```
1. Data Acquisition
2. Text Preparation
2.1 text cleanup
a. html tag
b. emoji
c. spelling checking

2.2 Basic Preprocessing
a basic
-a.1 tokenization
--A-Sentance
--B-word

2.3 optimal/ optional
-a.1 stop words remover
-b.2 stemming  or Lemmatizationn
-c.3 removing digits and punctuation
-d.4 lower case
-e.5 language detection

2.3 Advance Preprocessing
-POS tagging
-Parsing
-core    resolution

3. Feature Engineering


4. Modeling
4.1 Model Building
4.2 Evaluation

5 Devloyment
5.1 Deployment
5.2 Monitoring
5.3 Model Update
```

1. Data Acquisition

1. https://archive.ics.uci.edu/ml/index.php (https://archive.ics.uci.edu/ml/index.php)
2. https://medium.com/swlh/top-20-websites-for-machine-learning-and-data-science-d0b113130068 (https://medium.com/swlh/top-20-websites-for-machine-learning-and-data-science-d0b113130068)
3. https://medium.com/codex/how-to-collect-data-for-a-machine-learning-model-2b152752a15b (https://medium.com/codex/how-to-collect-data-for-a-machine-learning-model-2b152752a15b)

https://www.kaggle.com/datasets?tags=13204-NLP (https://www.kaggle.com/datasets?tags=13204-NLP)

2. Text Preparation

- https://medium.com/product-ai/text-preprocessing-in-python-steps-tools-and-examples-bf025f872908 (https://medium.com/product-ai/text-preprocessing-in-python-steps-tools-and-examples-bf025f872908)

2.1 text cleanup -- a. html tag, b. emoji, c. spelling checking etc.

In [1]:
```
# https://pynative.com/python-regex-compile/#:~:text=Return%20value-,The%20re
```

In [2]:
```
text = '<h2> HTML Element</h2><p> <The class"w3> ,glt;supsgt'
```

In [3]:
```
#Removing HTML Tags
import re
def html(data):
    p=re.compile('(<.*?>)')
    return p.sub('',data)
```

In [4]:
```
html(text)
#html(data)
```

Out[4]:
```
' HTML Element  ,glt;supsgt'
```

In [5]:
```
#b. emoji
emoji='😀💁 People • 🐻🌻 Animals • 🍔🍹 Food • 🎷⚽ Activities • 🚘🌇 T
emoji.encode('utf-8') # converrt in Machine understandable text
```

Out[5]:
```
b'\xf0\x9f\x98\x83\xf0\x9f\x92\x81 People \xe2\x80\xa2 \xf0\x9f\x90\xbb\xf0
\x9f\x8c\xbb Animals \xe2\x80\xa2 \xf0\x9f\x8d\x94\xf0\x9f\x8d\xb9 Food \xe
2\x80\xa2 \xf0\x9f\x8e\xb7\xe2\x9a\xbd\xef\xb8\x8f Activities \xe2\x80\xa2
\xf0\x9f\x9a\x98\xf0\x9f\x8c\x87 Travel \xe2\x80\xa2 \xf0\x9f\x92\xa1\xf0\x
9f\x8e\x89 Objects \xe2\x80\xa2 \xf0\x9f\x92\x96\xf0\x9f\x94\xa3 Symbols \x
e2\x80\xa2 \xf0\x9f\x8e\x8c\xf0\x9f\x8f\xb3\xef\xb8\x8f\xe2\x80\x8d\xf0\x9f
\x8c\x88 Flags'
```

In [6]:
```
pip install textblob
```

```
Requirement already satisfied: textblob in c:\programdata\anaconda3\lib\sit
e-packages (0.17.1)
Requirement already satisfied: nltk>=3.1 in c:\programdata\anaconda3\lib\si
te-packages (from textblob) (3.6.5)
Requirement already satisfied: click in c:\programdata\anaconda3\lib\site-p
ackages (from nltk>=3.1->textblob) (8.0.3)
Requirement already satisfied: joblib in c:\programdata\anaconda3\lib\site-
packages (from nltk>=3.1->textblob) (1.1.0)
Requirement already satisfied: regex>=2021.8.3 in c:\programdata\anaconda3
\lib\site-packages (from nltk>=3.1->textblob) (2021.8.3)
Requirement already satisfied: tqdm in c:\programdata\anaconda3\lib\site-pa
ckages (from nltk>=3.1->textblob) (4.62.3)
Requirement already satisfied: colorama in c:\programdata\anaconda3\lib\sit
e-packages (from click->nltk>=3.1->textblob) (0.4.4)
Note: you may need to restart the kernel to use updated packages.

WARNING: Ignoring invalid distribution -quests (c:\programdata\anaconda3\li
b\site-packages)
WARNING: Ignoring invalid distribution - (c:\programdata\anaconda3\lib\site
-packages)
WARNING: Ignoring invalid distribution -equests (c:\programdata\anaconda3\l
ib\site-packages)
WARNING: Ignoring invalid distribution -quests (c:\programdata\anaconda3\li
b\site-packages)
WARNING: Ignoring invalid distribution - (c:\programdata\anaconda3\lib\site
-packages)
WARNING: Ignoring invalid distribution -equests (c:\programdata\anaconda3\l
ib\site-packages)
WARNING: Ignoring invalid distribution -quests (c:\programdata\anaconda3\li
b\site-packages)
WARNING: Ignoring invalid distribution - (c:\programdata\anaconda3\lib\site
-packages)
WARNING: Ignoring invalid distribution -equests (c:\programdata\anaconda3\l
ib\site-packages)
WARNING: Ignoring invalid distribution -quests (c:\programdata\anaconda3\li
b\site-packages)
WARNING: Ignoring invalid distribution - (c:\programdata\anaconda3\lib\site
-packages)
WARNING: Ignoring invalid distribution -equests (c:\programdata\anaconda3\l
ib\site-packages)
WARNING: Ignoring invalid distribution -quests (c:\programdata\anaconda3\li
b\site-packages)
WARNING: Ignoring invalid distribution - (c:\programdata\anaconda3\lib\site
-packages)
WARNING: Ignoring invalid distribution -equests (c:\programdata\anaconda3\l
ib\site-packages)
WARNING: Ignoring invalid distribution -quests (c:\programdata\anaconda3\li
b\site-packages)
WARNING: Ignoring invalid distribution - (c:\programdata\anaconda3\lib\site
-packages)
WARNING: Ignoring invalid distribution -equests (c:\programdata\anaconda3\l
ib\site-packages)
```

In [7]:
```
# c. spelling checker\
spelling_mistakes= 'hello, how aree you, what aree you doing these days'
```

In [8]:
```
from textblob import TextBlob
```

In [9]:
```
correct_spelling=TextBlob(spelling_mistakes)
```

In [10]:
```
correct_spelling
```

Out[10]:
```
TextBlob("hello, how aree you, what aree you doing these days")
```

In [11]:
```
correct_spelling.correct()
```

Out[11]:
```
TextBlob("hello, how are you, what are you doing these days")
```

2.2 Basic Preprocessing

a basic -a.1 tokenization --A-Sentance --B-word

In [12]:
```
data='''Text-based emoticons are great because they work on most devices
and browsers. Whether you’re commenting on a post or texting with a friend, t
handy if your perfect response isn’t a word but an emotion.'''
```

In [13]:
```
from nltk.tokenize import sent_tokenize, word_tokenize
sents=sent_tokenize(data)
```

In [14]:
```
sents
```

Out[14]:
```
['Text-based emoticons are great because they work on most devices \nand br
owsers.',
 'Whether you’re commenting on a post or texting with a friend, this page m
ay come in\nhandy if your perfect response isn’t a word but an emotion.']
```

In [15]:
```
word=word_tokenize(data)
word
```

Out[15]:
```
['Text-based',
 'emoticons',
 'are',
 'great',
 'because',
 'they',
 'work',
 'on',
 'most',
 'devices',
 'and',
 'browsers',
 '.',
 'Whether',
 'you',
 "'",
 're',
 'commenting',
 'on',
 'a',
 'post',
 'or',
 'texting',
 'with',
 'a',
 'friend',
 ',',
 'this',
 'page',
 'may',
 'come',
 'in',
 'handy',
 'if',
 'your',
 'perfect',
 'response',
 'isn',
 "'",
 't',
 'a',
 'word',
 'but',
 'an',
 'emotion',
 '.']
```

2.3 optimal/ optional

- a.1 stop words remover

- https://www.ibm.com/docs/en/watson-explorer/11.0.0?topic=analytics-stop-word-removal (https://www.ibm.com/docs/en/watson-explorer/11.0.0?topic=analytics-stop-word-removal)
- b.2 stemming or Lemmatization
- c.3 removing digits and punctuation
- d.4 lower case

In [16]:
```
# a.1 stop words remover
# https://www.tutorialspoint.com/python_text_processing/python_remove_stopwor
# https://www.geeksforgeeks.org/removing-stop-words-nltk-python/
import nltk
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\omkan\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[16]:
```
True
```

In [17]:
```
from nltk.corpus import stopwords
stopwords = set(stopwords.words('english'))
```

In [18]:
```
words=[]
for w in word:
    if w not in stopwords:
        print(w)
        words.append(w)
```

