Twitter Hate Speech Detection-Project Notebook

Notebook Contributors: mumar, msakir

In [ ]:

```
!pip install tweet-preprocessor
```

```
Collecting tweet-preprocessor
  Downloading https://files.pythonhosted.org/packages/17/9d/71bd016a9edcef8860c607e531f30
bd09b13103c7951ae73dd2bf174163c/tweet_preprocessor-0.6.0-py3-none-any.whl
Installing collected packages: tweet-preprocessor
Successfully installed tweet-preprocessor-0.6.0
```

In [ ]:

```
!pip install demoji
```

```
Collecting demoji
  Downloading https://files.pythonhosted.org/packages/7b/fd/265f1ad2d745d6f46d1ede83d0054
327e87154e9f14b252c1e272749e657/demoji-0.3.0-py2.py3-none-any.whl
Requirement already satisfied: requests<3.0.0 in /usr/local/lib/python3.6/dist-packages (
from demoji) (2.23.0)
Collecting colorama
  Downloading https://files.pythonhosted.org/packages/44/98/5b86278fbbf250d239ae0ecb724f8
572af1c91f4a11edf4d36a206189440/colorama-0.4.4-py2.py3-none-any.whl
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /usr/local/lib/
python3.6/dist-packages (from requests<3.0.0->demoji) (1.24.3)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.6/dist-packages (fr
om requests<3.0.0->demoji) (2.10)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.6/dist-packag
es (from requests<3.0.0->demoji) (2020.11.8)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.6/dist-package
s (from requests<3.0.0->demoji) (3.0.4)
Installing collected packages: colorama, demoji
Successfully installed colorama-0.4.4 demoji-0.3.0
```

In [ ]:

```
import pandas as pd
import nltk
from nltk.corpus import stopwords
import preprocessor as p
import string
from textblob import TextBlob
import demoji
import re
from gensim.models import Word2Vec
import gensim.downloader
```

In [ ]:

```
nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[ ]:

```
True
```

In [ ]:

```
nltk.download()
```

```
NLTK Downloader
---------------------------------------------------------------------------
    d) Download   l) List    u) Update   c) Config   h) Help   q) Quit
```

```
---------------------------------------------------------------------------
Downloader> d

Download which package (l=list; x=cancel)?
  Identifier> all
    Downloading collection 'all'
       |
       | Downloading package abc to /root/nltk_data...
       |   Unzipping corpora/abc.zip.
       | Downloading package alpino to /root/nltk_data...
       |   Unzipping corpora/alpino.zip.
       | Downloading package biocreative_ppi to /root/nltk_data...
       |   Unzipping corpora/biocreative_ppi.zip.
       | Downloading package brown to /root/nltk_data...
       |   Unzipping corpora/brown.zip.
       | Downloading package brown_tei to /root/nltk_data...
       |   Unzipping corpora/brown_tei.zip.
       | Downloading package cess_cat to /root/nltk_data...
       |   Unzipping corpora/cess_cat.zip.
       | Downloading package cess_esp to /root/nltk_data...
       |   Unzipping corpora/cess_esp.zip.
       | Downloading package chat80 to /root/nltk_data...
       |   Unzipping corpora/chat80.zip.
       | Downloading package city_database to /root/nltk_data...
       |   Unzipping corpora/city_database.zip.
       | Downloading package cmudict to /root/nltk_data...
       |   Unzipping corpora/cmudict.zip.
       | Downloading package comparative_sentences to
       |     /root/nltk_data...
       |   Unzipping corpora/comparative_sentences.zip.
       | Downloading package comtrans to /root/nltk_data...
       | Downloading package conll2000 to /root/nltk_data...
       |   Unzipping corpora/conll2000.zip.
       | Downloading package conll2002 to /root/nltk_data...
       |   Unzipping corpora/conll2002.zip.
       | Downloading package conll2007 to /root/nltk_data...
       | Downloading package crubadan to /root/nltk_data...
       |   Unzipping corpora/crubadan.zip.
       | Downloading package dependency_treebank to /root/nltk_data...
       |   Unzipping corpora/dependency_treebank.zip.
       | Downloading package dolch to /root/nltk_data...
       |   Unzipping corpora/dolch.zip.
       | Downloading package europarl_raw to /root/nltk_data...
       |   Unzipping corpora/europarl_raw.zip.
       | Downloading package floresta to /root/nltk_data...
       |   Unzipping corpora/floresta.zip.
       | Downloading package framenet_v15 to /root/nltk_data...
       |   Unzipping corpora/framenet_v15.zip.
       | Downloading package framenet_v17 to /root/nltk_data...
       |   Unzipping corpora/framenet_v17.zip.
       | Downloading package gazetteers to /root/nltk_data...
       |   Unzipping corpora/gazetteers.zip.
       | Downloading package genesis to /root/nltk_data...
       |   Unzipping corpora/genesis.zip.
       | Downloading package gutenberg to /root/nltk_data...
       |   Unzipping corpora/gutenberg.zip.
       | Downloading package ieer to /root/nltk_data...
       |   Unzipping corpora/ieer.zip.
       | Downloading package inaugural to /root/nltk_data...
       |   Unzipping corpora/inaugural.zip.
       | Downloading package indian to /root/nltk_data...
       |   Unzipping corpora/indian.zip.
       | Downloading package jeita to /root/nltk_data...
       | Downloading package kimmo to /root/nltk_data...
       |   Unzipping corpora/kimmo.zip.
       | Downloading package knbc to /root/nltk_data...
       | Downloading package lin_thesaurus to /root/nltk_data...
       |   Unzipping corpora/lin_thesaurus.zip.
       | Downloading package mac_morpho to /root/nltk_data...
       |   Unzipping corpora/mac_morpho.zip.
       | Downloading package machado to /root/nltk_data...
       | Downloading package masc_tagged to /root/nltk_data...
```

