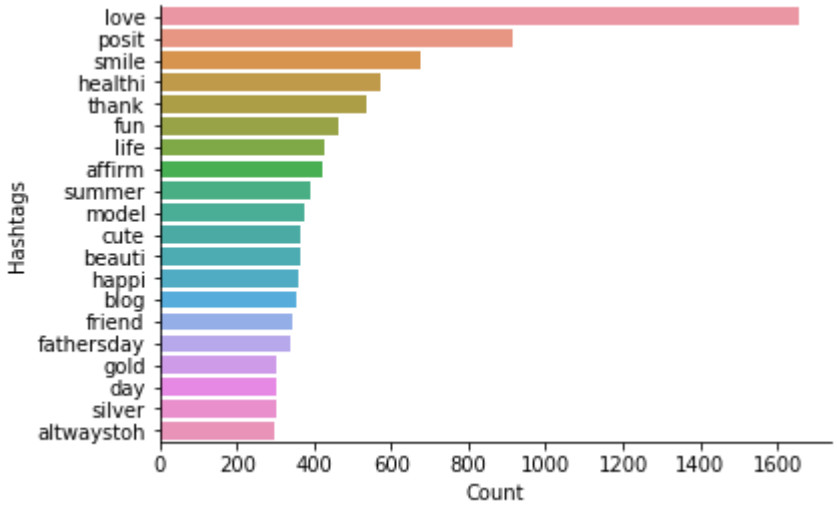
Creating a dataframe for the most frequently used words in hashtags

Out[33]:

	Hashtags	Count
0	run	72
1	lyft	2
2	disapoint	1
3	getthank	2
4	model	375
5	motiv	202
6	allshowandnogo	1
7	school	30
8	exam	9
9	hate	27

Plotting the barplot for the 20 most frequent words used for hashtags

For Negative Tweets in the dataset



Count BarPlot

Counting the frequency of the words having Negative Sentiment

OUTPUT :

FreqDist({'trump': 136, 'polit': 95, 'allahsoil': 92, 'liber': 81, 'libtard': 77, 'sjw': 75, 'retweet': 63, 'black': 46, 'miami': 46, 'hate': 37, ...})

Creating a dataframe for the most frequently used words in hashtags

Out[39]:

	Hashtags	Count
0	cnn	10
1	michigan	2
2	tcot	14
3	australia	6
4	opkillingbay	5
5	seashepherd	22
6	helpcovedolphin	3
7	thecov	4
8	neverump	8
9	xenophobia	12

Plotting the barplot for the 20 most frequent words used for hashtags