```
Text-based
emoticons
great
work
devices
browsers
.
Whether
’
commenting
post
texting
friend
,
page
may
come
handy
perfect
response
’
word
emotion
.
```

In [19]:
```
words
```

Out[19]:
```
['Text-based',
 'emoticons',
 'great',
 'work',
 'devices',
 'browsers',
 '.',
 'Whether',
 ',',
 'commenting',
 'post',
 'texting',
 'friend',
 ',',
 'page',
 'may',
 'come',
 'handy',
 'perfect',
 'response',
 ',',
 'word',
 'emotion',
 '.']
```

- b.2 stemming or Lemmatization

- https://www.analyticsvidhya.com/blog/2022/06/stemming-vs-lemmatization-in-nlp-must-know-differences/#:~:text=Stemming%20is%20a%20process%20that,'%20would%20return%20'((https://www.analyticsvidhya.com/blog/2022/06/stemming-vs-lemmatization-in-nlp-must-know-differences/#:~:text=Stemming%20is%20a%20process%20that,'%20would%20return%20'(
- https://www.guru99.com/stemming-lemmatization-python-nltk.html (https://www.guru99.com/stemming-lemmatization-python-nltk.html)
- https://www.geeksforgeeks.org/introduction-to-nltk-tokenization-stemming-lemmatization-pos-tagging/ (https://www.geeksforgeeks.org/introduction-to-nltk-tokenization-stemming-lemmatization-pos-tagging/)
- https://www.tutorialspoint.com/natural_language_toolkit/natural_language_toolkit_stemming (https://www.tutorialspoint.com/natural_language_toolkit/natural_language_toolkit_stemmin
- https://towardsdatascience.com/stemming-vs-lemmatization-2daddabcb221 (https://towardsdatascience.com/stemming-vs-lemmatization-2daddabcb221)

In [20]:
```
from nltk.stem import PorterStemmer
# create an object of class PorterStemmer
porter = PorterStemmer()
print(porter.stem("play"))
print(porter.stem("playing"))
print(porter.stem("plays"))
print(porter.stem("played"))
```

```
play
play
play
play
```

In [21]:
```
from nltk.stem import PorterStemmer
# create an object of class PorterStemmer
porter = PorterStemmer()
print(porter.stem("Communication"))
```

```
commun
```

https://towardsdatascience.com/stemming-vs-lemmatization-2daddabcb221 (https://towardsdatascience.com/stemming-vs-lemmatization-2daddabcb221)

In [22]:
```
import nltk
from nltk.stem.porter import *
p_stemmer = PorterStemmer()
for word in words:
    print(word+' --> '+p_stemmer.stem(word))
```

```
Text-based --> text-bas
emoticons --> emoticon
great --> great
work --> work
devices --> devic
browsers --> browser
. --> .
Whether --> whether
’ --> ’
commenting --> comment
post --> post
texting --> text
friend --> friend
, --> ,
page --> page
may --> may
come --> come
handy --> handi
perfect --> perfect
response --> respons
’ --> ’
word --> word
emotion --> emot
. --> .
```

In [23]:
```
import nltk
from nltk.stem.porter import *
p_stemmer = PorterStemmer()
for word in words:
    print(p_stemmer.stem(word))
```

```
text-bas
emoticon
great
work
devic
browser
.
whether
’
comment
post
text
friend
,
page
may
come
handi
perfect
respons
’
word
emot
.
```

In [24]:
```
import nltk
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
for word in words:
    print(word+' --> '+lemmatizer.lemmatize(word))
```

```
Text-based --> Text-based
emoticons --> emoticon
great --> great
work --> work
devices --> device
browsers --> browser
. --> .
Whether --> Whether
’ --> ’
commenting --> commenting
post --> post
texting --> texting
friend --> friend
, --> ,
page --> page
may --> may
come --> come
handy --> handy
perfect --> perfect
response --> response
’ --> ’
word --> word
emotion --> emotion
. --> .
```

In [25]:
```
import nltk
words_l=[]
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
for word in words:
    print(lemmatizer.lemmatize(word))
    words_l.append(word)
```

```
Text-based
emoticon
great
work
device
browser
.
Whether
’
commenting
post
texting
friend
,
page
may
come
handy
perfect
response
’
word
emotion
.
```

In [26]:
```
words_l
```

Out[26]:
```
['Text-based',
 'emoticons',
 'great',
 'work',
 'devices',
 'browsers',
 '.',
 'Whether',
 '’',
 'commenting',
 'post',
 'texting',
 'friend',
 ',',
 'page',
 'may',
 'come',
 'handy',
 'perfect',
 'response',
 '’',
 'word',
 'emotion',
 '.']
```

c.3 removing digits and punctuation

- https://www.programiz.com/python-programming/examples/remove-punctuation (https://www.programiz.com/python-programming/examples/remove-punctuation)

In [27]:
```
# define punctuation
punctuations = '''!()-[]{};:'"\,<>./?@#$%^&*_~'''

# remove punctuation from the string
no_punct = ""
for char in words_l:
    if char not in punctuations:
        no_punct = no_punct + char

# display the unpunctuated string
print(no_punct)
```

```
Text-basedemoticonsgreatworkdevicesbrowsersWhether’commentingposttextingfri
endpagemaycomehandyperfectresponse’wordemotion
```

In [28]:
```
digit_1_10='''Text-bas19edemoticonsgreatworkdevi11cesbrowser12sWhether’commen
4gposttex16ti17gfriendpag366756756818emayco7376575619mehand7658679yperfectres

digits='''0-1'''
# remove punctuation from the string
no_digits = ""
for char in words_l:
    if char not in digits:
        no_digits = no_digits + char

# display the unpunctuated string
print(no_digits)
```

```
Text-basedemoticonsgreatworkdevicesbrowsers.Whether’commentingposttextingfr
iend,pagemaycomehandyperfectresponse’wordemotion.
```

lowercase

- https://www.oreilly.com/library/view/python-natural-language/9781787121423/9742008f-6384-42a4-9711-2721dd6fd382.xhtml (https://www.oreilly.com/library/view/python-natural-language/9781787121423/9742008f-6384-42a4-9711-2721dd6fd382.xhtml)

In [29]:
```
def wordlowercase():
    text='An Easy Way To Change Uppercase to Lowercase And Title Capitalizati
    return text.lower()
```

In [30]:
```
wordlowercase()
```

Out[30]:
```
'an easy way to change uppercase to lowercase and title capitalization'
```

2.3 Advance Preprocessing

- POS tagging - Note: don't does after stowords
- Parsing
- Coreference resolution- is the task of finding all expressions that refer to the same entity in a text.
- https://www.mygreatlearning.com/blog/pos-tagging/#sh1 (https://www.mygreatlearning.com/blog/pos-tagging/#sh1)
- https://byteiota.com/pos-tagging/ (https://byteiota.com/pos-tagging/)
- https://www.geeksforgeeks.org/nlp-part-of-speech-default-tagging/ (https://www.geeksforgeeks.org/nlp-part-of-speech-default-tagging/)

NLP Classification

- The Official Twitter account for the Myers-Briggs (#MBTI) personality assessment, published by The Myers-Briggs Company

In [31]:
```
# import liberary
import numpy as np
import pandas as pd
# read data set
df=pd.read_csv('twitter_MBTI.csv')
df.head()
```

Out[31]:

	Unnamed: 0	text	label
0	0	@Pericles216 @HierBeforeTheAC @Sachinettiyil T...	intj
1	1	@Hispanthicckk Being you makes you look cute||...	intj
2	2	@Alshymi Les balles sont réelles et sont tirée...	intj
3	3	I'm like entp but idiotic|||Hey boy, do you wa...	intj
4	4	@kaeshurr1 Give it to @ZargarShanif ... He has...	intj

In [32]:
```
y=df['label']
```

MBTI stands for Myers Briggs Type Indicator. This is a tool which is frequently used to help individuals understand their own communication preference and how they interact with others.

https://clockify.me/blog/managing-time/productivity-based-on-personality-type/ (https://clockify.me/blog/managing-time/productivity-based-on-personality-type/)

https://www.crystalknows.com/personality-type/relationship (https://www.crystalknows.com/personality-type/relationship)

-https://www.16personalities.com/personality-types (https://www.16personalities.com/personality-types)

https://www.psychologyjunkie.com/heres-the-flower-youd-be-based-on-your-myers-briggs-personality-type/ (https://www.psychologyjunkie.com/heres-the-flower-youd-be-based-on-your-myers-briggs-personality-type/)

https://www.psychologyjunkie.com/discover-your-true-personality-type-free-test-included/ (https://www.psychologyjunkie.com/discover-your-true-personality-type-free-test-included/)

In [33]:
```
df['label'].value_counts(),df['label'].nunique()
```

Out[33]:
```
(infp    1282
 infj    1057
 intp     811
 intj     781
 enfp     729
 entp     577
 enfj     518
 isfp     367
 isfj     364
 istp     327
 entj     279
 istj     259
 esfp     174
 esfj     105
 estp     100
 estj      81
 Name: label, dtype: int64,
 16)
```

In [34]:
```
from sklearn.preprocessing import LabelEncoder
lb=LabelEncoder()
y=lb.fit_transform(y)
y
```

Out[34]:
```
array([10, 10, 10, ...,  3,  8, 15])
```

In [35]:
```
df.shape
```

Out[35]:
```
(7811, 3)
```

In [36]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7811 entries, 0 to 7810
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Unnamed: 0  7811 non-null   int64 
 1   text        7811 non-null   object
 2   label       7811 non-null   object
dtypes: int64(1), object(2)
memory usage: 183.2+ KB
```