```
| Downloading package moses_sample to /root/nltk_data...
|   Unzipping models/moses_sample.zip.
| Downloading package movie_reviews to /root/nltk_data...
|   Unzipping corpora/movie_reviews.zip.
| Downloading package names to /root/nltk_data...
|   Unzipping corpora/names.zip.
| Downloading package nombank.1.0 to /root/nltk_data...
| Downloading package nps_chat to /root/nltk_data...
|   Unzipping corpora/nps_chat.zip.
| Downloading package omw to /root/nltk_data...
|   Unzipping corpora/omw.zip.
| Downloading package opinion_lexicon to /root/nltk_data...
|   Unzipping corpora/opinion_lexicon.zip.
| Downloading package paradigms to /root/nltk_data...
|   Unzipping corpora/paradigms.zip.
| Downloading package pil to /root/nltk_data...
|   Unzipping corpora/pil.zip.
| Downloading package pl196x to /root/nltk_data...
|   Unzipping corpora/pl196x.zip.
| Downloading package ppattach to /root/nltk_data...
|   Unzipping corpora/ppattach.zip.
| Downloading package problem_reports to /root/nltk_data...
|   Unzipping corpora/problem_reports.zip.
| Downloading package propbank to /root/nltk_data...
| Downloading package ptb to /root/nltk_data...
|   Unzipping corpora/ptb.zip.
| Downloading package product_reviews_1 to /root/nltk_data...
|   Unzipping corpora/product_reviews_1.zip.
| Downloading package product_reviews_2 to /root/nltk_data...
|   Unzipping corpora/product_reviews_2.zip.
| Downloading package pros_cons to /root/nltk_data...
|   Unzipping corpora/pros_cons.zip.
| Downloading package qc to /root/nltk_data...
|   Unzipping corpora/qc.zip.
| Downloading package reuters to /root/nltk_data...
| Downloading package rte to /root/nltk_data...
|   Unzipping corpora/rte.zip.
| Downloading package semcor to /root/nltk_data...
| Downloading package senseval to /root/nltk_data...
|   Unzipping corpora/senseval.zip.
| Downloading package sentiwordnet to /root/nltk_data...
|   Unzipping corpora/sentiwordnet.zip.
| Downloading package sentence_polarity to /root/nltk_data...
|   Unzipping corpora/sentence_polarity.zip.
| Downloading package shakespeare to /root/nltk_data...
|   Unzipping corpora/shakespeare.zip.
| Downloading package sinica_treebank to /root/nltk_data...
|   Unzipping corpora/sinica_treebank.zip.
| Downloading package smultron to /root/nltk_data...
|   Unzipping corpora/smultron.zip.
| Downloading package state_union to /root/nltk_data...
|   Unzipping corpora/state_union.zip.
| Downloading package stopwords to /root/nltk_data...
|   Unzipping corpora/stopwords.zip.
| Downloading package subjectivity to /root/nltk_data...
|   Unzipping corpora/subjectivity.zip.
| Downloading package swadesh to /root/nltk_data...
|   Unzipping corpora/swadesh.zip.
| Downloading package switchboard to /root/nltk_data...
|   Unzipping corpora/switchboard.zip.
| Downloading package timit to /root/nltk_data...
|   Unzipping corpora/timit.zip.
| Downloading package toolbox to /root/nltk_data...
|   Unzipping corpora/toolbox.zip.
| Downloading package treebank to /root/nltk_data...
|   Unzipping corpora/treebank.zip.
| Downloading package twitter_samples to /root/nltk_data...
|   Unzipping corpora/twitter_samples.zip.
| Downloading package udhr to /root/nltk_data...
|   Unzipping corpora/udhr.zip.
| Downloading package udhr2 to /root/nltk_data...
|   Unzipping corpora/udhr2.zip.
```

```
       | Downloading package unicode_samples to /root/nltk_data...
       |   Unzipping corpora/unicode_samples.zip.
       | Downloading package universal_treebanks_v20 to
       |     /root/nltk_data...
       | Downloading package verbnet to /root/nltk_data...
       |   Unzipping corpora/verbnet.zip.
       | Downloading package verbnet3 to /root/nltk_data...
       |   Unzipping corpora/verbnet3.zip.
       | Downloading package webtext to /root/nltk_data...
       |   Unzipping corpora/webtext.zip.
       | Downloading package wordnet to /root/nltk_data...
       |   Unzipping corpora/wordnet.zip.
       | Downloading package wordnet_ic to /root/nltk_data...
       |   Unzipping corpora/wordnet_ic.zip.
       | Downloading package words to /root/nltk_data...
       |   Unzipping corpora/words.zip.
       | Downloading package ycoe to /root/nltk_data...
       |   Unzipping corpora/ycoe.zip.
       | Downloading package rslp to /root/nltk_data...
       |   Unzipping stemmers/rslp.zip.
       | Downloading package maxent_treebank_pos_tagger to
       |     /root/nltk_data...
       |   Unzipping taggers/maxent_treebank_pos_tagger.zip.
       | Downloading package universal_tagset to /root/nltk_data...
       |   Unzipping taggers/universal_tagset.zip.
       | Downloading package maxent_ne_chunker to /root/nltk_data...
       |   Unzipping chunkers/maxent_ne_chunker.zip.
       | Downloading package punkt to /root/nltk_data...
       |   Unzipping tokenizers/punkt.zip.
       | Downloading package book_grammars to /root/nltk_data...
       |   Unzipping grammars/book_grammars.zip.
       | Downloading package sample_grammars to /root/nltk_data...
       |   Unzipping grammars/sample_grammars.zip.
       | Downloading package spanish_grammars to /root/nltk_data...
       |   Unzipping grammars/spanish_grammars.zip.
       | Downloading package basque_grammars to /root/nltk_data...
       |   Unzipping grammars/basque_grammars.zip.
       | Downloading package large_grammars to /root/nltk_data...
       |   Unzipping grammars/large_grammars.zip.
       | Downloading package tagsets to /root/nltk_data...
       |   Unzipping help/tagsets.zip.
       | Downloading package snowball_data to /root/nltk_data...
       | Downloading package bllip_wsj_no_aux to /root/nltk_data...
       |   Unzipping models/bllip_wsj_no_aux.zip.
       | Downloading package word2vec_sample to /root/nltk_data...
       |   Unzipping models/word2vec_sample.zip.
       | Downloading package panlex_swadesh to /root/nltk_data...
       | Downloading package mte_teip5 to /root/nltk_data...
       |   Unzipping corpora/mte_teip5.zip.
       | Downloading package averaged_perceptron_tagger to
       |     /root/nltk_data...
       |   Unzipping taggers/averaged_perceptron_tagger.zip.
       | Downloading package averaged_perceptron_tagger_ru to
       |     /root/nltk_data...
       |   Unzipping taggers/averaged_perceptron_tagger_ru.zip.
       | Downloading package perluniprops to /root/nltk_data...
       |   Unzipping misc/perluniprops.zip.
       | Downloading package nonbreaking_prefixes to
       |     /root/nltk_data...
       |   Unzipping corpora/nonbreaking_prefixes.zip.
       | Downloading package vader_lexicon to /root/nltk_data...
       | Downloading package porter_test to /root/nltk_data...
       |   Unzipping stemmers/porter_test.zip.
       | Downloading package wmt15_eval to /root/nltk_data...
       |   Unzipping models/wmt15_eval.zip.
       | Downloading package mwa_ppdb to /root/nltk_data...
       |   Unzipping misc/mwa_ppdb.zip.
       |
     Done downloading collection all

