In [1]:

```
#importing all necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sb
import re
import string
from tqdm import tqdm
from tqdm.notebook import tqdm_notebook
tqdm_notebook.pandas()
from nltk.corpus import stopwords
from nltk.util import ngrams
stop=set(stopwords.words('english'))
from collections import Counter
```

In [2]:

```
#importing data from csv files using Pandas
train_df = pd.read_csv('train.csv')
test_df = pd.read_csv('test.csv')
```

In [3]:

```
train_df.shape,test_df.shape
```

Out[3]:

```
((27481, 4), (3534, 3))
```

In [4]:

```
train_df.head()
```

Out[4]:

	textID	text	selected_text	sentiment
0	cb774db0d1	I`d have responded, if I were going	I`d have responded, if I were going	neutral
1	549e992a42	Sooo SAD I will miss you here in San Diego!!!	Sooo SAD	negative
2	088c60f138	my boss is bullying me...	bullying me	negative
3	9642c003ef	what interview! leave me alone	leave me alone	negative
4	358bd9e861	Sons of ****, why couldn`t they put them on t...	Sons of ****,	negative

In [5]:

```
test_df.head()
```

Out[5]:

	textID	text	sentiment
0	f87dea47db	Last session of the day http://twitpic.com/67ezh	neutral
1	96d74cb729	Shanghai is also really exciting (precisely -...	positive
2	eee518ae67	Recession hit Veronique Branquinho, she has to...	negative
3	01082688c6	happy bday!	positive
4	33987a8ee5	http://twitpic.com/4w75p - I like it!!	positive

In [6]:

```
train_df.isnull().sum(),test_df.isnull().sum()
```

Out[6]:

```
(textID           0
 text             1
 selected_text    1
 sentiment        0
 dtype: int64, textID       0
 text         0
 sentiment    0
 dtype: int64)
```

Checking for null and duplicate and remove if any

In [7]:

```
train_df[train_df.text.isna()]
```

Out[7]:

	textID	text	selected_text	sentiment
314	fdb77c3752	NaN	NaN	neutral

In [8]:

```
train_df.dropna(axis=0,inplace=True)
train_df.shape,test_df.shape
```

Out[8]:

```
((27480, 4), (3534, 3))
```

In [9]:

```
train_df.duplicated().sum(),test_df.duplicated().sum()
```

Out[9]:

```
(0, 0)
```

Data Cleaning

In [10]:

```
#text cleaning or text preprocessing
def clean_text(text):
    #converting to lower case
    text = str(text).lower()
    #Remove the text in square brackets
    text = re.sub('\[.*?\]', ' ', text)
    #remove links in given text
    text = re.sub('https?://\S+|www\.\S+', ' ', text)
    #remove text/chars within angular brackets
    text = re.sub('<.*?>+', ' ', text)
    #replace ***** with <CURSE>
    text = re.sub("\*+", "<CURSE>", text)
    #remove punctuations from a string
    string.punctuation =  '!"#$%&\'()*+,-./:;<=>?@[\\]^_{|}~'
    text = re.sub('[%s]' % re.escape(string.punctuation), ' ', text)
    #remove expressions like \n from text
    text = re.sub('\n', ' ', text)
    #remove words containing numbers
    text = re.sub('\w*\d\w*', ' ', text)

    text = text.split()
    text = " ".join(text)
    return text
```

In [11]:

```
train_df['text'] = train_df['text'].progress_apply(lambda x : clean_text(x))
train_df['selected_text'] = train_df['selected_text'].progress_apply(lambda x : clean_text(x))
```

In [12]:

```
test_df['text'] = test_df['text'].progress_apply(lambda x : clean_text(x))
```

In [13]:
```
# Check for any null values.

train_df.isnull().sum(), test_df.isnull().sum()
```

Out[13]:
```
(textID           0
 text             0
 selected_text    0
 sentiment        0
 dtype: int64, textID       0
 text         0
 sentiment    0
 dtype: int64)
```

In [14]:
```
train_df[train_df.text=='']
```

Out[14]:

	textID	text	selected_text	sentiment
24926	0872ed0f00			neutral
26005	0b3fe0ca78			neutral

In [15]:
```
train_df[train_df.selected_text=='']
```

Out[15]:

	textID	text	selected_text	sentiment
9533	c6149b7abf	its castiel		positive
10997	6d85945e2c	hussein i have to wake up earlier than i thoug...		negative
13728	c7b78c1b26	btw and i ordered some of yer merch yesterday		positive
24348	a21d9c38a8	i`ll keep working on her lol good idea		positive
24926	0872ed0f00			neutral
25455	b6bf74a5c8	europe sounds will finish my exam on teus and ...		positive
25637	370880f242	it`s realy my book is on the side i`m not stud...		negative
26005	0b3fe0ca78			neutral

In [16]:
```
test_df[test_df.text=='']
```

Out[16]:

	textID	text	sentiment
196	b47c430fda		neutral

From the above, we could see that there are few rows where the text and selected_text columns are empty. we need to fix those values