In [37]:
```
# Lowercase for 4rth row
df['text'][4].lower()
```

Out[37]: "@kaeshurr1 give it to @zargarshanif ... he has pica since childhood|||@dannnyaaaa say qubool hai in my dm ❤️|||@dannnyaaaa get married asap|||@ubaid_rasool two decades actually 🤔|||@kaeshurr1 @umer__ayoub subas ?|||@kaeshurr1 @umer__ayoub wayne kar niyiv meati panas seyt|||@kaeshurr1 @umer__ayoub mea gov akh hafte czeaf dith|||@kaeshurr1 @umer__ayoub meati niyziha panas seyt|||okay 🥺 https://t.co/3fh1tqbfl2|||@hareembhat_ (https://t.co/3fh1tqbfl2|||@hareembhat_) https://t.co/0nve1cdbqd|||@pingaapongaa (https://t.co/0nve1cdbqd|||@pingaapongaa) hye|||@bint_e_tasleem i had a lot to speak out, but you ignored me 🥺|||@whytaniya welcome back|||just a simple thought, what's stopping from having icu facilities at dh shopian? 50-70% deaths occur due to referra… https://t.co/wpcgmdzsar|||@toyibanabi (https://t.co/wpcgmdzsar|||@toyibanabi) laxmi chit fund bank|||same bro same 🥺🤔 https://t.co/2g5zorfhsf|||@yawar_banday (https://t.co/2g5zorfhsf|||@yawar_banday) expecting a grand opening soon in sha allah???|||@dannnyaaaa aameeeeeeeennnn|||@babra_yousuf mohabat|||@yawar_banday thanks for bringing the coolest place ever seen ❤️|||@aadilsam118 mujhe b le jao tandoori chicken khane...🥺|||#newprofilepic https://t.co/dcilyzkxkv|||@muskan_parhlo (https://t.co/dcilyzkxkv|||@muskan_parhlo) @2 @penstagrammer 🥺|||monro kellie hypothesis 🥺 https://t.co/olf8wldpuj|||@ruqayawar (https://t.co/olf8wldpuj|||@ruqayawar) quch healthy sa khaya karo didi\nkamzori bohot hai apko|||medicos use twitter and snapchat just to show their books, apron, stethoscope and college.\n#crowning|||@bashir_zarnab ye lo https://t.co/wexbr3j8dr|||@urvaak_ (https://t.co/wexbr3j8dr|||@urvaak_) @kaeshurr1 bull means saand 🥺|||@kaeshurr1 bell|||@kaeshurr1 hil bill 🥺|||@mir_reyazz badshah|||@shazilturey bhai toilet mai karleta|||@aadilsam118 looking new pandemic|||@d_otherkhaleesi effects of ignoring my hi 🥺|||@amnaghaffar777 me n who 🥺|||@waseems72520362 bellisa champ|||@penstagrammer inviting ?|||@bleeehh_ kaise bro kaise|||@penstagrammer https://t.co/wkwqccfdcf|||@_____alif (https://t.co/wkwqccfdcf|||@_____alif) get well soon 🙂|||@khargooooooosssh ok bye 👋|||@saniazehra123 ok|||@arshiequreshi bring some to delhi too 🥺|||@umi07khan @shaziltu rey hahahahaha bhai he's the best poser|||@shazilturey 13 khoon maaf|||@mehnaz077 pehle he bola tha 🥺|||@kaeshurr1 mumbai wala good job leaving leaving kese join karega 🥺|||@_cheese_corn @kalam_ki_noke 😊|||@__ali07 @2 @m_zainab_k 🥺|||@maham_32 @amnaghaffar777 astagfirullah 🙄|||@amnaghaffar777 @baezaarrr same dua for me too bro 🥺|||@kaeshurr1 bhai time travel kar k aaya mai tere bache khelte dikh gaye 😪|||@kaeshurr1 ok bye 👋 https://t.co/teihtevd2m|||@shinvanuk (https://t.co/teihtevd2m|||@shinvanuk) https://t.co/racxo2amr4|||@giggle_44 (https://t.co/racxo2amr4|||@giggle_44) faridabad is green|||@giggle_44 no|||@umi07khan @_cheese_corn https://t.co/lgpeyrtulx|||@_cheese_corn (https://t.co/lgpeyrtulx|||@_cheese_corn) send over 😪|||somewhere in world, someone is having toothache, headache, earache. some woman is undergoing labour pain, some male… https://t.co/zb7s9yq5du|||@giggle_44 (https://t.co/zb7s9yq5du|||@giggle_44) dil wale nahi hai west delhi wale|||@kaeshurr1 hahahahaha you're his secret admirer 🤣🤣🤣|||@kaeshurr1 your workout reason 🤣 https://t.co/jkdorof4ih|||@jutti__hun (https://t.co/jkdorof4ih|||@jutti__hun) @umi07khan kya bolte ho bhai|||@impov same bro same 🥺|||@nawab__zadiii @sakon_e_qalb naya lelo 🥺|||@uzma_nisar @warofchains masha allah|||@ammar_xaidi @1st @heya_ambivert 🥺|||@___maida__ wo kya hota hai 🥺|||@sarahsa62844321 text me once you reach home doc 🥺|||@kaiiiisus @ladygaga 🥺|||in aankhon ki masti mai .... dark circles hazaroon hai 🥺\n#darkcirclesarehot|||@giggle_44 sikhao 🤔|||@giggle_44 what's cp 🥺|||@giggle_44 you can have 9 after 9 months 🤔|||they talk about golden or diamond rings.\ni talk about fleischer ring, weiss ring and vossious ring.\n#zlife \U0001f979|||@toyibanabi credits \U0001f979|||@sadiaakbarrr thanks (3)|||@hudaibaj left hand \U0001f979|||@bint_e_tasleem tajamul soab \U0001f979|||@_cheese_corn mirzapur|||@kifayat677671 saal kar chu kheon🤔|||@kifayat677671 mubarak|||@giggle_44 @ranamishka tumhe kab mili|||@d_otherkhaleesi astagfir

ullah|||@d_otherkhaleesi https://t.co/vbp0iwejee|||@d_otherkhaleesi (http
s://t.co/vbp0iwejee|||@d_otherkhaleesi) alhamdulillah you don't talk anymor
e|||@d_otherkhaleesi no it's your current relationship status|||@d_otherkha
leesi alhamdulillah|||@muskurategham +1|||@d_otherkhaleesi @gowhar_ you're
29+ 😊|||@bhavyaxoxo @elonmusk bhai batao inko \U0001f979|||@bhavyaxoxo oka
y|||@bint_e_tasleem \U0001f979|||@_dabosschick chalo meaning samjhao sabka
🥺|||life was all good then \naphagia, asphyxia, apraxia, aphasia, ageusia,
dysphagia, dysarthria, dyspraxia happened. 🥺\n#frigid_thoughts.|||@hudaiba
j shaatiraana harkatein by @kaeshurr1 🥺|||@giggle_44 ye bhi theek hai 🥺|
||@na_jao_na @1st @mir_reyazz|||@bakalhafsa @jandktourism @kashmir_weather
@jktourism_corpn mai bhi 🥺🥺🥺|||@mir_reyazz @sadiyabhat_ @shazilturey i
t's haraam to text gair mehram 🥺|||@giggle_44 https://t.co/f72e8dkok3|||@g
iggle_44 (https://t.co/f72e8dkok3|||@giggle_44) ghar k bahar wali under 18
bachu ki park mai jane ko bahar jana bolte hai west delhi mai ? \U0001f979|
||@giggle_44 as if you're hanging out with friends. you just tweeted while
brushing 🤭|||for pediatricians, the word baby isn't romantic at all. \nthe
y will end up thinking whether you're talking about neo… https://t.co/1c5i8
sofpx|||@___maida__ (https://t.co/1c5i8sofpx|||@___maida__) kaha 🤭|||@m_za
inab_k paani bhi garam aaraha gaaaaaawwwaaaizzz 🤭|||@farkhanda_shah je veu
x m'enfuir que tout recommence\noh ma douce souffrance \U0001f979|||@theylu
vvshine 6'1\n27\nwidower|||@nowreeeeeeeen https://t.co/rai0bvalyy|||@nowree
eeeeeen (https://t.co/rai0bvalyy|||@nowreeeeeeeen) https://t.co/gwafdw3ptn|
||@nowreeeeeeeen (https://t.co/gwafdw3ptn|||@nowreeeeeeeen) @giggle_44 http
s://t.co/dexau092zz|||@nowreeeeeeeen (https://t.co/dexau092zz|||@nowreeeeee
een) how much ? \U0001f979|||@giggle_44 tab xxy or xyy hoga so in both case
s extra x and y can come from male partner so de dana dhan \U0001f979|||@gi
ggle_44 you means parents or offspring?|||if you wanna beat someone because
a girl is born, beat your son.\n#girlaareablessing|||men have xy chromosome
s and females have xx chromosomes. female always gives x and it is on the m
ale genome which d… https://t.co/bu1s5v20n3|||@maham_32 (https://t.co/bu1s5
v20n3|||@maham_32) @maybeeevirgo @amnaghaffar777 astagfirullah 🙄|||@sakon_
e_qalb @3rd @afreenshowkat 🙄|||@theyluvvshine replied|||forget bad memorie
s of life just like you forgot muscle attachments in anatomy and drug class
ification in pharma. \U0001f979… https://t.co/jbca3bsjrl|||@altafg22 (http
s://t.co/jbca3bsjrl|||@altafg22) @4th @14harneet \U0001f979|||@kaeshurr1 ht
tps://t.co/u5qvjdpfc2|||@kaeshurr1 (https://t.co/u5qvjdpfc2|||@kaeshurr1) a
s if you've plans for weekend \U0001f979|||@toyibanabi @youtube give her ad
free subscription \U0001f979|||@aadilsam118 married tulips of shopian|||@gi
ggle_44 padhai likhayi kro ashu\nia y a s karo|||@kaeshurr1 😂|||@giggle_44
show off 🥺🥺🥺|||@kaeshurr1 pet mai kitne jayege bhai ... one tree produc
es 500 boxes on average|||@kaeshurr1 true that. but it's shelf life sucks||
|@nowreeeeeeeen @noorhopes @3rd what's bt 😪|||@bint_e_tasleem sending hi f
rom 46° 😊|||@_dabosschick 🥺|||@khanra_tiyasa ok bye 👋|||don't get attac
hed to many people.\nyou're human not humerus that can hold 13 muscle attac
hments alone.🤭\n#mentalhealthawarenessweek|||@seydquraiba @sonusood bhai i
ske liye flight book krdo 😪|||@d_otherkhaleesi not again / not against
😪|||@umi07khan @shazilturey kanijung ka khatra 😪|||@nad1an4v33d not me
😵|||@shazilturey looking omicron|||@noushibahilal i'm scared of email atta
chments 😵|||@syed_shoiabb @minicotinee ok bhaijaan|||@minicotinee @syed_sh
oiabb toh mai best friend b ni hu 🙄|||@minicotinee mai kya hun fir 🧐|||@
dissociativeee theek hai|||@_illusion98 your nails and pant are twinning
😊|||@minicotinee order from @amazonin|||@_dabosschick https://t.co/yklxw26
q76|||@seydquraiba (https://t.co/yklxw26q76|||@seydquraiba) aunts 🤭|||@mus
segareeb_ welcome to india|||@zaira_aaa @nowreeeeeeeen \U0001f979|||@haddha
i_ replied|||treat yourself like someone you loved \U0001f979\n#mood|||@its
bitchma you're pretty|||@itsbitchma ok|||@afaxima masha allah|||@m_zainab_k
congratulations 👏|||@afaxima mujhe bhi sath lejate \U0001f979|||@mir_reyaz

```
z @sheen_piipin wadnas ruduy na waar|||@bint_e_tasleem shall i recite bismi
llah and dm you then|||@mir_reyazz @sheen_piipin kya hamle kornay mulazim s
oaba \U0001f979\U0001f979|||@bint_e_tasleem i always fantasise dming you. b
ut then i reject myself and stop fantasising even. \ni've this inferiority
complex.|||@asraafarooque we can be anything? 🙄|||@bint_e_tasleem reply my
all pending dms|||@umi07khan @nun_chaai @isaguha while working in your appl
e orchard may be \U0001f979|||@sheen_piipin zakhmi pipin|||@stanfordmbb @ja
meskeefe22 @brandon_angel3 goodluck❤️|||@youlovemeikthat ab may dump k sath
krna|||@afaxima pray for me too|||@ammar_xaidi @2nd @_dabosschick|||@aadils
am118 covid baradari|||@dr_mun24 you want a doll? choti wali ?|||@kheramahi
ra21 kyu \U0001f979|||@kheramahira21 ap krte ho ? \U0001f979|||@eramwani sl
apping is not good|||@nusrat193 beshak|||@rukhsanajalani tajamul islam|||@m
ir_reyazz @_maryambhat_ hahahahaha.... beha govus bore yeti wallah ....|||@
irtiqaayoub keep smiling didi|||@mir_reyazz @_maryambhat_ yi kya waatiy mul
azim soaba ?|||@_maryambhat_ how old are you|||@_cheese_corn separated ?|||
@jaane_bhai_ tajamul___islam|||@minicotinee happy birthday 🎊🎁|||@theycal
lmepoem 👋|||@_illusion98 @sheen_piipin thanks|||@sheen_piipin nice wohvov
\U0001f979|||@sarahafreenmal1 what do you do on weekends?|||@sarahafreenmal
1 still you said no to me ?|||@sarahafreenmal1 sarah afreen. lemme take you
out someday. works ? \U0001f979|||@sarahafreenmal1 so it's fact hot girls d
on't date ugly guys 🙂|||@sarahafreenmal1 agreed mam. so can we date or not
? \U0001f979|||@sarahafreenmal1 so we can still date ?|||@themessybreeze co
ngratulations|||@sarahafreenmal1 you're 35 and hot \ni'm 27 and ugly. \nle
t's date and break the myth\U0001f979|||@fasaadd khilao fir 😋|||@fasaadd h
ttps://t.co/v7qhurrehd|||@shazilturey (https://t.co/v7qhurrehd|||@shaziltur
ey) looking new variant.|||"
```