---------------------------------------------------------------------------
    d) Download   l) List    u) Update   c) Config   h) Help   q) Quit
```

```
--------------------------------------------------------------------------------
Downloader> q
```

Out[ ]:

```
True
```

In [ ]:

```
demoji.download_codes()
```

```
Downloading emoji data ...
... OK (Got response in 0.09 seconds)
Writing emoji data to /root/.demoji/codes.json ...
... OK
```

In [ ]:

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

Shallow Machine Learning

In [ ]:

```
df = pd.read_csv("/content/drive/MyDrive/Project/hate_speech_data/train_tweets.csv")
df
```

Out[ ]:

	id	label	tweet
0	1	0	@user when a father is dysfunctional and is s...
1	2	0	@user @user thanks for #lyft credit i can't us...
2	3	0	bihday your majesty
3	4	0	#model i love u take with u all the time in ...
4	5	0	factsguide: society now #motivation
...	...	...	...
31957	31958	0	ate @user isz that youuu?ðððððð...
31958	31959	0	to see nina turner on the airwaves trying to...
31959	31960	0	listening to sad songs on a monday morning otw...
31960	31961	1	@user #sikh #temple vandalised in in #calgary,...
31961	31962	0	thank you @user for you follow

31962 rows × 3 columns

Deleting the ID column

In [ ]:

```
del df['id']
df
```

Out[ ]:

	label	tweet
0	0	@user when a father is dysfunctional and is s...
1	0	@user @user thanks for #lyft credit i can't us...
2	0	bihday your majesty

	label	tweet
3	0	#model i love u take with u all the time in ...
4	0	factsguide: society now #motivation
...	...	...
31957	0	ate @user isz that youuu?ðððððð...
31958	0	to see nina turner on the airwaves trying to...
31959	0	listening to sad songs on a monday morning otw...
31960	1	@user #sikh #temple vandalised in in #calgary,...
31961	0	thank you @user for you follow

31962 rows × 2 columns

In [ ]:

```
df.describe()
```

Out[ ]:

	label
count	31962.000000
mean	0.070146
std	0.255397
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	1.000000

Converting Emoji to Text format

In [ ]:

```
def repEmoji(text):
  emoji = demoji.findall(text)
  for emotes, tex in emoji.items():
    text.replace(emotes, tex)
  return text
```

In [ ]:

```
df['cleaned'] = df['tweet'].apply(repEmoji)
df
```

Out[ ]:

	label	tweet	cleaned
0	0	@user when a father is dysfunctional and is s...	@user when a father is dysfunctional and is s...
1	0	@user @user thanks for #lyft credit i can't us...	@user @user thanks for #lyft credit i can't us...
2	0	bihday your majesty	bihday your majesty
3	0	#model i love u take with u all the time in ...	#model i love u take with u all the time in ...
4	0	factsguide: society now #motivation	factsguide: society now #motivation
...	...	...	...
31957	0	ate @user isz that youuu?ðððððð...	ate @user isz that youuu?ðððððð...
31958	0	to see nina turner on the airwaves trying to...	to see nina turner on the airwaves trying to...
31959	0	listening to sad songs on a monday morning otw...	listening to sad songs on a monday morning otw...

	label	tweet	cleaned
31960	1	@user #sikh #temple vandalised in in #calgary,...	@user #sikh #temple vandalised in in #calgary,...
31961	0	thank you @user for you follow	thank you @user for you follow

31962 rows × 3 columns

Cleaning tweets using Tweet-preprocess library

In [ ]:

```
p.set_options(p.OPT.URL, p.OPT.MENTION, p.OPT.RESERVED, p.OPT.EMOJI, p.OPT.SMILEY, p.OPT
.NUMBER)
df['cleaned'] = df['cleaned'].apply(p.clean)
df
```

Out[ ]:

	label	tweet	cleaned
0	0	@user when a father is dysfunctional and is s...	when a father is dysfunctional and is so selfi...
1	0	@user @user thanks for #lyft credit i can't us...	thanks for #lyft credit i can't use cause they...
2	0	bihday your majesty	bihday your majesty
3	0	#model i love u take with u all the time in ...	#model i love u take with u all the time in ur!!!
4	0	factsguide: society now #motivation	factsguide: society now #motivation
...	...	...	...
31957	0	ate @user isz that youuu?ðððððð...	ate isz that youuu?
31958	0	to see nina turner on the airwaves trying to...	to see nina turner on the airwaves trying to w...
31959	0	listening to sad songs on a monday morning otw...	listening to sad songs on a monday morning otw...
31960	1	@user #sikh #temple vandalised in in #calgary,...	#sikh #temple vandalised in in #calgary, #wso ...
31961	0	thank you @user for you follow	thank you for you follow

31962 rows × 3 columns

In [ ]:

```
contraction_mapping = {"ain't": "is not", "aren't": "are not","can't": "cannot", "'cause"
: "because", "could've": "could have", "couldn't": "could not",
                           "didn't": "did not",  "doesn't": "does not", "don't": "do not
", "hadn't": "had not", "hasn't": "has not", "haven't": "have not",
                           "he'd": "he would","he'll": "he will", "he's": "he is", "how'
d": "how did", "how'd'y": "how do you", "how'll": "how will", "how's": "how is",
                           "I'd": "I would", "I'd've": "I would have", "I'll": "I will",
"I'll've": "I will have","I'm": "I am", "I've": "I have", "i'd": "i would",
                           "i'll": "i will",  "i'll've": "i will have","i'm": "i am", "i
've": "i have", "isn't": "is not", "it'd": "it would",
                           "it'll": "it will", "it'll've": "it will have","it's": "it is
", "let's": "let us", "ma'am": "madam",
                           "might've": "might have","mightn't": "might not","mightn't've
": "might not have", "must've": "must have",
                           "mustn't": "must not", "mustn't've": "must not have", "needn'
t": "need not", "needn't've": "need not have","o'clock": "of the clock",
                           "oughtn't": "ought not", "oughtn't've": "ought not have", "sh
an't": "shall not", "sha'n't": "shall not", "shan't've": "shall not have",
                           "she'd": "she would", "she'd've": "she would have", "she'll":
"she will", "she'll've": "she will have", "she's": "she is",
                           "should've": "should have", "shouldn't": "should not", "shoul
dn't've": "should not have", "so've": "so have","so's": "so as",
                           "this's": "this is","that'd": "that would", "that'd've": "tha
t would have", "that's": "that is", "there'd": "there would",
                           "there'd've": "there would have", "there's": "there is", "her
e's": "here is","they'd": "they would", "they'd've": "they would have",
                           "they'll": "they will", "they'll've": "they will have", "they
're": "they are", "they've": "they have", "to've": "to have",
                           "wasn't": "was not", "we'd": "we would", "we'd've": "we would
```

```
have", "we'll": "we will", "we'll've": "we will have", "we're": "we are",
                            "we've": "we have", "weren't": "were not", "what'll": "what w
ill", "what'll've": "what will have", "what're": "what are",
                            "what's": "what is", "what've": "what have", "when's": "when
is", "when've": "when have", "where'd": "where did", "where's": "where is",
                            "where've": "where have", "who'll": "who will", "who'll've":
"who will have", "who's": "who is", "who've": "who have",
                            "why's": "why is", "why've": "why have", "will've": "will hav
e", "won't": "will not", "won't've": "will not have",
                            "would've": "would have", "wouldn't": "would not", "wouldn't'
ve": "would not have", "y'all": "you all",
                            "y'all'd": "you all would","y'all'd've": "you all would have"
,"y'all're": "you all are","y'all've": "you all have",
                            "you'd": "you would", "you'd've": "you would have", "you'll":
"you will", "you'll've": "you will have",
                            "you're": "you are", "you've": "you have", "ive": "i have"}
```

selfProcess(text) removes # symbol, converts words to their contracted form and removes single letters

In [ ]:

```
def selfPreprocess(text):
  tok = text.split()
  for i in range(0, len(tok)):
    if tok[i][0] == '#':
      new = tok[i][1:]
      tok[i] = new

  contr = [contraction_mapping[w] if w in contraction_mapping.keys() else w for w in tok
]

  sent = ' '.join(contr)
  reTok = sent.split()
  noSingle = [word for word in reTok if len(word)>1]

  return ' '.join(noSingle)
```

In [ ]:

```
df['cleaned'] = df['cleaned'].apply(selfPreprocess)
df
```

Out[ ]:

	label	tweet	cleaned
0	0	@user when a father is dysfunctional and is s...	when father is dysfunctional and is so selfish...
1	0	@user @user thanks for #lyft credit i can't us...	thanks for lyft credit cannot use cause they d...
2	0	bihday your majesty	bihday your majesty
3	0	#model i love u take with u all the time in ...	model love take with all the time in ur!!!
4	0	factsguide: society now #motivation	factsguide: society now motivation
...	...	...	...
31957	0	ate @user isz that youuu?ðððððð...	ate isz that youuu?
31958	0	to see nina turner on the airwaves trying to...	to see nina turner on the airwaves trying to w...
31959	0	listening to sad songs on a monday morning otw...	listening to sad songs on monday morning otw t...
31960	1	@user #sikh #temple vandalised in in #calgary,...	sikh temple vandalised in in calgary, wso cond...
31961	0	thank you @user for you follow	thank you for you follow

31962 rows × 3 columns

In [ ]:

```
def countWords(text):
    tok = text.split()
```

```
    return len(tok)
```

In [ ]:

```
df['word_count'] = df['cleaned'].apply(countWords)
df
```

Out[ ]:

	label	tweet	cleaned	word_count
0	0	@user when a father is dysfunctional and is s...	when father is dysfunctional and is so selfish...	16
1	0	@user @user thanks for #lyft credit i can't us...	thanks for lyft credit cannot use cause they d...	17
2	0	bihday your majesty	bihday your majesty	3
3	0	#model i love u take with u all the time in ...	model love take with all the time in ur!!!	9
4	0	factsguide: society now #motivation	factsguide: society now motivation	4
...	...	...	...	...
31957	0	ate @user isz that youuu?ðððððð...	ate isz that youuu?	4
31958	0	to see nina turner on the airwaves trying to...	to see nina turner on the airwaves trying to w...	22
31959	0	listening to sad songs on a monday morning otw...	listening to sad songs on monday morning otw t...	12
31960	1	@user #sikh #temple vandalised in in #calgary,...	sikh temple vandalised in in calgary, wso cond...	9
31961	0	thank you @user for you follow	thank you for you follow	5

31962 rows × 4 columns

In [ ]:

```
def uniqWordCount(text):
    tok = text.split()
    uniq = set(tok)
    return len(uniq)
```

In [ ]:

```
df['unique_words_count'] = df['cleaned'].apply(uniqWordCount)
df
```

Out[ ]:

	label	tweet	cleaned	word_count	unique_words_count
0	0	@user when a father is dysfunctional and is s...	when father is dysfunctional and is so selfish...	16	14
1	0	@user @user thanks for #lyft credit i can't us...	thanks for lyft credit cannot use cause they d...	17	17
2	0	bihday your majesty	bihday your majesty	3	3
3	0	#model i love u take with u all the time in ...	model love take with all the time in ur!!!	9	9
4	0	factsguide: society now #motivation	factsguide: society now motivation	4	4
...	...	...	...	...	...
31957	0	ate @user isz that youuu?ðððððð...	ate isz that youuu?	4	4
31958	0	to see nina turner on the airwaves trying to...	to see nina turner on the airwaves trying to w...	22	20
31959	0	listening to sad songs on a monday morning otw...	listening to sad songs on monday morning otw t...	12	10
31960	1	@user #sikh #temple vandalised in in #calgary,...	sikh temple vandalised in in calgary, wso cond...	9	8
31961	0	thank you @user for you follow	thank you for you follow	5	4

31962 rows × 5 columns

In [ ]:

```
def remStop(text):
  stop = stopwords.words('english')
  tok = text.split()
  noStop = [word for word in tok if word not in stop]
  return ' '.join(noStop)
```

In [ ]:

```
df['cleaned'] = df['cleaned'].apply(remStop)
df
```

Out[ ]:

	label	tweet	cleaned	word_count	unique_words_count
0	0	@user when a father is dysfunctional and is s...	father dysfunctional selfish drags kids dysfun...	16	14
1	0	@user @user thanks for #lyft credit i can't us...	thanks lyft credit cannot use cause offer whee...	17	17
2	0	bihday your majesty	bihday majesty	3	3
3	0	#model i love u take with u all the time in ...	model love take time ur!!!	9	9
4	0	factsguide: society now #motivation	factsguide: society motivation	4	4
...	...	...	...	...	...
31957	0	ate @user isz that youuu?ðððððð...	ate isz youuu?	4	4
31958	0	to see nina turner on the airwaves trying to...	see nina turner airwaves trying wrap mantle ge...	22	20
31959	0	listening to sad songs on a monday morning otw...	listening sad songs monday morning otw work sad	12	10
31960	1	@user #sikh #temple vandalised in in #calgary,...	sikh temple vandalised calgary, wso condemns act	9	8
31961	0	thank you @user for you follow	thank follow	5	4

31962 rows × 5 columns

In [ ]:

```
def punCount(text):
  tok = text.split()
  count = 0
  for word in tok:
    for char in word:
      if char in string.punctuation:
        count += 1
  return count
```

In [ ]:

```
df['puntuation_count'] = df['cleaned'].apply(punCount)
df
```

Out[ ]:

	label	tweet	cleaned	word_count	unique_words_count	puntuation_count
0	0	@user when a father is dysfunctional and is s...	father dysfunctional selfish drags kids dysfun...	16	14	1
1	0	@user @user thanks for #lyft credit i can't us...	thanks lyft credit cannot use cause offer whee...	17	17	1
2	0	bihday your majesty	bihday majesty	3	3	0

3	0	#model i love u take with u all the time in ...	model love take time ur!!!	9	9	3
4	0	factsguide: society now #motivation	factsguide: society motivation	4	4	1
...	...	...	...	...	...	...
31957	0	ate @user isz that youuu?ðððððð...	ate isz youuu?	4	4	1
31958	0	to see nina turner on the airwaves trying to...	see nina turner airwaves trying wrap mantle ge...	22	20	1
31959	0	listening to sad songs on a monday morning otw...	listening sad songs monday morning otw work sad	12	10	0
31960	1	@user #sikh #temple vandalised in in #calgary,...	sikh temple vandalised calgary, wso condemns act	9	8	1
31961	0	thank you @user for you follow	thank follow	5	4	0

31962 rows × 6 columns

In [ ]:

```
df['polarity'] = df['cleaned'].apply(lambda x: TextBlob(x).sentiment.polarity)
df
```

Out[ ]:

	label	tweet	cleaned	word_count	unique_words_count	puntuation_count	polarity
0	0	@user when a father is dysfunctional and is s...	father dysfunctional selfish drags kids dysfun...	16	14	1	-0.500000
1	0	@user @user thanks for #lyft credit i can't us...	thanks lyft credit cannot use cause offer whee...	17	17	1	0.200000
2	0	bihday your majesty	bihday majesty	3	3	0	0.000000
3	0	#model i love u take with u all the time in ...	model love take time ur!!!	9	9	3	0.976562
4	0	factsguide: society now #motivation	factsguide: society motivation	4	4	1	0.000000
...	...	...	...	...	...	...	...
31957	0	ate @user isz that youuu?ðððððð...	ate isz youuu?	4	4	1	0.000000
31958	0	to see nina turner on the airwaves trying to...	see nina turner airwaves trying wrap mantle ge...	22	20	1	0.400000
31959	0	listening to sad songs on a monday morning otw...	listening sad songs monday morning otw work sad	12	10	0	-0.500000
31960	1	@user #sikh #temple vandalised in in #calgary,...	sikh temple vandalised calgary, wso condemns act	9	8	1	0.000000
31961	0	thank you @user for you follow	thank follow	5	4	0	0.000000

31962 rows × 7 columns

In [ ]:

```
df['POS_tagged'] = nltk.pos_tag_sents(df['cleaned'].apply(nltk.word_tokenize).tolist())
df
```

Out[ ]:

	label	tweet	cleaned	word_count	unique_words_count	puntuation_count	polarity	POS_tagged
0	0	@user when a father is dysfunctional and is s...	father dysfunctional selfish drags kids dysfun...	16	14	1	-0.500000	[(father, RBR), (dysfunctional, JJ), (selfish,...
1	0	@user @user thanks for #lyft credit i can't us...	thanks lyft credit cannot use cause offer whee...	17	17	1	0.200000	[(thanks, NNS), (lyft, VBP), (credit, NN), (ca...
2	0	bihday your majesty	bihday majesty	3	3	0	0.000000	[(bihday, NN), (majesty, NN)]
3	0	#model i love u take with u all the time in ...	model love take time ur!!!	9	9	3	0.976562	[(model, NN), (love, VB), (take, NN), (time, N...
4	0	factsguide: society now #motivation	factsguide: society motivation	4	4	1	0.000000	[(factsguide, NN), (:, :), (society, NN), (mot...
...	...	...	...	...	...	...	...	...
31957	0	ate @user isz that youuu?ðððððð...	ate isz youuu?	4	4	1	0.000000	[(ate, NN), (isz, NN), (youuu, NN), (?, .)]
31958	0	to see nina turner on the airwaves trying to...	see nina turner airwaves trying wrap mantle ge...	22	20	1	0.400000	[(see, VB), (nina, JJ), (turner, NN), (airwave...
31959	0	listening to sad songs on a monday morning otw...	listening sad songs monday morning otw work sad	12	10	0	-0.500000	[(listening, VBG), (sad, JJ), (songs, NNS), (m...
31960	1	@user #sikh #temple vandalised in in #calgary,...	sikh temple vandalised calgary, wso condemns act	9	8	1	0.000000	[(sikh, JJ), (temple, NNS), (vandalised, VBD),...
31961	0	thank you @user for you follow	thank follow	5	4	0	0.000000	[(thank, NN), (follow, NN)]

31962 rows × 8 columns

Converting all POS tags to a sentence form

In [ ]:

```
def sepPOS(text):
    keep = []
    for tup in text:
        keep.append(tup[1])

    tags = ' '.join(keep)
    return tags
```

In [ ]:

```
df['POS_tagged'] = df['POS_tagged'].apply(sepPOS)
df
```

Out[ ]:

	label	tweet	cleaned	word_count	unique_words_count	puntuation_count	polarity	POS_tagged
0	0	@user when a father is dysfunctional and	father dysfunctional selfish drags	16	14	1	-0.500000	RBR JJ JJ NNS NNS

	label	tweet	cleaned	word_count	unique_words_count	puntuation_count	polarity	POS_tagged
1	0	@user @user thanks for #lyft credit i can't us...	thanks lyft credit cannot use cause offer whee...	17	17	1	0.200000	NNS VBP NN MD RB VB NN NN NN NNS VBP . VBN VBD
2	0	bihday your majesty	bihday majesty	3	3	0	0.000000	NN NN
3	0	#model i love u take with u all the time in ...	model love take time ur!!!	9	9	3	0.976562	NN VB NN NN JJ . . .
4	0	factsguide: society now #motivation	factsguide: society motivation	4	4	1	0.000000	NN : NN NN
...	...	...	...	...	...	...	...	...
31957	0	ate @user isz that youuu?ðððððð...	ate isz youuu?	4	4	1	0.000000	NN NN NN .
31958	0	to see nina turner on the airwaves trying to...	see nina turner airwaves trying wrap mantle ge...	22	20	1	0.400000	VB JJ NN NNS VBG NN FW JJ NN IN NN NN . NN NN
31959	0	listening to sad songs on a monday morning otw...	listening sad songs monday morning otw work sad	12	10	0	-0.500000	VBG JJ NNS JJ NN NN NN NN
31960	1	@user #sikh #temple vandalised in in #calgary,...	sikh temple vandalised calgary, wso condemns act	9	8	1	0.000000	JJ NNS VBD JJ , JJ NN NN
31961	0	thank you @user for you follow	thank follow	5	4	0	0.000000	NN NN

31962 rows × 8 columns

Downloading word2vec pretrained model 'glove-wiki-gigaword-300' and using it to calculate cosine similarity

In [ ]:

```
glove_vectors = gensim.downloader.load('glove-wiki-gigaword-300')
```

In [ ]:

```
def similarity(sen1,sen2):
    new=""
    for word in sen1.split():
        if(word) in glove_vectors.vocab:
            new= new + " " + word
    if(len(new)) < 1:
        new = ";"
    sim = glove_vectors.wv.n_similarity(new.split(),sen2.split())
    return sim
```

List of hate speech words. These are taken out manually from our train data.

In [ ]:

```
hate_speech = "fight xenophobia people black supremacy racism fuck bitch jewish trash pr
ejudice leftist jew trump mock racist fake sex evil violation genocide leftist zionism ca
pitalism communist islamic extremists islam nude nigger nigga nazis"
df['cosine_similarity'] = df['cleaned'].apply(lambda x: similarity(x, hate_speech))
df
```

```
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:8: DeprecationWarning: Call
```

```
to deprecated `wv` (Attribute will be removed in 4.0.0, use self instead).

/usr/local/lib/python3.6/dist-packages/gensim/matutils.py:737: FutureWarning: Conversion
of the second argument of issubdtype from `int` to `np.signedinteger` is deprecated. In f
uture, it will be treated as `np.int64 == np.dtype(int).type`.
  if np.issubdtype(vec.dtype, np.int):
```

Out[ ]:

	label	tweet	cleaned	word_count	unique_words_count	puntuation_count	polarity	POS_tagged	cosi
0	0	@user when a father is dysfunctional and is s...	father dysfunctional selfish drags kids dysfun...	16	14	1	-0.500000	RBR JJ JJ NNS NNS NN . VB	
1	0	@user @user thanks for #lyft credit i can't us...	thanks lyft credit cannot use cause offer whee...	17	17	1	0.200000	NNS VBP NN MD RB VB NN NN NN NNS VBP . VBN VBD	
2	0	bihday your majesty	bihday majesty	3	3	0	0.000000	NN NN	
3	0	#model i love u take with u all the time in ...	model love take time ur!!!	9	9	3	0.976562	NN VB NN NN JJ . . .	
4	0	factsguide: society now #motivation	factsguide: society motivation	4	4	1	0.000000	NN : NN NN	
...	...	...	...	...	...	...	...	...	
31957	0	ate @user isz that youuu?ðððððð...	ate isz youuu?	4	4	1	0.000000	NN NN NN .	
31958	0	to see nina turner on the airwaves trying to...	see nina turner airwaves trying wrap mantle ge...	22	20	1	0.400000	VB JJ NN NNS VBG NN FW JJ NN IN NN NN . NN NN	
31959	0	listening to sad songs on a monday morning otw...	listening sad songs monday morning otw work sad	12	10	0	-0.500000	VBG JJ NNS JJ NN NN NN NN	
31960	1	@user #sikh #temple vandalised in in #calgary,...	sikh temple vandalised calgary, wso condemns act	9	8	1	0.000000	JJ NNS VBD JJ , JJ NN NN	
31961	0	thank you @user for you follow	thank follow	5	4	0	0.000000	NN NN	

31962 rows × 9 columns

In [ ]:

```
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(df[['cleaned', 'word_count', 'unique
_words_count', 'POS_tagged', 'puntuation_count', 'polarity']], df['label'], random_state
= 1, train_size = 0.8)
```

In [ ]:

```
x_train
```

Out[ ]:

	cleaned	word_count	unique_words_count	POS_tagged	puntuation_count	polarity
2051	females worry good niggaz take good care den kids	19	17	NNS VBP JJ NNS VBP JJ NN JJ NNS	0	0.70000
20151	euro2016 marseille england russia france tearg...	11	11	NN NNS VBP JJ NN NN NNS VBP NN , JJ NN	1	0.00000
6595	ego, suppose? happening "me" happening..	9	9	NN , VB . VBG `` PRP '' NN	6	0.00000
8676	love puppy labicha yelbicho model puppy barcel...	12	11	VB JJ NN NN NN JJ NN NN NN	0	0.50000
13588	lighttherapy help depression? altwaystoheal he...	11	11	NN NN NN . JJ JJ JJ . .	3	0.75000
...	...	...	...	...	...	...
17289	remember lost empire dreams success goals aim ...	11	11	VB VBN NN NN NN NNS VBP NN NN NN	0	0.30000
5192	justice served bosmatrial	5	5	NN VBD JJ	0	0.00000
12172	repurposed former mustard jar beaut little vas...	13	13	VBN JJ NN NN NN JJ JJ NN NN	0	-0.09375
235	happiest baby ever known cute smiles babygirl ...	13	13	NN NN RB VBN NN NNS VBP JJ RB VBD NN	0	0.67500
29733	ased bull up: dominate bull direct whatever wa...	21	15	VBN NN IN : JJ NN JJ WDT VBP VBP .	2	0.10000

25569 rows × 6 columns

In [ ]:

```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn_pandas import DataFrameMapper
mapper = DataFrameMapper([
    (['word_count', 'unique_words_count', 'puntuation_count'], None),
    ('cleaned',CountVectorizer()),
    ('POS_tagged', CountVectorizer())
])
train1 = mapper.fit_transform(x_train)
test1 = mapper.transform(x_test)
```

In [ ]:

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn_pandas import DataFrameMapper
mapper = DataFrameMapper([
    (['word_count', 'unique_words_count', 'puntuation_count', 'polarity'], None),
    ('cleaned',TfidfVectorizer()),
    ('POS_tagged', TfidfVectorizer())
])
train2 = mapper.fit_transform(x_train)
test2 = mapper.transform(x_test)
```

In [ ]:

```
train1.shape
```

Out[ ]:

```
(25569, 33801)
```

In [ ]:

```
train2.shape
```

Out[ ]:

```
(25569, 33802)
```

Naive Bayes (Baseline)

In [ ]:

```
from sklearn.naive_bayes import MultinomialNB
nv = MultinomialNB()
nv.fit(train1, y_train)
```

Out[ ]:

```
MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True)
```

In [ ]:

```
pred_nb = nv.predict(test1)
```

In [ ]:

```
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
print('Accuracy score: ', format(accuracy_score(y_test, pred_nb)))
print('Precision score: ', format(precision_score(y_test, pred_nb)))
print('Recall score: ', format(recall_score(y_test, pred_nb)))
print('F1 score: ', format(f1_score(y_test, pred_nb)))
```

```
Accuracy score:  0.9505709369623025
Precision score:  0.9915966386554622
Recall score:  0.27251732101616627
F1 score:  0.427536231884058
```

linearSVC (1st Approach)

In [ ]:

```
from sklearn import svm
lin_clf = svm.LinearSVC(max_iter=10000)
lin_clf.fit(train2, y_train)
```

```
/usr/local/lib/python3.6/dist-packages/sklearn/svm/_base.py:947: ConvergenceWarning: Libl
inear failed to converge, increase the number of iterations.
  "the number of iterations.", ConvergenceWarning)
```

Out[ ]:

```
LinearSVC(C=1.0, class_weight=None, dual=True, fit_intercept=True,
          intercept_scaling=1, loss='squared_hinge', max_iter=10000,
          multi_class='ovr', penalty='l2', random_state=None, tol=0.0001,
          verbose=0)
```

In [ ]:

```
pred_svc = lin_clf.predict(test2)
```

In [ ]:

```
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
print('Accuracy score: ', format(accuracy_score(y_test, pred_svc)))
print('Precision score: ', format(precision_score(y_test, pred_svc)))
print('Recall score: ', format(recall_score(y_test, pred_svc)))
print('F1 score: ', format(f1_score(y_test, pred_svc)))
```

```
Accuracy score:  0.9643359924917879
Precision score:  0.8317152103559871
Recall score:  0.5935334872979214
F1 score:  0.692722371967655
```

Deep learning/Neural network

In [ ]:

```
import numpy as np
from tensorflow import keras
from tensorflow.keras import layers
```

In [ ]:

```
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(df['cleaned'], df['label'], random_s
tate = 1, train_size = 0.85)
```

In [ ]:

```
y_train.value_counts()
```

Out[ ]:

```
0    25264
1     1903
Name: label, dtype: int64
```

In [ ]:

```
from tensorflow.keras.preprocessing.text import Tokenizer  # to encode text to integers

# ENCODE Spanish source sentences
tok = Tokenizer(lower=False)
tok.fit_on_texts(x_train)

x_train = tok.texts_to_sequences(x_train) # encode the train data into integers

x_test = tok.texts_to_sequences(x_test) # encode the test data into integers

total_words = len(tok.word_index) + 1   # adding 1 because of 0 padding
```

In [ ]:

```
max_features = 30000  # Only consider the top 30k words
maxlen = 100  # Only consider the first 100 words of each tweet
EMBEDDING_SIZE=32
VOCAB_SIZE=total_words
```