In [17]:
```
train_df.drop(labels = train_df[train_df.text==''].index,inplace=True)
```

In [18]:
```
train_df.drop(labels = train_df[train_df.selected_text==''].index,inplace=True)
```

In [19]:
```
test_df.drop(labels = test_df[test_df.text==''].index,inplace=True)
```

In [20]:
```
train_df[train_df.text=='']
```

Out[20]:

	textID	text	selected_text	sentiment

In [21]:
```
train_df[train_df.selected_text=='']
```

Out[21]:

	textID	text	selected_text	sentiment

In [22]:
```
test_df[test_df.text=='']
```

Out[22]:

	textID	text	sentiment

Find Misspelled words from selected_text column

In [23]:
```
def misspelled_words(x):
    text,sel_text = x[0],x[1]
    misspelled =[]
    text = text.split()
    sel_text = sel_text.split()
    for each in sel_text:
        if each not in text:
            misspelled.append(each)

    if len(misspelled)>0:
        return " ".join(misspelled)
    else:
        return 'No'
```

In [24]:
```
train_df['misspelled'] = train_df[['text','selected_text']].progress_apply(misspelled_words,axis=1)
```

In [25]:
```
train_df[train_df.misspelled != 'No']
```

Out[25]:

	textID	text	selected_text	sentiment	misspelled
18	af3fed7fc3	is back home now gonna miss every one	onna	negative	onna
32	1c31703aef	if it is any consolation i got my bmi tested h...	well so much for being unhappy for about minute	negative	minute
39	2863f435bd	a little happy for the wine jeje ok it`sm my f...	a little happy fo	positive	fo
48	3d9d4b0b55	i donbt like to peel prawns i also dont like g...	dont like go	negative	go
49	3fcea4debc	which case i got a new one last week and i`m n...	d i`m not thrilled at all with mine	negative	d
...	...	...	...	...	...
27426	132e051fe8	my cousins moved there like years ago and i mi...	m sad	negative	m
27456	d32efe060f	i wanna leave work already not feelin it	wanna leave work al	negative	al
27470	778184dff1	lol i know and haha did you fall asleep or jus...	t bored	negative	t
27474	8f14bb2715	so i get up early and i feel good about the da...	i feel good ab	positive	ab
27476	4eac33d1c0	wish we could come see u on denver husband los...	d lost	negative	d

1443 rows × 5 columns

In the given training data, there are around 1443 rows in which the selected_text **contains chars/words that needs to be cleaned**

I am trying to filter out the rows which contains only 1 char that needs to be cleaned

In [26]:
```
train_df[train_df['misspelled'].apply(lambda x: len(x)) ==1]
```

Out[26]:

	textID	text	selected_text	sentiment	misspelled
49	3fcea4debc	which case i got a new one last week and i`m n...	d i`m not thrilled at all with mine	negative	d
66	95e12b1cb1	he`s awesome have you worked with him before h...	s awesome	positive	s
129	94f67cfa6d	hey mia totally adore your music when will you...	y adore	positive	y
134	6903cb08f2	nice to see you tweeting it`s sunday may and w...	e nice	positive	e
166	c78bf59e67	lichfield tweetup sounds like fun hope to see ...	p sounds like fun	positive	p
...	...	...	...	...	...
27153	a044ed928d	enjoy nola definitely one of my favorite citie...	y one of my favorite cities in the world	positive	y
27240	40143b692e	who knows it makes me sad lol	e sad	negative	e
27426	132e051fe8	my cousins moved there like years ago and i mi...	m sad	negative	m
27470	778184dff1	lol i know and haha did you fall asleep or jus...	t bored	negative	t
27476	4eac33d1c0	wish we could come see u on denver husband los...	d lost	negative	d

688 rows × 5 columns

Around **688 rows** has single misspelled character in selected_text column. We can just remove those single chars thats causing this issue

In [27]:
```
def clean_sel_text(x):
    text,sel_text,misspelled = x[0],x[1],x[2]
    sel_text = sel_text.split()
    sel_text.remove(misspelled)
    return " ".join(sel_text)
```

In [28]:
```
train_df['selected_text']=train_df[['text','selected_text','misspelled']].progress_apply(lambda x: clean_sel_text(x)
                                                    if len(x['misspelled']) == 1 else x['selected_text'], axis=1)
```

Again finding the misspelled words in char column, after removing the single characters

In [29]:
```
train_df['misspelled'] = train_df[['text','selected_text']].progress_apply(misspelled_words,axis=1)
train_df[train_df.misspelled != 'No']
```