In [38]:
```
text=df['text']
```

In [39]:
```
# convert in lowercase
text=text.str.lower()
```

In [40]:
```
 # remove html tags
import re
def remove_html_tags(text):
    pattern = re.compile('<.*?>')
    return pattern.sub(r'',text)

# df['text']=df['text'].apply(remove_html_tags)
```

In [41]:
```
text=text.apply(remove_html_tags)
text
```

Out[41]:
```
0       @pericles216 @hierbeforetheac @sachinettiyil t...
1       @hispanthicckk being you makes you look cute||...
2       @alshymi les balles sont réelles et sont tirée...
3       i'm like entp but idiotic|||hey boy, do you wa...
4       @kaeshurr1 give it to @zargarshanif ... he has...
                              ...
7806    @sobsjjun god,,pls take care 😟|||@sobsjjun hir...
7807    @ignis_02 wow last time i got intp https://t.c... (https://t.c...)
7808    @akupilled a 100%|||@akupilled that someone wi...
7809    if you’re #intj this one is for you | what is ...
7810    @harry__lambert @gucci hey can you dm me a pic...
Name: text, Length: 7811, dtype: object
```

In [42]:
```
# clear url
def remove_url(text):
    pattern = re.compile(r'https?://\S+|www\.\s+')
    return pattern.sub(r'', text)
# df['text']=df['text'].apply(remove_url)
```

In [43]:
```
text=text.apply(remove_url)
text
```

Out[43]:
```
0       @pericles216 @hierbeforetheac @sachinettiyil t...
1       @hispanthicckk being you makes you look cute||...
2       @alshymi les balles sont réelles et sont tirée...
3       i'm like entp but idiotic|||hey boy, do you wa...
4       @kaeshurr1 give it to @zargarshanif ... he has...
                              ...
7806    @sobsjjun god,,pls take care 😟|||@sobsjjun hir...
7807    @ignis_02 wow last time i got intp    i think ...
7808    @akupilled a 100%|||@akupilled that someone wi...
7809    if you’re #intj this one is for you | what is ...
7810    @harry__lambert @gucci hey can you dm me a pic...
Name: text, Length: 7811, dtype: object
```

https://www.geeksforgeeks.org/python-remove-punctuation-from-string/ (https://www.geeksforgeeks.org/python-remove-punctuation-from-string/)

In [44]:
```
# remove punctuations - method - 1

import string,time
string.punctuation
```

Out[44]:
```
'!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'
```

In [45]:
```
exclude = string.punctuation
```

In [46]:
```
def remove_punctuation(text):
    for char in exclude:
        text = text.replace(char,'')
    return text
```

In [47]:
```
start = time.time()
print(remove_punctuation(text))
time1 = time.time() - start
print(time1)
```

```
0       @pericles216 @hierbeforetheac @sachinettiyil t...
1       @hispanthicckk being you makes you look cute||...
2       @alshymi les balles sont réelles et sont tirée...
3       i'm like entp but idiotic|||hey boy, do you wa...
4       @kaeshurr1 give it to @zargarshanif ... he has...
                              ...
7806    @sobsjjun god,,pls take care 😟|||@sobsjjun hir...
7807    @ignis_02 wow last time i got intp    i think ...
7808    @akupilled a 100%|||@akupilled that someone wi...
7809    if you’re #intj this one is for you | what is ...
7810    @harry__lambert @gucci hey can you dm me a pic...
Name: text, Length: 7811, dtype: object
0.03207516670227051
```

In [48]:
```
# remove punctuations - method - 2
def remove_punc1(text):
    return text.translate(str.maketrans('','',exclude))
```

In [49]:
```
text=text.apply(remove_punc1)
```

In [50]:
```
start = time.time()
text=text.apply(remove_punc1)
time2 = time.time() - start
print(time2)
```

```
4.269453287124634
```

In [51]:
```
time1/time2
```

Out[51]:
```
0.007512710538138316
```

In [ ]:
```

```

In [52]:
```
# remove punctuations - method - 3
punc = '''!()-[]{};:'"\,<>./?|@#$%^&*_~'''

def remove_punc(text):
    for char in punc:
        text = text.replace(char,'')
    return text
# df['text']=df['text'].apply(remove_punc)
```

In [53]:
```
text=text.apply(remove_punc)
text
```

Out[53]:
```
0       pericles216 hierbeforetheac sachinettiyil the ...
1       hispanthicckk being you makes you look cutethi...
2       alshymi les balles sont réelles et sont tirées...
3       im like entp but idiotichey boy do you want to...
4       kaeshurr1 give it to zargarshanif  he has pica...
                              ...
7806    sobsjjun godpls take care 😟sobsjjun hiro emerg...
7807    ignis02 wow last time i got intp    i think u ...
7808    akupilled a 100akupilled that someone will get...
7809    if you’re intj this one is for you  what is ne...
7810    harrylambert gucci hey can you dm me a pic of ...
Name: text, Length: 7811, dtype: object
```

In [54]:
```
import time
start = time.time()
remove_punc(text)
time3=time.time()-start
print(time3)
```

```
0.028387784957885742
```

In [55]:
```
time3/time2
```

Out[55]:
```
0.0066490445143163
```

```
# spelling checker & correct_string
from textblob import TextBlob
#tb=TextBlob(single text)
#tb.correct().string
def correct_string(text):
    for char in TextBlob:
        text = text.replace(char,'')
    return text
```

```
text=text.apply(correct_string)
```

