Week 8

Dataset cleaning and preprocessing

Team Members:

Member no.	Name	Email	Country	college	Specialization
1	Zyad Hussein	Zyadrashad262@yahoo.com	Egyptian	University of York	NLP

Problem description:

To train a deep learning model for hate speech detection, a good, clean and precise dataset is needed. The dataset provided contained special characters, numbers mixed with letters, Unicode and other unidentifiable characters. It is believed that these characters represented a meaningful sentences and words before adding them to the csv file. It is also believed that the csv file might have lost some of these symbols or misinterpreted them which resulted in unidentifiable characters with no meaning. The task is to clean the dataset and prepare it for model training to achieve high accuracy results.

While data preparation was in process, it was noticed that some tweets along the dataset had mixed feelings and some other tweets were mislabelled. It was believed that these tweets can be removed to avoid any loss in accuracy. However, it is currently believed that leaving them might yield a robust model that can detect hard identifying (mixed feeling) tweets. The decision to remove them or leave them is dependent on the outcome of the model. Therefore, these tweets were left until model accuracy results are obtained. The method of removing these tweets automatically is still being discovered. Figure 1 shows the dataset loaded and the anomalies, special characters, and meaningless characters in it. Figure 2 shows the result of the data after applying regex and other python methods to clean the dataset. It is worth noting that the hashtag symbol was left as it is believed that hashtags hold meaningful information that can help the model identify the speech type.

loading data using dask as it is fast enough to work with and more organised method besides that it has many functions that can help in tthe manipulation of the dataset.

In [33]:
```
import dask.dataframe as dd
import re
train=dd.read_csv('train_E6oV3lV.csv')
test=dd.read_csv('test_tweets_anuFYb8.csv')
```

In [6]:
```
train.compute()
```

Out[6]:

	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?ð□□□ð□□□ð□□□ð□□□ð...
31958	31959	0	to see nina turner on the airwaves trying to...
31959	31960	0	listening to sad songs on a monday morning otw...
31960	31961	1	@user #sikh #temple vandalised in in #calgary,...
31961	31962	0	thank you @user for you follow

31962 rows × 3 columns

Fig. 1: data loading.

In [*]:
```
train['tweet'].compute()
for i in range(0,31962):

    x=train.loc[i,'tweet'].values
    f=x.compute()
    #this is cleaning a sample text extracted from the dataset using regex as requested.
    #
    cleaning = re.sub(r'[^\w\s\d#]', '', f[0])#remove anything that s not words, numbers or spaces
    #but not hashtags as they contain good information
    cleaning = re.sub(r'@\w+', '', cleaning)# removing @user (optional)
    #cleaning = re.sub(r'\s+', ' ', cleaning).strip()#removing whitespaces
    cleaning = re.sub(r'http\S+|www\S+|https\S+', '', cleaning)#removing urls
    cleaned = cleaning.encode('ascii', 'ignore').decode('ascii')#remove unicode characters
    data.append(cleaned)
```

data

In [57]:
```
data
```

Out[57]:
```
[' user when a father is dysfunctional and is so selfish he drags his kids into his dysfunction   run',
 'user user thanks for lyft credit i cant use cause they dont offer wheelchair vans in pdx    disapointed getthanked',
 '  bihday your majesty',
 'model   i love u take with u all the time in ur   ',
 ' factsguide society now    motivation',
 '22 huge fan fare and big talking before they leave chaos and pay disputes when they get there allshowandnogo  ',
 ' user camping tomorrow user user user user user user danny',
 'the next school year is the year for exams cant think about that  school exams   hate imagine actorslife revolutionschoo
l girl',
 'we won love the land allin cavs champions cleveland clevelandcavaliers   ',
 ' user user welcome here   im   its so gr8  ',
 '  ireland consumer price index mom climbed from previous 02 to 05 in may   blog silver gold forex',
 'we are so selfish orlando standwithorlando pulseshooting orlandoshooting biggerproblems selfish heabreaking   values lov
e ',
 'i get to see my daddy today   80days gettingfed',
 'user cnn calls michigan middle school build the wall chant  tcot  ',
 'no comment  in australia   opkillingbay seashepherd helpcovedolphins thecove  helpcovedolphins',
```

Fig.2: data cleaning using regex and its results.

In [75]:
```
data2=[]
def correct(word):
    # Apply the necessary correction logic to fix the word
    # Example: Replace 'q' with 'g'
    corrected_word = word.replace('user', "")
    return corrected_word
for row in data:
    cleaner=row.replace('user', "")
    data2.append(cleaner)
```

Now the data could be considered as eady for model training. model that will be used is BERT which is short for Bidirectional Encoder Representations from Transformers. such model has been a revolutionary point in NLP when introduced and it laid foundation for other models uch as ROBERTa, XLNet and many more NLP text classification models. it utilizes transformer architecture which was a requirement by Data Glacier. the architecture will be explained more in detail in implementation code file.

In [76]:
```
data2
```

Out[76]:
```
['  when a father is dysfunctional and is so selfish he drags his kids into his dysfunction   #run',
 '  thanks for #lyft credit i cant use cause they dont offer wheelchair vans in pdx    #disapointed #getthanked',
 '  bihday your majesty',
 '#model   i love u take with u all the time in ur   ',
 ' factsguide society now    #motivation',
 '22 huge fan fare and big talking before they leave chaos and pay disputes when they get there #allshowandnogo  ',
 '  camping tomorrow        danny',
 'the next school year is the year for exams cant think about that  #school #exams   #hate #imagine #actorslife #revolutio
nschool #girl',
 'we won love the land #allin #cavs #champions #cleveland #clevelandcavaliers   ',
 '   welcome here   im   its so #gr8  ',
 '  #ireland consumer price index mom climbed from previous 02 to 05 in may   #blog #silver #gold #forex',
 'we are so selfish #orlando #standwithorlando #pulseshooting #orlandoshooting #biggerproblems #selfish #heabreaking   #va
lues #love #',
 'i got to see my daddy today   #80days #gettingfed'
```

Fig 3: removing user from all the dataset.