Out[29]:

	textID	text	selected_text	sentiment	misspelled
18	af3fed7fc3	is back home now gonna miss every one	onna	negative	onna
32	1c31703aef	if it is any consolation i got my bmi tested h...	well so much for being unhappy for about minute	negative	minute
39	2863f435bd	a little happy for the wine jeje ok it`sm my f...	a little happy fo	positive	fo
48	3d9d4b0b55	i donbt like to peel prawns i also dont like g...	dont like go	negative	go
247	ce69e99e71	i`m not sleeping at all until accepts my appology	i`m not sleeping at all un	negative	un
...	...	...	...	...	...
27362	7b82d63ee4	just found out i won`t be tweeting from my pho...	c sorr	negative	c sorr
27386	e149ebd3a1	would one of the vwllers want to add this even...	ch appreciat	positive	ch appreciat
27401	261e064dd4	oh silence verona i am wanting to go jaja enjo...	ja enjoyyitverymu	positive	ja enjoyyitverymu
27456	d32efe060f	i wanna leave work already not feelin it	wanna leave work al	negative	al
27474	8f14bb2715	so i get up early and i feel good about the da...	i feel good ab	positive	ab

755 rows × 5 columns

In [30]:
```
train_df[train_df['misspelled'].apply(lambda x: len(x)) ==1]
```

Out[30]:

	textID	text	selected_text	sentiment	misspelled

In [31]:
```
train_df[(train_df.misspelled != 'No')]
```

Out[31]:

	textID	text	selected_text	sentiment	misspelled
18	af3fed7fc3	is back home now gonna miss every one	onna	negative	onna
32	1c31703aef	if it is any consolation i got my bmi tested h...	well so much for being unhappy for about minute	negative	minute
39	2863f435bd	a little happy for the wine jeje ok it`sm my f...	a little happy fo	positive	fo
48	3d9d4b0b55	i donbt like to peel prawns i also dont like g...	dont like go	negative	go
247	ce69e99e71	i`m not sleeping at all until accepts my appology	i`m not sleeping at all un	negative	un
...	...	...	...	...	...
27362	7b82d63ee4	just found out i won`t be tweeting from my pho...	c sorr	negative	c sorr
27386	e149ebd3a1	would one of the vwllers want to add this even...	ch appreciat	positive	ch appreciat
27401	261e064dd4	oh silence verona i am wanting to go jaja enjo...	ja enjoyyitverymu	positive	ja enjoyyitverymu
27456	d32efe060f	i wanna leave work already not feelin it	wanna leave work al	negative	al
27474	8f14bb2715	so i get up early and i feel good about the da...	i feel good ab	positive	ab

755 rows × 5 columns

https://www.nltk.org/api/nltk.metrics.html#:~:text=The%20edit%20distance%20is%20the,%E2%80%9D%20%2D%3E%20%E2%80%9Cshine%E2%80%9D (https://www.nltk.org/api/nltk.metrics.html#:~:text=The%20edit%20distance%20is%20the,%E2%80%9D%20%2D%3E%20%E2%80%9Cshine%E2%80%9D). Here we are gonna use a new fuction called **edit_distance** available in nltk module.The edit distance is the number of #characters that need to be substituted, inserted, or deleted, to transform s1 into s2.

In [32]:
```
#import nltk
#nltk.edit_distance("la", "a")
```

In [33]:

```
#import nltk
#def correct_spelling(x):

#    text,sel_text,misspelled = x[0],x[1],x[2]
#    sel_text = sel_text.split()
#    text = text.split()
#    misspelled = misspelled.split()
#    misspelled = [each for each in misspelled if len(each)>1]
#    for mis in misspelled:
#        for each in text:
#            if mis in sel_text:
#                if (nltk.edit_distance(each,mis)==1):
#                    index = sel_text.index(mis)
#                    sel_text[index]=each

#    return " ".join(sel_text)
```

In [34]:

```
#train_df['selected_text']=train_df[['text','selected_text','misspelled']].progress_apply(lambda x: correct_spelling(x)
#                                                      if (x['misspelled']) != 'No' else x['selected_text'], axis=1)
```

the nltk.edit_distance function didnt work very well. Hence using fuzz.ratio() function

In [35]:

```
from fuzzywuzzy import fuzz
def check_fuzzy(x):
    text,sel_text,misspelled = x[0],x[1],x[2]
    sel_text = sel_text.split()
    text = text.split()
    misspelled = misspelled.split()
    for mis in misspelled:
        for each in text:
            if mis in sel_text:
                if (fuzz.ratio(each,mis)>70):
                    index = sel_text.index(mis)
                    sel_text[index]=each

    return " ".join(sel_text)
```

In [36]:

```
train_df['selected_text']=train_df[['text','selected_text','misspelled']].progress_apply(lambda x: check_fuzzy(x)
                                                      if (x['misspelled']) != 'No' else x['selected_text'], axis=1)
```

In [37]:

```
train_df['misspelled'] = train_df[['text','selected_text']].progress_apply(misspelled_words,axis=1)
train_df[(train_df.misspelled != 'No')]
```