In [56]:
```
text[7806]
```

Out[56]: 'sobsjjun godpls take care 😕sobsjjun hiro emergency room are you okay wait dafahoe zeharrrrrrr😭😭😭😭😭😭😭😭txtlomlsss ye0nyang oh my god thats 💔ye0nyang txtlomlsss did i liesoobinspolaroid i knownano h—\nsoule grandm a r so cute😭my whole tl just 4 oomfs doing sum shit but i love itsoobins polaroid is this u0x1bts plus i took a test yesterday and i am an intp ur s o annoying0x1bts dude atleast speak in english what if he minds wtfsoobinsp olaroid sobsjjun oh my godsoobinspolaroid sobsjjun who tf tells before stea ling just steal when hes sleeping or sum shit 😭0x1bts soobinspolaroid what the fuck0x1bts soobinspolaroid im intpi should stop making jokes helpbeom thyusiast1 nooooo😭ye0nyang snakskdmkwkwmdmbeomthyusiast1 who said im jokin gyeehawnjun im so funnytxtlomlsss rightt im so geniusox1gyuu sooo truecacti gumi im so funnytxtlomlsss im geniuschoiseeker yesyuniberri yunnie twtstole n btwif a shark bites ubite back youll still d1e but the shark would be li ke yo wtfinsoobinspolaroid i love kids too every kid i meet gets very attac hed to me and idk why😭yeonjun is no ones shadow shut the fuck up loser gu ess whos the anonsoobinspolaroid im intp ✨soobinspolaroid anon ur so funny soobinspolaroid wait fried egg raw egg boiled egg or smthsoobinspolaroid e ggsoobinspolaroid oh ny god samwsoobinspolaroid like yoursoobinspolaroid ye s his name is sunoosoobinspolaroid sunoo lomlsoobinspolaroid vacations righ t same 💀dafahoe helpppbeomgyuspabomoa ok👁️👄👁️man i love him did u try making pancakes againsobsjjun yes lmao im okay0x1bts teray pas aur kuch bol nay ko hi nahi hai 💀0x1bts darne do lawl0x1bts mein fazool baton pe nahi r oti \U0001faf6 sobsjjun buy a lot of snacks and eat a lotsobsjjun yeasssso bsjjun ps5sobsjjun shirt or smthdude me and my cousin were standing on the street and there was a big wild cat running towards us and it was fast a… keep telling me that it wont effect me bc ur not my mom \U0001faf6 forssera txt wtf shes a whole minorhyyhkai exactly💀yawnzzniverted yesrkwnzzn exactl y shes 15 like wtfbtxtrealm ikr likei hate yall leave them alonewho tf is s hipping eunche and 🐻forsseratxt what happened 💀taehyun0x1bts ur a nuguur just like me or let me carry these trash bags for yoy waitgjnzzn omg nood iswifey fucj mathsrekimimi17 no its ewwwwww5oobinluv yes i oened weverse an d boomodiswifey dpes this scare you wtf is fhis history science maths e w maths5oobinluv noo i wont dont worry 😭😭5oobinluv yes we always watch t hrillers 😭5oobinluv yeasss again 😾sobsjjun okay give them to me then 💀o diswifey yay thank u nona sobsjjun he is cooler than u \U0001faf6 odiswifey i dont know yet but i watched the insidious yesterday \U0001faf6 sobsjjun h es cuter than u 😕sobsjjun they do that one baby u posted was so cuteso im watching a horror movie again today 😾sobsjjun and tallsobsjjun ur so coolx cyj gm lovi have a great day good morning yes yesi love u \U0001faf6 john cena hi non oomf pls fb actually u got caught in bl0ckchainsjjuniverses yes my meows i actually did notice but i thought ur sleeping sooooo☹️lol i wa nt this or else ill make one myself 😭😭ningningie nauuryryrytried to pul l my blanket but i hit my face insteadbeozip yawnwdz 2 likeyawnwdz why r ra ndom ppl retweeting my taemin tweets helloyawnwdz the rt 💀yawnwdz wait tae min from shineeyawnwdz i think its shinee from taeminsoobinspolaroid ikr li ke lmao😭the tweet that started it all all know its odi dudebeomgyuspabo moa block him these people r fucking annoyingbeomgyuspabomoa is that some l ocaljafferyyy phir letay rahosaucegamez soobnf4iry ianndior so jao omgiann wanted to collab with them you too lucas 😔soobinspolaroid inoor will n ever let you flop and will defend u till i d1e cringessoobinspolaroid i wil l never let u flop gothic fontsoobinspolaroid luflop jtxtlomlsss same lmaot xtlomlsss helpppi miss yeonjundude yes\U0001faf6 its not on nwtflix omg i love this\n most disturbing movie i ever watchedwhy r the voices m0aning isnt this a horror moviethere are too many jumpscares in insidious im crydi dbokay christians daughters r pretty not gonna liecreepy as fuckd sksim tra umatizedglorysoob jkdkth you guys im watching insidious and im traumatizede lise ur sooh shit they found this whistleyeonkook7 yesthe woman is so creep

ysobsjjun im watching insidious right now \U0001faf6 biych a fucking womab standing there and u dont see itjkdkth right the first few scenes were a bi t scary but its a bit confusing 😭glorysoob omgyou guys im fucking screamin g its so scary oh😭yeonkook7 omg ok im watching right now 😭is insidious scaryguys is orphan scarymagnifityun yes lmskskdk with ny cousin though 😭y eonssim the movie exorcistmagnifityun yes the 1973 one omg i think i should watch ot tonightratio omg is that scaryrkwnzzn yes yes ill wztch this th ank yoult3my mom said its too scary☹️has anyone watched the exorcist ☹️rkw nzzn omg yesis gram innocentsobsjjun oki lets intwract more \U0001faf6 sobs jjun im so short wtf 💀28yeonsthetic ayeeee28yeonsthetic lmfao what happene dis this me ☹️ 164cm'

https://stackoverflow.com/questions/73472173/modulenotfounderror-while-importing-emoji-in-jupyter (https://stackoverflow.com/questions/73472173/modulenotfounderror-while-importing-emoji-in-jupyter)

In [57]:
```
#pip install emoji
```

In [58]:
```
#pip install emoji --upgrade
```

Handle emoji

In [59]:
```
### Handle emoji
import emoji
print(emoji.demojize(text[7606]))
```

guys omg so happy they got to meet again sayangku :loudly_crying_face::loudly_crying_face::loudly_crying_face::loudly_crying_face:wetheboyz sayangi hope their arms is the last thing i see before i sleep and then i will be dreaming about sanghak arms toogoing crazy one chance please give me one chance :face_with_hand_over_mouth::grinning_face_with_smiling_eyes::rolling_on_the_floor_laughing::beaming_face_with_smiling_eyes::face_holding_back_tears::grinning_squinting_face:guys i love haknyeon so muchsanghaakkkkkkkkkkk armsssssssssss:grinning_face_with_smiling_eyes::grinning_face_with_smiling_eyes::grinning_face_with_smiling_eyes::grinning_face_with_smiling_eyes::grinning_face_with_smiling_eyes::grinning_face_with_smiling_eyes::grinning_face_with_smiling_eyes: ilywetheboyz omgwetheboyz juyeonliterally what the fucklord shut up hak ily:grinning_face_with_smiling_eyes::grinning_face_with_smiling_eyes: i love youwetheboyz omgoh my bbangnyuman… am i suppose to move on what baby i love you so muchcravitystarship thank you youngtae :heart_hands::heart_hands::heart_hands:hey god its me again my lovesweetheboyz thank youwetheboyz oh my godbbangnyuanother win for me today as a fengfan and juyeon enthusiast way my god my god :heart_hands::heart_hands::heart_hands::heart_hands::heart_hands::heart_hands: i love him so much yay so proud of everyone today :heart_hands::heart_hands:couldn’t watch dream concert earlier when it was streaming but i just watch cix performance and they served once again :heart_hands::heart_hands:seungyounification of juyeon :heart_hands::heart_hands: i dont miss when cix is on the red carpet im not home and my data is :face_holding_back_tears:i really hope hes doing well boyz served once againsweet :heart_hands::heart_hands::heart_hands::heart_hands::heart_hands:bloom bloom 20 :heart_hands::heart_hands::heart_hands::heart_hands:ist i need you to announce literally right now if the boyz is coming to sg so i can find money and plan my timelomlllllll lomlhappy seeun day lt3 cannot do this today… have hope for the boyz coming to sg plsplsplsplsplsplsnct 127 in sg… so tempted to buy the tickets :loudly_crying_face::loudly_crying_face:love well soon jacob good morning to me omg for juyeon kevin and changmin fast recovery no way omgwhat…because like wtf morning to me what the fuck did i just wake up towetheboyz qtqueen congrats loves so proud of everyoneso proud of them they deserve it so much lt3 omg hi :heart_hands::heart_hands::heart_hands:my love omgwetheboyz goodnight kev lt3spent an hour watching bbangnyu vlive i think life is at its peakvityprnt happy birthday cor doh my i love youwetheboyz youmy bbangnyuwhat the fuckshut up morning to me yay sunwoowetheboyz ilyisttheboyz thank you hyunjaeoh my god world stop omgisttheboyz whati love him my hak yay hakwetheboyz my loveforever 11 lt3okay but no bloom bloom :face_holding_back_tears::face_holding_back_tears:i would have fucking collapsed what proud of thembet on youddd my belovedtheyre the cutest fr actually gonna cry bloom bloom… no bloom bloomthis :face_holding_back_tears::face_holding_back_tears::face_holding_back_tears::face_holding_back_tears::beaming_face_with_smiling_eyes:also the birthday surprise for jacob thats so cutei keep on saying that i will log off for my mental health but the thought of missing out when they are in the red o… have survived this irl red outfits are so prettywhat my chanhee cannot i cannot i cannot checkmate omgi need to see haklord juyeon:face_holding_back_tears: off :loudly_crying_face::loudly_crying_face::loudly_crying_face::loudly_crying_face::loudly_crying_face::loudly_crying_face::loudly_crying_face:may we get a lot of haknyeon pics today or i will screamone day one day they will come to singapore jacob day :red_heart:

따스함으로물든제이콥의스물여섯
ourangeljacobdayhappy jacob day lt3 going to sleep through the whole thebzone concert tmr definitely not because i cannot go to oneokay younghoon has an undercut and i only know about it nowkev hyunjae younghoon :face_holdi

ng_back_tears::face_holding_back_tears::face_holding_back_tears:not now ple
ase why is always when im outside:loudly_crying_face::loudly_crying_face::l
oudly_crying_face::loudly_crying_face: my god prettiest hak hak my favour
ite hakcherry hak wdym wdym wdymmy loves the prettiest mf late to the pin
k chanhee but omg pink chanheewetheboyz i love you so so much :heart_hand
s::heart_hands::heart_hands:jichang ily