In [ ]:

```
from keras.models import Sequential
from keras.layers import Dense, Softmax, Dropout, Activation
from keras.layers import SimpleRNN, LSTM, Embedding, Bidirectional
from keras.utils import to_categorical

model = Sequential()
model.add(Embedding(VOCAB_SIZE, EMBEDDING_SIZE, input_length=100))
model.add(Bidirectional(SimpleRNN(25), merge_mode='concat'))
model.add(Dense(2, activation = 'sigmoid'))
model.add(Dropout(0.2)) #Adding dropout of 0.2
model.add(Dense(2, activation = 'sigmoid'))
model.summary()
```

```
Model: "sequential_4"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_4 (Embedding)      (None, 100, 32)           1142432
_________________________________________________________________
bidirectional_4 (Bidirection (None, 50)                2900
_________________________________________________________________
dense_7 (Dense)              (None, 2)                 102
_________________________________________________________________
dropout_4 (Dropout)          (None, 2)                 0
```

```
_________________________________________________________________
dense_8 (Dense)              (None, 2)                 6
=================================================================
Total params: 1,145,440
Trainable params: 1,145,440
Non-trainable params: 0
_________________________________________________________________
```

In [ ]:

```
from keras.preprocessing.sequence import pad_sequences

x_train = pad_sequences(x_train, maxlen=maxlen, value=0, padding='post', truncating='pos
t')
x_test = pad_sequences(x_test, maxlen=maxlen, value=0, padding='post', truncating='post'
)
```

Adding Early Stopping

In [ ]:

```
from tensorflow.keras.callbacks import EarlyStopping

model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy', 'Precis
ion', 'Recall'])
callback = EarlyStopping(monitor='loss', patience=5)
bi_history = model.fit(x_train, to_categorical(y_train), epochs=50, validation_split=0.1
0, callbacks = [callback])
```

```
Epoch 1/50
765/765 [==============================] - 106s 138ms/step - loss: 0.0709 - accuracy: 0.9
627 - precision: 0.9627 - recall: 0.9627 - val_loss: 0.3254 - val_accuracy: 0.9382 - val_
precision: 0.9382 - val_recall: 0.9382
Epoch 2/50
765/765 [==============================] - 106s 138ms/step - loss: 0.0697 - accuracy: 0.9
625 - precision: 0.9625 - recall: 0.9625 - val_loss: 0.3221 - val_accuracy: 0.9260 - val_
precision: 0.9257 - val_recall: 0.9260
Epoch 3/50
765/765 [==============================] - 105s 137ms/step - loss: 0.0706 - accuracy: 0.9
617 - precision: 0.9617 - recall: 0.9617 - val_loss: 0.3285 - val_accuracy: 0.9194 - val_
precision: 0.9194 - val_recall: 0.9194
Epoch 4/50
765/765 [==============================] - 107s 139ms/step - loss: 0.0704 - accuracy: 0.9
617 - precision: 0.9617 - recall: 0.9617 - val_loss: 0.3521 - val_accuracy: 0.9396 - val_
precision: 0.9396 - val_recall: 0.9396
Epoch 5/50
765/765 [==============================] - 108s 141ms/step - loss: 0.0694 - accuracy: 0.9
630 - precision: 0.9630 - recall: 0.9630 - val_loss: 0.3951 - val_accuracy: 0.9404 - val_
precision: 0.9404 - val_recall: 0.9404
Epoch 6/50
765/765 [==============================] - 107s 140ms/step - loss: 0.0696 - accuracy: 0.9
629 - precision: 0.9629 - recall: 0.9629 - val_loss: 0.3409 - val_accuracy: 0.9260 - val_
precision: 0.9260 - val_recall: 0.9260
Epoch 7/50
765/765 [==============================] - 107s 140ms/step - loss: 0.0683 - accuracy: 0.9
636 - precision: 0.9636 - recall: 0.9636 - val_loss: 0.3740 - val_accuracy: 0.9308 - val_
precision: 0.9308 - val_recall: 0.9308
Epoch 8/50
765/765 [==============================] - 108s 141ms/step - loss: 0.0701 - accuracy: 0.9
620 - precision: 0.9620 - recall: 0.9620 - val_loss: 0.3893 - val_accuracy: 0.9282 - val_
precision: 0.9282 - val_recall: 0.9282
Epoch 9/50
765/765 [==============================] - 108s 141ms/step - loss: 0.0728 - accuracy: 0.9
596 - precision: 0.9596 - recall: 0.9596 - val_loss: 0.3831 - val_accuracy: 0.9308 - val_
precision: 0.9308 - val_recall: 0.9308
Epoch 10/50
765/765 [==============================] - 110s 144ms/step - loss: 0.0684 - accuracy: 0.9
636 - precision: 0.9636 - recall: 0.9636 - val_loss: 0.3967 - val_accuracy: 0.9326 - val_
precision: 0.9327 - val_recall: 0.9330
Epoch 11/50
765/765 [==============================] - 109s 143ms/step - loss: 0.0708 - accuracy: 0.9
613 - precision: 0.9613 - recall: 0.9613 - val_loss: 0.4112 - val_accuracy: 0.9293 - val_
```

```
precision: 0.9293 - val_recall: 0.9293
Epoch 12/50
765/765 [==============================] - 111s 145ms/step - loss: 0.0686 - accuracy: 0.9
633 - precision: 0.9633 - recall: 0.9633 - val_loss: 0.3982 - val_accuracy: 0.9319 - val_
precision: 0.9316 - val_recall: 0.9319
```

In [ ]:

```
loss, acc, precision, recall = model.evaluate(x_test, to_categorical(y_test))
print("Test accuracy: %0.2f%%"%(acc*100))
print("Test precision: %0.2f%%"%(precision*100))
print("Test recall: %0.2f%%"%(recall*100))
```

```
150/150 [==============================] - 2s 16ms/step - loss: 0.3856 - accuracy: 0.9449
- precision: 0.9449 - recall: 0.9449
Test accuracy: 94.49%
Test precision: 94.49%
Test recall: 94.49%
```