Out[37]:

	textID	text	selected_text	sentiment	misspelled
48	3d9d4b0b55	i donbt like to peel prawns i also dont like g...	dont like go	negative	go
247	ce69e99e71	i`m not sleeping at all until accepts my apology	i`m not sleeping at all un	negative	un
251	77ba0fee75	powerblog what is this powerblog challenge you...	g what is this powerblog challenge you keep ta...	neutral	g
349	322b61740c	degrees gross skies and thunderstorms perfect ...	perfect ma	positive	ma
362	b94aaf845e	please review sunehre ad placement	please re	positive	re
...	...	...	...	...	...
27362	7b82d63ee4	just found out i won`t be tweeting from my pho...	c sorry	negative	c
27386	e149ebd3a1	would one of the vwllers want to add this even...	ch appreciate	positive	ch
27401	261e064dd4	oh silence verona i am wanting to go jaja enjo...	ja enjoyyitverymuch	positive	ja
27456	d32efe060f	i wanna leave work already not feelin it	wanna leave work al	negative	al
27474	8f14bb2715	so i get up early and i feel good about the da...	i feel good ab	positive	ab

321 rows × 5 columns

In [38]:

```
train_df[train_df['misspelled'].apply(lambda x: len(x)) ==1]
```

Out[38]:

	textID	text	selected_text	sentiment	misspelled
251	77ba0fee75	powerblog what is this powerblog challenge you...	g what is this powerblog challenge you keep ta...	neutral	g
1077	3d5c1ed21b	up is out i didn`t get the memo it looks amazing	o it looks amazing	positive	o
1363	4eec486ad7	hey i loved acs but i had to see it online is ...	y i loved acs but i had to see it online is no...	positive	y
1548	9b5db85d83	had a little family party tonight hope it rocked	t hope it rocked	positive	t
4729	ae1ada9158	tell me about it what is obvious in real life ...	e makes for great entertainment on tv it is gr...	positive	e
4764	43e6d9aeaa	well i guess they think of everything thanks s...	g think	positive	g
5213	2814c8f7c4	its not a starfleet one its not even a romulin...	t lame	negative	t
5358	12f2262ade	me too i was in florida last weekend for the r...	n terrible	negative	n
5385	213632bbd4	hopping in the shower you can help me tidy my ...	y my room its a CURSE hole	negative	y
6939	a50c8dc573	o its feels like a hot box and no matter where...	s like a hot box and	negative	s
6948	7b25c09b0f	i`m sure lots of that studio equipment was col...	y condolences	negative	y
9063	3e775363a1	we can`t even call you from belgium sucks	m sucks	negative	m
9539	825b22b853	wait and electrik red or richgirl i`m a sucker...	l i`m a sucker for the later	negative	l
11431	e158424933	nope san leandro marina how are you hope you`r...	u hope you`re well	positive	u
11837	20c8263baf	i have been getting CURSE ones as i mentioned ...	d all these girls seem to be at a loss	negative	d
13445	af0ac6b470	presentation went well yes i also met a buch o...	h of cool people checked your portfolio nice w...	positive	h
14571	7abcaab000	i`m out of town next week we`ll have to party ...	k happy	positive	k
14611	ae2f20ed25	wow that is so awesome andre it will be great ...	s so awesome	positive	s
17627	e040d9f166	what did i do to you sheesh	u sheesh	negative	u
18314	842b557aab	if any one is looking for he is now at and if ...	s good guy	positive	s
20305	be2de1ed61	erincharde hey you guys should invite me out m...	e hey you guys should invite me out moody is m...	neutral	e
21495	0d0959690c	were you able to watch it online i hope you we...	h belinda jensen was really good	positive	h
22928	3c68f6963c	lol on blip fm is not the on twitter i hate th...	r i hate that on	negative	r
23269	0ca5e24eb3	does she like it or does she roll her eyes now...	s now i`m curious as hell	positive	s
25391	7972092a15	monday funday wake up people and keep me awake...	t eww	negative	t
27362	7b82d63ee4	just found out i won`t be tweeting from my pho...	c sorry	negative	c

Now too, we could see there are few rows where there is only 1 char misspelled. We are gonna remoe those characters

In [39]:
```
train_df['selected_text']=train_df[['text','selected_text','misspelled']].progress_apply(lambda x: clean_sel_text(x)
                                                    if len(x['misspelled']) == 1 else x['selected_text'], axis=1)
```

In [40]:
```
train_df['misspelled'] = train_df[['text','selected_text']].progress_apply(misspelled_words,axis=1)
train_df[train_df.misspelled != 'No']
```

Out[40]:

	textID	text	selected_text	sentiment	misspelled
48	3d9d4b0b55	i donbt like to peel prawns i also dont like g...	dont like go	negative	go
247	ce69e99e71	i`m not sleeping at all until accepts my apology	i`m not sleeping at all un	negative	un
349	322b61740c	degrees gross skies and thunderstorms perfect ...	perfect ma	positive	ma
362	b94aaf845e	please review sunehre ad placement	please re	positive	re
366	b751f39570	yea i should know but tell me everything ps se...	yea i should know but tell me everything ps se...	neutral	ha
...	...	...	...	...	...
27302	90c8aa60db	have i ever told you i absolutly hate writing ...	hate wr	negative	wr
27386	e149ebd3a1	would one of the vwllers want to add this even...	ch appreciate	positive	ch
27401	261e064dd4	oh silence verona i am wanting to go jaja enjo...	ja enjoyyitverymuch	positive	ja
27456	d32efe060f	i wanna leave work already not feelin it	wanna leave work al	negative	al
27474	8f14bb2715	so i get up early and i feel good about the da...	i feel good ab	positive	ab