In [60]:
```
text[7606].encode('utf-8')
```

Out[60]: b'guys omg so happy they got to meet again sayangku \xf0\x9f\x98\xad\xf0\x9f\x98\xad\xf0\x9f\x98\xad\xf0\x9f\x98\xadwetheboyz sayangi hope their arms is the last thing i see before i sleep and then i will be dreaming about sanghak arms toogoing crazy one chance please give me one chance \xf0\x9f\xa4\xad\xf0\x9f\x98\x84\xf0\x9f\xa4\xa3\xf0\x9f\x98\x81\xf0\x9f\xa5\xb9\xf0\x9f\x98\x86guys i love haknyeon so muchsanghaakkkkkkkkkkkk armsssssssssss\xf0\x9f\x98\x84\xf0\x9f\x98\x84\xf0\x9f\x98\x84\xf0\x9f\x98\x84\xf0\x9f\x98\x84\xf0\x9f\x98\x84\xf0\x9f\x98\x84 ilywetheboyz omgwetheboyz juyeonliterally what the fucklord shut up hak ily\xf0\x9f\x98\x84\xf0\x9f\x98\x84 i love youwetheboyz omgoh my bbangnyuman\xe2\x80\xa6 am i suppose to move on what baby i love you so muchcravitystarship thank you youngtae \xf0\x9f\xab\xb6\xf0\x9f\xab\xb6\xf0\x9f\xab\xb6hey god its me again my loveswetheboyz thank youwetheboyz oh my godbbangnyuanother win for me today as a fengfan and juyeon enthusiast way my god my god \xf0\x9f\xab\xb6\xf0\x9f\xab\xb6\xf0\x9f\xab\xb6\xf0\x9f\xab\xb6\xf0\x9f\xab\xb6\xf0\x9f\xab\xb6 i love him so much yay so proud of everyone today \xf0\x9f\xab\xb6\xf0\x9f\xab\xb6couldn\xe2\x80\x99t watch dream concert earlier when it was streaming but i just watch cix performance and they served once again \xf0\x9f\xab\xb6\xf0\x9f\xab\xb6seungyounification of juyeon \xf0\x9f\xab\xb6\xf0\x9f\xab\xb6 i dont miss when cix is on the red carpet im not home and my data is \xf0\x9f\xa5\xb9i really hope hes doing well boyz served once againsweet \xf0\x9f\xab\xb6\xf0\x9f\xab\xb6\xf0\x9f\xab\xb6\xf0\x9f\xab\xb6\xf0\x9f\xab\xb6\xf0\x9f\xab\xb6bloom bloom 20 \xf0\x9f\xab\xb6\xf0\x9f\xab\xb6\xf0\x9f\xab\xb6\xf0\x9f\xab\xb6ist i need you to announce literally right now if the boyz is coming to sg so i can find money and plan my timelomlllllll lomlhappy seeun day lt3 cannot do this today\xe2\x80\xa6 have hope for the boyz coming to sg plsplsplsplsplsplsnct 127 in sg\xe2\x80\xa6 so tempted to buy the tickets \xf0\x9f\x98\xad\xf0\x9f\x98\xadlove well soon jacob good morning to me omg for juyeon kevin and changmin fast recovery no way omgwhat\xe2\x80\xa6because like wtf morning to me what the fuck did i just wake up towetheboyz qtqueen congrats loves so proud of everyoneso proud of them they deserve it so much lt3 omg hi \xf0\x9f\xab\xb6\xf0\x9f\xab\xb6\xf0\x9f\xab\xb6my love omgwetheboyz goodnight kev lt3spent an hour watching bbangnyu vlive i think life is at its peakvityprnt happy birthday cor doh my i love you wetheboyz youmy bbangnyuwhat the fuckshut up morning to me yay sunwooweth eboyz ilyistheboyz thank you hyunjaeoh my god world stop omgistheboyz whati love him my hak yay hakwetheboyz my loveforever 11 lt3okay but no bloom bloom \xf0\x9f\xa5\xb9\xf0\x9f\xa5\xb9i would have fucking collapsed what proud of thembet on youddd my belovedtheyre the cutest fr actually gonna cry bloom bloom\xe2\x80\xa6 no bloom bloomthis \xf0\x9f\xa5\xb9\xf0\x9f\xa5\xb9\xf0\x9f\xa5\xb9\xf0\x9f\xa5\xb9\xf0\x9f\x98\x81also the birthday surprise for jacob thats so cutei keep on saying that i will log off for my mental health but the thought of missing out when they are in the red o\xe2\x80\xa6 have survived this irl red outfits are so prettywhat my chanhee cannot i cannot i cannot checkmate omgi need to see haklord juyeon\xf0\x9f\xa5\xb9 off \xf0\x9f\x98\xad\xf0\x9f\x98\xad\xf0\x9f\x98\xad\xf0\x9f\x98\xad\xf0\x9f\x98\xad\xf0\x9f\x98\xad\xf0\x9f\x98\xadmay we get a lot of haknyeon pics today or i will screamone day one day they will come to singapore jacob day \xe2\x9d\xa4\xef\xb8\x8f\n\n\xeb\x94\xb0\xec\x8a\xa4\xed\x95\xa8\xec\x9c\xbc\xeb\xa1\x9c\xeb\xac\xbc\xeb\x93\xa0\xec\xa0\x9c\xec\x9d\xb4\xec\xbd\xa5\xec\x9d\x98\xec\x8a\xa4\xeb\xac\xbc\xec\x97\xac\xec\x84\xaf\nourangeljacobdayhappy jacob day lt3 going to sleep through the whole theb zone concert tmr definitely not because i cannot go to oneokay younghoon has an undercut and i only know about it nowkev hyunjae younghoon \xf0\x9f\xa5\xb9\xf0\x9f\xa5\xb9\xf0\x9f\xa5\xb9not now please why is always when im outside\xf0\x9f\x98\xad\xf0\x9f\x98\xad\xf0\x9f\x98\xad\xf0\x9f\x98\xad my god prettiest hak hak my favourite hakcherry hak wdym wdym

```
wdymmy loves  the prettiest  mf late to the pink chanhee but omg pink chanh
eewetheboyz i love you so so much \xf0\x9f\xab\xb6\xf0\x9f\xab\xb6\xf0\x9f
\xab\xb6jichang  ily'
```

- https://stackoverflow.com/questions/57514169/how-can-i-remove-emojis-from-a-dataframe (https://stackoverflow.com/questions/57514169/how-can-i-remove-emojis-from-a-dataframe)

In [61]:
```
type(text)
```

Out[61]:
```
pandas.core.series.Series
```

In [62]:
```
test=[[text]]
```

In [63]:
```
text
```

Out[63]:
```
0       pericles216 hierbeforetheac sachinettiyil the ...
1       hispanthicckk being you makes you look cutethi...
2       alshymi les balles sont réelles et sont tirées...
3       im like entp but idiotichey boy do you want to...
4       kaeshurr1 give it to zargarshanif  he has pica...
                              ...
7806    sobsjjun godpls take care 😕sobsjjun hiro emerg...
7807    ignis02 wow last time i got intp    i think u ...
7808    akupilled a 100akupilled that someone will get...
7809    if you’re intj this one is for you  what is ne...
7810    harrylambert gucci hey can you dm me a pic of ...
Name: text, Length: 7811, dtype: object
```

In [64]:
```
text=pd.DataFrame(text)
text
```

Out[64]:

	text
0	pericles216 hierbeforetheac sachinettiyil the ...
1	hispanthicckk being you makes you look cutethi...
2	alshymi les balles sont réelles et sont tirées...
3	im like entp but idiotichey boy do you want to...
4	kaeshurr1 give it to zargarshanif he has pica...
...	...
7806	sobsjjun godpls take care 😟sobsjjun hiro emerg...
7807	ignis02 wow last time i got intp i think u ...
7808	akupilled a 100akupilled that someone will get...
7809	if you're intj this one is for you what is ne...
7810	harrylambert gucci hey can you dm me a pic of ...

7811 rows × 1 columns

In [65]:
```
type(text)
```

Out[65]:
```
pandas.core.frame.DataFrame
```

In [66]:
```
import emoji
text=text.astype(str).apply(lambda x: x.str.encode('ascii', 'ignore').str.dec
```

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

Out[67]:
```
(7811, 1)
```

In [68]:
```
text[7606:7607]
```

Out[68]:

	text
7606	guys omg so happy they got to meet again say...

In [69]:
```
text=text['text']
```

In [70]:
```
text.shape
```

Out[70]:
```
(7811,)
```

In [71]:
```
type(text)
```

Out[71]:
```
pandas.core.series.Series
```

In [72]:
```
text
```

Out[72]:
```
0       pericles216 hierbeforetheac sachinettiyil the ...
1       hispanthicckk being you makes you look cutethi...
2       alshymi les balles sont relles et sont tires t...
3       im like entp but idiotichey boy do you want to...
4       kaeshurr1 give it to zargarshanif  he has pica...
                              ...
7806    sobsjjun godpls take care sobsjjun hiro emerge...
7807    ignis02 wow last time i got intp    i think u ...
7808    akupilled a 100akupilled that someone will get...
7809    if youre intj this one is for you  what is nev...
7810    harrylambert gucci hey can you dm me a pic of ...
Name: text, Length: 7811, dtype: object
```

In [73]:
```
df['text']
```

Out[73]:
```
0       @Pericles216 @HierBeforeTheAC @Sachinettiyil T...
1       @Hispanthicckk Being you makes you look cute||...
2       @Alshymi Les balles sont réelles et sont tirée...
3       I'm like entp but idiotic|||Hey boy, do you wa...
4       @kaeshurr1 Give it to @ZargarShanif ... He has...
                              ...
7806    @sobsjjun God,,pls take care 😕|||@sobsjjun Hir...
7807    @Ignis_02 wow last time i got intp https://t.c... (https://t.c...)
7808    @akupilled A 100%|||@akupilled That SOMEONE wi...
7809    If you’re #INTJ this one is for you | What is ...
7810    @harry__lambert @gucci hey can you dm me a pic...
Name: text, Length: 7811, dtype: object
```

In [ ]:
```

```

Tokenization

In [74]:
```
from nltk.tokenize import word_tokenize,sent_tokenize
```

In [ ]:
```

```

In [75]:
```
text=pd.DataFrame(text)
text
```

Out[75]:

	text
0	pericles216 hierbeforetheac sachinettiyil the ...
1	hispanthicckk being you makes you look cutethi...
2	alshymi les balles sont relles et sont tires t...
3	im like entp but idiotichey boy do you want to...
4	kaeshurr1 give it to zargarshanif he has pica...
...	...
7806	sobsjjun godpls take care sobsjjun hiro emerge...
7807	ignis02 wow last time i got intp i think u ...
7808	akupilled a 100akupilled that someone will get...
7809	if youre intj this one is for you what is nev...
7810	harrylambert gucci hey can you dm me a pic of ...