295 rows × 5 columns

In [41]:
```
train_df[train_df['misspelled'].apply(lambda x: len(x)) ==1]
```

Out[41]:

	textID	text	selected_text	sentiment	misspelled

We still have around **295 words** that has misspelled chars/words in selected_text column and it needs to be corrected

In [42]:
```
train_df[(train_df.misspelled != 'No') & (train_df.sentiment =='neutral')]
```

Out[42]:

	textID	text	selected_text	sentiment	misspelled
366	b751f39570	yea i should know but tell me everything ps se...	yea i should know but tell me everything ps se...	neutral	ha
2136	cc4a151d1d	ï¿½anisalovesu me too i am so so upset especia...	lovesu me too i am so so upset especially beca...	neutral	lovesu
2385	13d0be3942	must be time of month watering eyes again spec...	st be time of month watering eyes again	neutral	st
3187	e14d41126f	photo got my prints a few days ago ready for t...	got my prints a few days ago ready for the no	neutral	no
5697	e8c90dee68	im kinda bored anyone else i think ill listen ...	im kinda bored anyone else i think ill listen ...	neutral	ww
5799	13889237c6	another day and i couldn`t find you i ended up...	another day and i couldn`t find you i ended up...	neutral	`si
7780	594162e9a0	tat you looks beautiful and are a heck of a ma...	tat you looks beautiful and are a heck of a ma...	neutral	da
9385	b1bc6d98f9	wish i was famous like some of d people im fol...	wish i was famous li	neutral	li
10411	6c3f13ed50	horten moss kï¿½	horten moss kï	neutral	kï
10892	3e4b17c28b	also check out and other wonderful causes that...	also check out and other wonderful causes that...	neutral	he
12787	879f64f30d	bullet train from tokyo the gf and i have been...	bullet train from tokyo the gf and i have been...	neutral	go
15651	bdf5a8fe4c	eh you`re a really nice girl how are you miss ...	ly nice girl how are you miss youuuuuuu	neutral	ly
16423	02d8181035	well im gonna go shower now gotta get rdy movi...	well im gonna go shower now gotta get rdy movi...	neutral	kath
17467	8731f9f41f	officially moss doesnt work on vista but unoff...	officially moss doesnt work on vista but unoff...	neutral	fi
18011	dd638181f9	yea i will haha	yea i will ha	neutral	ha
20584	caabd948d9	was gonna dm you but it says you`re not follow...	was gonna dm you but it says you`re not follow...	neutral	hah
20872	d46919e9b3	is wondering where all her friends went	is wondering wh	neutral	wh
23784	fd0493acff	i miss jack n box and whataburger and oooo tac...	i miss jack n box and whataburger and oooo tac...	neutral	lm
25947	0216f668c9	woot lol it`s gonna be hard to send this one back	woot lol it`s gonna be hard to send this one ba	neutral	ba
26687	c5de5361d2	does anyone out there want to be really awesom...	does anyone out there want to be really awesom...	neutral	ht
26870	ff77427519	will do it in a couple od days when i have mor...	will do it in a couple od days when i have mor...	neutral	pe
26882	336b9cfc93	xdxdxd you crazy little thing why didnï¿½t you...	dxd you crazy little thing why didnï¿½t you ge...	neutral	dxd

From the above, we could conclude that for **Neutral** text, the data in text and selected_text columns are almost same. So we can just copy the whole content in text column to selected_text column for these nuetral texts

In [43]:
```
train_df['selected_text'] = train_df.progress_apply(lambda x: x['text']
                                if ( (x['misspelled'] != 'No') & (x['sentiment'] =='neutral')) else x['selected_text'],axis=1)
```

In [44]:
```
train_df['misspelled'] = train_df[['text','selected_text']].progress_apply(misspelled_words,axis=1)
```

In [45]:
```
train_df[(train_df.misspelled != 'No') & (train_df.sentiment =='neutral')]
```

Out[45]:

	textID	text	selected_text	sentiment	misspelled

Now all the neutral texts has **No Misspelled words** in selected text column

In [46]:
```
train_df[(train_df.misspelled != 'No') & (train_df.sentiment =='positive')]
```

Out[46]:

	textID	text	selected_text	sentiment	misspelled
349	322b61740c	degrees gross skies and thunderstorms perfect ...	perfect ma	positive	ma
362	b94aaf845e	please review sunehre ad placement	please re	positive	re
809	31fa81e0ae	yes you should go see star trek it`s sooooo mu...	ch fun	positive	ch
931	568ad4a905	happy star wars day may the be with you read f...	happy st	positive	st
1105	5419aaf31e	packing up and leaving inlaws house heading ho...	nice we	positive	we
...	...	...	...	...	...
25908	fdd49e735a	ok time for bed good night twitter	good night tw	positive	tw
26377	54c1aaf23c	working on your birthday isn`t so bad when you...	isn`t so bad wh	positive	wh
27386	e149ebd3a1	would one of the vwllers want to add this even...	ch appreciate	positive	ch
27401	261e064dd4	oh silence verona i am wanting to go jaja enjo...	ja enjoyyitverymuch	positive	ja
27474	8f14bb2715	so i get up early and i feel good about the da...	i feel good ab	positive	ab

152 rows × 5 columns

In [47]:
```
train_df[(train_df.misspelled != 'No') & (train_df.sentiment =='negative')]
```

Out[47]:

	textID	text	selected_text	sentiment	misspelled
48	3d9d4b0b55	i donbt like to peel prawns i also dont like g...	dont like go	negative	go
247	ce69e99e71	i`m not sleeping at all until accepts my appology	i`m not sleeping at all un	negative	un
863	19d585c61b	my poor heather she didnt make the cheerleadin...	sorry ba	negative	ba
1159	62086f9961	is sad that she is not seeing basshunter at me...	is sad th	negative	th
1303	915c66ead8	acsvxdcbgfn soccer now shall see young phoebe ...	i don`t want her dressed up th	negative	th
...	...	...	...	...	...
26830	24d37f9ba7	my bracelet broke today too	elet br	negative	elet br
27121	a11ddb9e58	the geography was an exam today but turned out...	ry nervous	negative	ry
27209	30a1e8c2b4	guys i know my ability to read time telling de...	es failed	negative	es
27302	90c8aa60db	have i ever told you i absolutly hate writing ...	hate wr	negative	wr
27456	d32efe060f	i wanna leave work already not feelin it	wanna leave work al	negative	al

121 rows × 5 columns

We still have **152 rows** in positive sentiment texts and **121 rows** in Negative sentiment texts that needs to be cleaned

In [48]:
```
from fuzzywuzzy import fuzz
def check_fuzzy(x):
    text,sel_text,misspelled = x[0],x[1],x[2]
    sel_text = sel_text.split()
    text = text.split()
    misspelled = misspelled.split()
    for mis in misspelled:
        for each in text:
            if mis in sel_text:
                if (fuzz.ratio(each,mis)>65):
                    index = sel_text.index(mis)
                    sel_text[index]=each

    return " ".join(sel_text)
```

In [49]:
```
train_df['selected_text']=train_df[['text','selected_text','misspelled']].progress_apply(lambda x: check_fuzzy(x)
                                                     if (x['misspelled']) != 'No' else x['selected_text'], axis=1)
```

In [50]:
```
train_df['misspelled'] = train_df[['text','selected_text']].progress_apply(misspelled_words,axis=1)
```

In [51]:
```
train_df[(train_df.misspelled != 'No') & (train_df.sentiment =='positive')]
```

Out[51]:

	textID	text	selected_text	sentiment	misspelled
349	322b61740c	degrees gross skies and thunderstorms perfect ...	perfect ma	positive	ma
362	b94aaf845e	please review sunehre ad placement	please re	positive	re
1105	5419aaf31e	packing up and leaving inlaws house heading ho...	nice we	positive	we
1531	a0ee798944	waiting to go to bed had a great weekend	great we	positive	we
1580	8a159382ea	just woke up gonna have a shower and go to nan...	happy mo	positive	mo
...	...	...	...	...	...
24169	cc58093a1b	i know our cats could be family mikesh is so cute	sh is so cute	positive	sh
24682	fee6cd911d	wave looks interesting if we are going to live...	wave looks interesting ht	positive	ht
25760	61306b5f1e	bella these are my family girls love u guys CU...	ls love	positive	ls
25908	fdd49e735a	ok time for bed good night twitter	good night tw	positive	tw
27474	8f14bb2715	so i get up early and i feel good about the da...	i feel good ab	positive	ab

84 rows × 5 columns

In [52]:
```
train_df[(train_df.misspelled != 'No') & (train_df.sentiment =='negative')]
```