7811 rows × 1 columns

In [76]:
```
#text['text']=text['text'].apply(sent_tokenize)
```

In [77]:
```
text['text']
```

Out[77]:
```
0       pericles216 hierbeforetheac sachinettiyil the ...
1       hispanthicckk being you makes you look cutethi...
2       alshymi les balles sont relles et sont tires t...
3       im like entp but idiotichey boy do you want to...
4       kaeshurr1 give it to zargarshanif  he has pica...
                              ...
7806    sobsjjun godpls take care sobsjjun hiro emerge...
7807    ignis02 wow last time i got intp    i think u ...
7808    akupilled a 100akupilled that someone will get...
7809    if youre intj this one is for you  what is nev...
7810    harrylambert gucci hey can you dm me a pic of ...
Name: text, Length: 7811, dtype: object
```

In [78]:
```
sent_tokenize
```

Out[78]:
```
<function nltk.tokenize.sent_tokenize(text, language='english')>
```

In [79]:
```
df = pd.DataFrame({'sentences': ['This is a very good site. I will recommend
df
```

Out[79]:

	sentences
0	This is a very good site. I will recommend it ...
1	Can you please give me a call at 9983938428. h...
2	good work! keep it up

https://stackoverflow.com/questions/33098040/how-to-use-word-tokenize-in-data-frame (https://stackoverflow.com/questions/33098040/how-to-use-word-tokenize-in-data-frame)

In [80]:
```
df['tokenized_sents'] = df.apply(lambda row: nltk.word_tokenize(row['sentence
df
```

Out[80]:

	sentences	tokenized_sents
0	This is a very good site. I will recommend it ...	[This, is, a, very, good, site, ., I, will, re...
1	Can you please give me a call at 9983938428. h...	[Can, you, please, give, me, a, call, at, 9983...
2	good work! keep it up	[good, work, !, keep, it, up]

In [81]:
```
df['sents_length'] = df.apply(lambda row: len(row['tokenized_sents']), axis=1
df
```

Out[81]:

	sentences	tokenized_sents	sents_length
0	This is a very good site. I will recommend it ...	[This, is, a, very, good, site, ., I, will, re...	14
1	Can you please give me a call at 9983938428. h...	[Can, you, please, give, me, a, call, at, 9983...	15
2	good work! keep it up	[good, work, !, keep, it, up]	6

In [82]:
```
text['text']
```

Out[82]:
```
0       pericles216 hierbeforetheac sachinettiyil the ...
1       hispanthicckk being you makes you look cutethi...
2       alshymi les balles sont relles et sont tires t...
3       im like entp but idiotichey boy do you want to...
4       kaeshurr1 give it to zargarshanif  he has pica...
                              ...
7806    sobsjjun godpls take care sobsjjun hiro emerge...
7807    ignis02 wow last time i got intp    i think u ...
7808    akupilled a 100akupilled that someone will get...
7809    if youre intj this one is for you  what is nev...
7810    harrylambert gucci hey can you dm me a pic of ...
Name: text, Length: 7811, dtype: object
```

In [83]:
```
word_tokenize=text['text'].apply(word_tokenize)
```

In [84]:
```
word_tokenize
```

Out[84]:
```
0       [pericles216, hierbeforetheac, sachinettiyil, ...
1       [hispanthicckk, being, you, makes, you, look, ...
2       [alshymi, les, balles, sont, relles, et, sont,...
3       [im, like, entp, but, idiotichey, boy, do, you...
4       [kaeshurr1, give, it, to, zargarshanif, he, ha...
                              ...
7806    [sobsjjun, godpls, take, care, sobsjjun, hiro,...
7807    [ignis02, wow, last, time, i, got, intp, i, th...
7808    [akupilled, a, 100akupilled, that, someone, wi...
7809    [if, youre, intj, this, one, is, for, you, wha...
7810    [harrylambert, gucci, hey, can, you, dm, me, a...
Name: text, Length: 7811, dtype: object
```

In [85]:
```
text1=pd.DataFrame(word_tokenize)
text['text']
```

Out[85]:
```
0       pericles216 hierbeforetheac sachinettiyil the ...
1       hispanthicckk being you makes you look cutethi...
2       alshymi les balles sont relles et sont tires t...
3       im like entp but idiotichey boy do you want to...
4       kaeshurr1 give it to zargarshanif  he has pica...
                              ...
7806    sobsjjun godpls take care sobsjjun hiro emerge...
7807    ignis02 wow last time i got intp    i think u ...
7808    akupilled a 100akupilled that someone will get...
7809    if youre intj this one is for you  what is nev...
7810    harrylambert gucci hey can you dm me a pic of ...
Name: text, Length: 7811, dtype: object
```

In [ ]:

stop words

In [86]:
```
from nltk.corpus import stopwords
```

In [87]:
```
stopwords.words('English')
```

```
 "you're",
 "you've",
 "you'll",
 "you'd",
 'your',
 'yours',
 'yourself',
 'yourselves',
 'he',
 'him',
 'his',
 'himself',
 'she',
 "she's",
 'her',
 'hers',
 'herself',
 'it',
 "it's",
```

In [88]:
```
def remove_stopwords(text):
    for char in stopwords.words('English'):
        text = text.replace(char,'')
    return text
#df['text'].apply(remove_stopwords)
text.apply(remove_stopwords)
```

Out[88]:

	text
0	pericles216 hierbeforetheac sachinettiyil the ...
1	hispanthicckk being you makes you look cutethi...
2	alshymi les balles sont relles et sont tires t...
3	im like entp but idiotichey boy do you want to...
4	kaeshurr1 give it to zargarshanif he has pica...
...	...
7806	sobsjjun godpls take care sobsjjun hiro emerge...
7807	ignis02 wow last time i got intp i think u ...
7808	akupilled a 100akupilled that someone will get...
7809	if youre intj this one is for you what is nev...
7810	harrylambert gucci hey can you dm me a pic of ...

7811 rows × 1 columns

In [89]:
```
text
```

Out[89]:

	text
0	pericles216 hierbeforetheac sachinettiyil the ...
1	hispanthicckk being you makes you look cutethi...
2	alshymi les balles sont relles et sont tires t...
3	im like entp but idiotichey boy do you want to...
4	kaeshurr1 give it to zargarshanif he has pica...
...	...
7806	sobsjjun godpls take care sobsjjun hiro emerge...
7807	ignis02 wow last time i got intp i think u ...
7808	akupilled a 100akupilled that someone will get...
7809	if youre intj this one is for you what is nev...
7810	harrylambert gucci hey can you dm me a pic of ...

7811 rows × 1 columns

Feature Engineering

- https://www.analyticsvidhya.com/blog/2021/07/feature-extraction-and-embeddings-in-nlp-a-beginners-guide-to-understand-natural-language-processing/ (https://www.analyticsvidhya.com/blog/2021/07/feature-extraction-and-embeddings-in-nlp-a-beginners-guide-to-understand-natural-language-processing/)
- https://www.geeksforgeeks.org/feature-extraction-techniques-nlp/ (https://www.geeksforgeeks.org/feature-extraction-techniques-nlp/)

1. creating bag of words using countvectorizer

- https://www.kaggle.com/code/basilb2s/language-detection-using-nlp (https://www.kaggle.com/code/basilb2s/language-detection-using-nlp)

- CountVectorizer

- from sklearn.feature_extraction.text import CountVectorizer
- cv = CountVectorizer()
- X = cv.fit_transform(text_list).toarray()
- X.shape
- https://heartbeat.comet.ml/using-machine-learning-for-language-detection-517fa6e68f22 (https://heartbeat.comet.ml/using-machine-learning-for-language-detection-517fa6e68f22)

In [90]:
```
text[7809:7810]
```

Out[90]:

	text
7809	if youre intj this one is for you what is nev...

In [91]:
```
#CountVectorizer
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer()
x1 = cv.fit_transform(text[7809:7810])#.toarray()
x1.shape
```

Out[91]:
```
(1, 1)
```

https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html (https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html)

In [92]:
```
x1
```

Out[92]:
```
<1x1 sparse matrix of type '<class 'numpy.int64'>'
        with 1 stored elements in Compressed Sparse Row format>
```

https://stackoverflow.com/questions/48603888/print-all-columns-and-rows-of-a-numpy-array (https://stackoverflow.com/questions/48603888/print-all-columns-and-rows-of-a-numpy-array)

In [93]:
```
#np.set_printoptions(threshold=np.inf)
```

In [94]:
```
x1
```

Out[94]:
```
<1x1 sparse matrix of type '<class 'numpy.int64'>'
        with 1 stored elements in Compressed Sparse Row format>
```

In [95]:
```
text[['text']]
```

Out[95]:

	text
0	pericles216 hierbeforetheac sachinettiyil the ...
1	hispanthicckk being you makes you look cutethi...
2	alshymi les balles sont relles et sont tires t...
3	im like entp but idiotichey boy do you want to...
4	kaeshurr1 give it to zargarshanif he has pica...
...	...
7806	sobsjjun godpls take care sobsjjun hiro emerge...
7807	ignis02 wow last time i got intp i think u ...
7808	akupilled a 100akupilled that someone will get...
7809	if youre intj this one is for you what is nev...
7810	harrylambert gucci hey can you dm me a pic of ...