Out[52]:

	textID	text	selected_text	sentiment	misspelled
48	3d9d4b0b55	i donbt like to peel prawns i also dont like g...	dont like go	negative	go
247	ce69e99e71	i`m not sleeping at all until accepts my appology	i`m not sleeping at all un	negative	un
863	19d585c61b	my poor heather she didnt make the cheerleadin...	sorry ba	negative	ba
1303	915c66ead8	acsvxdcbgfn soccer now shall see young phoebe ...	i don`t want her dressed up th	negative	th
1850	8b25e05dff	i hate waiting in lines	i hate wa	negative	wa
2004	31098bb260	wishes he had realized his wife hadn`t held on...	his wife hadn`t held onto the debit card be	negative	be
2324	9ca53f2fb2	okay the man with the hook for a hand is kinda...	is kinda freaking me out ri	negative	ri
2671	65a48787a4	hollowbabeshere comes the utter shite bgt i co...	here comes the utter shite	negative	here
3034	8472323928	no so sad about that i`m from malta have you h...	so sad ab	negative	ab
3325	ee5eb5337b	heading to the gym the group of guys that used...	ly sad	negative	ly
3504	d5b33ff5f4	it was just true and you do cause me to having...	ng dirty	negative	ng
4405	ddf77049bc	haha its under so ive got no one to go with cu...	se none of my friends liking	negative	se
5241	4739eb265a	dude i will check again but i couldnt find any...	al CURSE e	negative	al
6113	2cb67e64b4	these dogs are going to die if somebody doesn`...	aam these dogs are going to die if somebody do...	negative	aam
6842	a61820a07f	right CURSE the whole twitter silence experime...	CURSE the whole twitter silence ex	negative	ex
8153	11404ea118	definition of senioritis me about to go to che...	not good cl	negative	cl
8917	f54cfdafa1	happy emox lmao lucky it`s minutes on foot for...	ewww	negative	ewww
8954	eee8f9c860	stuck in traffic on the on the way to costa mesa	stuck in tr	negative	tr
9272	cee8acf2eb	wtf twitter doesnt support messages from my ph...	doesnt support me	negative	me
9454	8aafc1514d	soch hah nooo she is the obly one that can aum...	i hate nick be	negative	be
9535	29dc42f844	i don`t feel so good after eating all that foo...	i don`t feel so good af	negative	af
9801	95f2b72c57	ashley tisdale i mean cuz shes in berlin on ju...	i wanna start cryin right now as	negative	as
9812	5dbe01a708	hates these khaki pants project to do todayy	hates th	negative	th
10092	c9ea30009c	ï¿½ï¿½ ï¿½ï¿½ i bet man i wish i coulda went s...	nd nyt life sux	negative	nd
12067	0fcb5f9931	up is the saddest movie i`ve ever seen	saddest mo	negative	mo
12563	24ad5b316e	sleeping would`ve been home sooner but we acci...	killed ba	negative	ba
12732	633e9e07ff	i did and i feel great but i still miss it	ll miss	negative	ll
13381	a206385e93	rest in peace marshall	rest in peace ma	negative	ma
13637	d83fd6c942	tweeets fgs tweekdeckkk hates me cryyyy	kk hates me cryyyy	negative	kk
13674	8607d4de1a	dans public transport again and have decided i...	utter CURSE th	negative	th
14551	5d5178af59	you can get into canada but i can`t wtf seriou...	ly i`m not allowed to cross the border	negative	ly
14986	708e67409c	bhb i went to that concert and i remember dere...	miss se	negative	se
15207	4b3fd4ec22	last day at dma over a million sad faces	ion	negative	ion
15940	0d1a051f0a	i`m pondering lunch at shane`s i think i can a...	whining ab	negative	ab
16468	061c01cc9b	is still pretty depressed about losing her hel...	is still pretty depressed ab	negative	ab
16493	f07e9d3c3e	this is not a happy tweet	this is not a happy tw	negative	tw
16777	98a7e00c8b	dammit i cant watch stadium music	dammit i cant watch st	negative	st
16798	5060d948d7	im a bad blogger i have not blogged in weeks oops	bad bl	negative	bl
17013	3fe6f7f128	is not a happy bunny	not a happy bu	negative	bu
18548	d4d8e06110	ugh i can`t access through my mobile web	ugh i can`t ac	negative	ac
19057	a690e02d3c	it`s gloomy as hell outside today	hell ou	negative	ou
19724	1e8548b565	i am very sad because i have gone on the show ...	i am very sad be	negative	be
19904	e772b4ccc1	dreambears were CURSE compared to their wicked...	CURSE co	negative	co
20078	52ca74468a	we`re idiots ok mostly i was skint but hell i ...	re idiots	negative	re
20778	9c8e56c60a	i`m so confused about the weather is it really...	i`m so confused ab	negative	ab
20851	b8c9ac356f	how do i get my cat killin rabbits another hea...	how do i get my cat killin ra	negative	ra
21339	9dd55e8b0e	how are you convinced that i have always wante...	i think i just lost an	negative	an
21632	933e0c5446	morning tweeple i`m a bit sneezy today	sneezy to	negative	to
21876	c18f679ceb	moving back home today pro obnoxiously closer ...	obnoxiously cl	negative	cl
22378	84c5466055	feels all kinds of not so well right now	not so well ri	negative	ri
24490	1d5066fb72	shoutz the mix on the site is gonna b nervvoouuss	na b nervvoouuss	negative	na
24886	c0dfcf858d	the northern clemency but i don`t recommend it...	on too slow	negative	on
25495	a24a95ffed	i`ve become one of those pathetic girls that f...	miss jo	negative	jo
26830	24d37f9ba7	my bracelet broke today too	bracelet br	negative	br
27209	30a1e8c2b4	guys i know my ability to read time telling de...	es failed	negative	es
27302	90c8aa60db	have i ever told you i absolutly hate writing ...	hate wr	negative	wr
27456	d32efe060f	i wanna leave work already not feelin it	wanna leave work al	negative	al