7811 rows × 1 columns

In [96]:
```
from sklearn.feature_extraction.text import CountVectorizer
cv = CountVectorizer()
x1 = cv.fit_transform(text['text'])
x1.shape
```

Out[96]:
```
(7811, 736582)
```

In [97]:
```
x1
```

Out[97]:
```
<7811x736582 sparse matrix of type '<class 'numpy.int64'>'
        with 4425658 stored elements in Compressed Sparse Row format>
```

2. TF-IDF (Term Frequency-Inverse Document Frequency)

- https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/ (https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/)
- https://www.learndatasci.com/glossary/tf-idf-term-frequency-inverse-document-frequency/ (https://www.learndatasci.com/glossary/tf-idf-term-frequency-inverse-document-frequency/)
- https://towardsdatascience.com/tf-term-frequency-idf-inverse-document-frequency-from-scratch-in-python-6c2b61b78558 (https://towardsdatascience.com/tf-term-frequency-idf-inverse-document-frequency-from-scratch-in-python-6c2b61b78558)
- https://blog.marketmuse.com/glossary/term-frequency-inverse-document-frequency-tf-idf-definition/ (https://blog.marketmuse.com/glossary/term-frequency-inverse-document-frequency-tf-idf-definition/)
- https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/ (https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/)
- https://monkeylearn.com/blog/what-is-tf-idf/ (https://monkeylearn.com/blog/what-is-tf-idf/)

In [98]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [99]:
```
tr_idf_model  = TfidfVectorizer()
tf_idf = tr_idf_model.fit_transform(text['text'])
```

In [100]:
```
tf_idf.size
```

Out[100]:
```
4425658
```

https://www.flightpedia.org/convert/44-gigabytes-to-bytes.html (https://www.flightpedia.org/convert/44-gigabytes-to-bytes.html)

In [101]:
```
print(type(tf_idf), tf_idf.shape)
```

```
<class 'scipy.sparse.csr.csr_matrix'> (7811, 736582)
```

```
tf_idf_array = tf_idf_vector.toarray()

print(tf_idf_array)
```

In [102]:
```
from keras.utils import np_utils
npy=np_utils.to_categorical(y)
print(npy[:5])
```

```
[[0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0.]
 [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0.]
 [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0.]
 [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0.]
 [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0.]]
```

In [103]:
```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x1,npy,test_size=.25, random_s
```

In [104]:
```
print(x_train.shape)
print(x_test.shape)
print(y_train.shape)
print(y_test.shape)
```

```
(5858, 736582)
(1953, 736582)
(5858, 16)
(1953, 16)
```

In [105]:
```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler(with_mean=False)
x_train = sc.fit_transform(x_train)
x_test = sc.transform(x_test)
```

In [106]:
```
print(x_train.shape)
print(x_test.shape)
print(y_train.shape)
print(y_test.shape)
```

```
(5858, 736582)
(1953, 736582)
(5858, 16)
(1953, 16)
```

In [107]:
```
import tensorflow as tf
from tensorflow import keras
from keras.models import Sequential
from keras.layers import Dense,Activation,Dropout
from keras.optimizers import Adam
from keras.metrics import categorical_crossentropy
```

In [108]:
```
model = Sequential()
model.add(Dense(units=160,activation='relu', input_dim=736582))
model.add(Dropout(0.1))
model.add(Dense(units=16,activation='relu'))
model.add(Dropout(0.1))
model.add(Dense(units=16,activation='sigmoid'))
```

In [109]:
```
model.compile(optimizer= 'Adam', loss='categorical_crossentropy',metrics=['ac
```

In [110]:
```
model.summary()
```

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 160)               117853280

 dropout (Dropout)           (None, 160)               0

 dense_1 (Dense)             (None, 16)                2576

 dropout_1 (Dropout)         (None, 16)                0

 dense_2 (Dense)             (None, 16)                272

=================================================================
Total params: 117,856,128
Trainable params: 117,856,128
Non-trainable params: 0
_________________________________________________________________
```

- https://machinelearningmastery.com/tutorial-first-neural-network-python-keras/ (https://machinelearningmastery.com/tutorial-first-neural-network-python-keras/)

In [111]:
```
model.fit(x_train,y_train, epochs=1, verbose=1)
```

```
C:\ProgramData\Anaconda3\lib\site-packages\tensorflow\python\framework\inde
xed_slices.py:444: UserWarning: Converting sparse IndexedSlices(IndexedSlic
es(indices=Tensor("gradient_tape/sequential/dense/embedding_lookup_sparse/R
eshape_1:0", shape=(None,), dtype=int32), values=Tensor("gradient_tape/sequ
ential/dense/embedding_lookup_sparse/Reshape:0", shape=(None, 160), dtype=f
loat32), dense_shape=Tensor("gradient_tape/sequential/dense/embedding_looku
p_sparse/Cast:0", shape=(2,), dtype=int32))) to a dense Tensor of unknown s
hape. This may consume a large amount of memory.
  warnings.warn(

184/184 [==============================] - 135s 732ms/step - loss: 2.8176 -
accuracy: 0.1408
```

Out[111]:
```
<keras.callbacks.History at 0x20d91a5b5e0>
```

In [ ]:
```

```

In [112]:
```
# import liberary
import numpy as np
import pandas as pd
# read data set
df=pd.read_csv('twitter_MBTI.csv')
df.head()

y=df['label']
```

In [113]:
```
from sklearn.feature_extraction.text import CountVectorizer
from nltk.tokenize import RegexpTokenizer
token = RegexpTokenizer(r'[a-zA-Z0-9]+')
cv = CountVectorizer(lowercase=True,stop_words='english',ngram_range=(1,1), t
X = cv.fit_transform(text['text'])
X
```

Out[113]:
```
<7811x736308 sparse matrix of type '<class 'numpy.int64'>'
        with 3666239 stored elements in Compressed Sparse Row format>
```

In [114]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
tf=TfidfVectorizer()
X=tf.fit_transform(text['text'])
X
```

Out[114]:
```
<7811x736582 sparse matrix of type '<class 'numpy.float64'>'
        with 4425658 stored elements in Compressed Sparse Row format>
```

In [115]:
```
from sklearn.preprocessing import LabelEncoder
lb=LabelEncoder()
y=lb.fit_transform(df['label'])
y
```

Out[115]:
```
array([10, 10, 10, ...,  3,  8, 15])
```

In [116]:
```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(X,y,test_size=.25)
```

In [117]:
```
print(x_train.shape)
print(x_test.shape)
print(y_train.shape)
print(y_test.shape)
```

```
(5858, 736582)
(1953, 736582)
(5858,)
(1953,)
```

In [118]:
```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler(with_mean=False)
x_train = sc.fit_transform(x_train)
x_test = sc.transform(x_test)
```

In [119]:
```
from sklearn.naive_bayes import MultinomialNB
from sklearn import metrics
nb= MultinomialNB().fit(x_train,y_train)
nb
```

Out[119]:
```
MultinomialNB()
```

In [120]:
```
nb.score(x_train,y_train)
```

Out[120]:
```
1.0
```

In [121]:
```
nb.score(x_test,y_test)
```

Out[121]:
```
0.18996415770609318
```

In [122]:
```
y_pred_nb= nb.predict(x_test)
print(len(y_pred_nb))
y_pred_nb
```

```
1953
```

Out[122]:
```
array([ 4, 10,  9, ...,  9, 13,  1])
```

In [123]:
```
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.metrics import accuracy_score
```

In [124]:
```
metrics.accuracy_score(y_test,y_pred_nb)
```

Out[124]:
```
0.18996415770609318
```

In [125]:
```
print(confusion_matrix (y_test,y_pred_nb))
```

```
[[20 12  2  7  0  4  6  5 21 16 18  8  7  3  2  1]
 [10 25  4  7  8  4  3  3 36 29 25 14  5  7  8  3]
 [ 3  5  5  3  1  1  3  0  9  9 21  3  3  1  0  3]
 [ 6 16  1 34  3  4  2  6 19 20 24 12  2  3  6  5]
 [ 0  1  0  3  3  0  0  0  3  2  1  2  2  3  0  0]
 [ 6  5  1  3  0  5  1  0  5 12  3  4  2  2  0  1]
 [ 4  2  1  2  0  0  1  0  4  2  6  1  0  1  0  1]
 [ 1  2  0  6  1  0  2  1  3  3  1  2  1  0  0  1]
 [ 9 17  8  9  6  4  6  1 70 34 47 17 19  6  5  5]
 [15 23 12 14  7 10  6  9 50 75 47 21 10 14  3  6]
 [ 7 10  6  6  6  2  6  4 36 22 50 10  7  5  6  4]
 [14  9  1 14  3  5  6  2 32 26 26 37 11  6  0  6]
 [ 6  5  2  2  1  2  2  1  9 16  3  2 17  1  2  1]
 [ 3  7  5  1  3  3  2  2 17 16  7 10  3 12  6  5]
 [ 1  0  1  3  3  3  4  1  7  8  6  4  4  4  5  1]
 [ 3  6  2  6  3  1  4  0  8 17  5  3  5  4  1 11]]
```

- https://github.com/omkantsharma/Suicide_Detection_Project (https://github.com/omkantsharma/Suicide_Detection_Project)
- https://github.com/omkantsharma/Spam-Ham-Classification (https://github.com/omkantsharma/Spam-Ham-Classification)
- https://github.com/omkantsharma/Sentiment-Analysis-NLP-Classification-Project (https://github.com/omkantsharma/Sentiment-Analysis-NLP-Classification-Project)