In [53]:
```
def clean_sel_text(x):
    text,sel_text,misspelled = x[0],x[1],x[2]
    sel_text = sel_text.split()
    if (len(sel_text)>1):
        sel_text.remove(misspelled)
    return " ".join(sel_text)
```

In [54]:
```
train_df['selected_text']=train_df[['text','selected_text','misspelled']].progress_apply(lambda x: clean_sel_text(x)
                                                    if (x['misspelled']) != 'No' else x['selected_text'], axis=1)
```

In [55]:
```
train_df['misspelled'] = train_df[['text','selected_text']].progress_apply(misspelled_words,axis=1)
train_df[(train_df.misspelled != 'No')]
```

Out[55]:

	textID	text	selected_text	sentiment	misspelled
3622	792063a20e	geronimo thanks for sharing with your friends	imo	positive	imo
5189	9df7f02404	cracking myself more more up phootoboothingisf...	sfunf	positive	sfunf
8917	f54cfdafa1	happy emox lmao lucky it`s minutes on foot for...	ewww	negative	ewww
15207	4b3fd4ec22	last day at dma over a million sad faces	ion	negative	ion
21376	eaf2942ee8	bud light up in massachusetts and no boston la...	ht	positive	ht

We still have a few misspelled examples which we would fix manually

In [56]:
```
train_df.loc[3622].selected_text = 'geronimo thanks'
train_df.loc[15207].selected_text = 'million sad faces'
train_df.drop(index=[5189,8917,21376],inplace=True)
```

In [57]:
```
train_df.isnull().sum()
```

Out[57]:
```
textID           0
text             0
selected_text    0
sentiment        0
misspelled       0
dtype: int64
```

In [58]:
```
train_df[train_df.selected_text == '']
```

Out[58]:

	textID	text	selected_text	sentiment	misspelled
8729	12f21c8f19	star wars is CURSE boo i wanna do your job han...		positive	No

In [59]:
```
train_df.drop(index=8729,inplace=True)
```

In [60]:
```
train_df['misspelled'] = train_df[['text','selected_text']].progress_apply(misspelled_words,axis=1)
train_df[(train_df.misspelled != 'No')]
```

Out[60]:

	textID	text	selected_text	sentiment	misspelled

After all the cleaning and preprocessing steps done, we have fixed all misspelled texts/phrases in selected_text columns.

Saving all the preprocessed text into a csv file

In [61]:
```
train_df.shape,test_df.shape
```

Out[61]:
```
((27468, 5), (3533, 3))
```

In [62]:
```
train_df.to_csv('preprocessed_train.csv',index=False)
```

In [63]:
```
df = pd.read_csv('preprocessed_train.csv')
df.head()
```

Out[63]:

	textID	text	selected_text	sentiment	misspelled
0	cb774db0d1	i`d have responded if i were going	i`d have responded if i were going	neutral	No
1	549e992a42	sooo sad i will miss you here in san diego	sooo sad	negative	No
2	088c60f138	my boss is bullying me	bullying me	negative	No
3	9642c003ef	what interview leave me alone	leave me alone	negative	No
4	358bd9e861	sons of CURSE why couldn`t they put them on th...	sons of CURSE	negative	No

In [64]:
```
df.shape
```

Out[64]:
```
(27468, 5)
```

In [65]:
```
test_df.to_csv('preprocessed_test.csv',index=False)
```

In [66]:
```
df = pd.read_csv('preprocessed_test.csv')
df.head()
```

Out[66]:

	textID	text	sentiment
0	f87dea47db	last session of the day	neutral
1	96d74cb729	shanghai is also really exciting precisely sky...	positive
2	eee518ae67	recession hit veronique branquinho she has to ...	negative
3	01082688c6	happy bday	positive
4	33987a8ee5	i like it	positive

In [67]:
```
df.shape
```

Out[67]:
```
(3533, 3)
```

In [68]:
```
train_df = pd.read_csv('preprocessed_train.csv')
train_df.loc[1375].selected_text = 'cool'
train_df.loc[2649].selected_text = 'a good'
train_df.loc[2748].selected_text = 'sad that'
train_df.loc[3805].selected_text = 'a fake competetion'
train_df.loc[4162].selected_text = 'i miss'
train_df.loc[4324].selected_text = 'was hoping'
train_df.loc[4472].selected_text = 'cute though'
train_df.loc[6375].selected_text = 'sounds amazing'
train_df.to_csv('preprocessed_train.csv',index=False)
```