ICP 4

Submitted By: Dhairya Chandra

What you learned in the ICP

I have learned data cleaning and preprocessing using NLP. Also I learned about data visualization, building TFIDF models and Deep learning models

ICP description what was the task you were performing

Data cleaning and preprocessing (at minimum have the following: Removing unnecessary columns or data, Removing Twitter Handles(@user), Removing punctuation, numbers, special characters, Removing stop words, Tokenization, and Stemming, TFIDF vectors, POS tagging, checking for missing values , train/test split of data), Data Visualization and analysis for critical steps, deep learning model building using liner SVM model and MLP Classifier

Challenges that you faced

I have some issue in building models but I got help from towards data science, medium and stack overflow.

Screen shots that shows the successful execution of each required step of your code :

Data Cleaning and Preprocessing

```
# Importing Dataset

data = pd.read_csv('https://raw.githubusercontent.com/dD2405/Twitter_Sentiment_Analysis/master/train.csv')
data.head(10)
```

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

```
# Removing Twitter Handles (@username)

data['cleaned_tweet'] = data['tweet'].replace(to_replace="@[\w]*",value='',regex=True)
data.head(10)
```

	id	label	tweet	cleaned_tweet
0	1	0	@user when a father is dysfunctional and is s...	when a father is dysfunctional and is so sel...
1	2	0	@user @user thanks for #lyft credit i can't us...	thanks for #lyft credit i can't use cause th...
2	3	0	bihday your majesty	bihday your majesty
3	4	0	#model i love u take with u all the time in ...	#model i love u take with u all the time in ...
4	5	0	factsguide: society now #motivation	factsguide: society now #motivation
5	6	0	[2/2] huge fan fare and big talking before the...	[2/2] huge fan fare and big talking before the...
6	7	0	@user camping tomorrow @user @user @user @use...	camping tomorrow dannyâ¦
7	8	0	the next school year is the year for exams.ð...	the next school year is the year for exams.ð...
8	9	0	we won!!! love the land!!! #allin #cavs #champ...	we won!!! love the land!!! #allin #cavs #champ...
9	10	0	@user @user welcome here ! i'm it's so #gr...	welcome here ! i'm it's so #gr8 !

```
# Removing Numbers

data['cleaned_tweet'] = data['cleaned_tweet'].apply(lambda x: re.sub(r'[^a-zA-Z]',' ',x))
data.head(10)
```

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

```
# Removing all special characters

data['cleaned_tweet'] = data['cleaned_tweet'].apply(lambda x: re.sub(r'[^a-zA-Z0-9]',' ',x))
data.head(10)
```

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

Preprocessing Data

```
# Tokenization

cleaned_text = data['cleaned_tweet']

preprocessed_data = cleaned_text.apply(lambda tweet:word_tokenize(tweet))
preprocessed_data.head(20)
```

```
0     [when, a, father, is, dysfunctional, and, is, ...
1     [thanks, for, lyft, credit, i, can, t, use, ca...
2                                [bihday, your, majesty]
3     [model, i, love, u, take, with, u, all, the, t...
4                [factsguide, society, now, motivation]
5     [huge, fan, fare, and, big, talking, before, t...
6                             [camping, tomorrow, danny]
7     [the, next, school, year, is, the, year, for, ...
8     [we, won, love, the, land, allin, cavs, champi...
9                   [welcome, here, i, m, it, s, so, gr]
10    [ireland, consumer, price, index, mom, climbed...
11    [we, are, so, selfish, orlando, standwithorlan...
12    [i, get, to, see, my, daddy, today, days, gett...
13    [cnn, calls, michigan, middle, school, build, ...
14    [no, comment, in, australia, opkillingbay, sea...
15    [ouch, junior, is, angry, got, junior, yugyoem...
16    [i, am, thankful, for, having, a, paner, thank...
17                             [retweet, if, you, agree]
18    [its, friday, smiles, all, around, via, ig, us...
19    [as, we, all, know, essential, oils, are, not,...
Name: cleaned_tweet, dtype: object
```

```
[52] # List of Stopwords

stop_words = set(stopwords.words('english'))
print(stop_words)
```

```
{'theirs', "mightn't", 'doing', 'don', 'for', "couldn't", 'herself', 'our', 'only', 'will', 'his', 'themselve
```

```
# Removing the Stopwords

preprocessed_data_sw = preprocessed_data.apply(lambda x: [word for word in x if not word in stop_words])
preprocessed_data_sw.head(20)
```

```
0     [father, dysfunctional, selfish, drags, kids, ...
1     [thanks, lyft, credit, use, cause, offer, whee...
2                                  [bihday, majesty]
3                  [model, love, u, take, u, time, ur]
4                    [factsguide, society, motivation]
5     [huge, fan, fare, big, talking, leave, chaos, ...
6                           [camping, tomorrow, danny]
7     [next, school, year, year, exams, think, schoo...
8     [love, land, allin, cavs, champions, cleveland...
9                                    [welcome, gr]
10    [ireland, consumer, price, index, mom, climbed...
11    [selfish, orlando, standwithorlando, pulseshoo...
12          [get, see, daddy, today, days, gettingfed]
13    [cnn, calls, michigan, middle, school, build, ...
14    [comment, australia, opkillingbay, seashepherd...
15     [ouch, junior, angry, got, junior, yugyoem, omg]
16               [thankful, paner, thankful, positive]
17                                 [retweet, agree]
18    [friday, smiles, around, via, ig, user, cookie...
19             [know, essential, oils, made, chemicals]
Name: cleaned_tweet, dtype: object
```

```
# Stemming

from nltk.stem import PorterStemmer
stemming = PorterStemmer()

data_stemming = preprocessed_data_sw.apply(lambda token: ' '.join([stemming.stem(i) for i in token]))
data_stemming.head(20)
```

```
0          father dysfunct selfish drag kid dysfunct run
1     thank lyft credit use caus offer wheelchair va...
2                                       bihday majesti
3                             model love u take u time ur
4                                factsguid societi motiv
5     huge fan fare big talk leav chao pay disput ge...
6                                    camp tomorrow danni
7     next school year year exam think school exam h...
8     love land allin cav champion cleveland clevela...
9                                             welcom gr
10    ireland consum price index mom climb previou m...
11    selfish orlando standwithorlando pulseshoot or...
12                    get see daddi today day gettingf
13    cnn call michigan middl school build wall chan...
14    comment australia opkillingbay seashepherd hel...
15             ouch junior angri got junior yugyoem omg
16                               thank paner thank posit
17                                        retweet agre
18     friday smile around via ig user cooki make peopl
19                           know essenti oil made chemic
Name: cleaned_tweet, dtype: object
```

PoS Tagging

```
# Converting string into words for first tweet

nltk_tokens = nltk.word_tokenize(data_stemming[0])
nltk_tokens
```

```
['father', 'dysfunct', 'selfish', 'drag', 'kid', 'dysfunct', 'run']
```

```
[96] for words in nltk_tokens:
         pos_words = nltk.pos_tag(nltk_tokens)
     pos_words
```

```
[('father', 'RB'),
 ('dysfunct', 'JJ'),
 ('selfish', 'JJ'),
 ('drag', 'NN'),
 ('kid', 'NN'),
 ('dysfunct', 'NN'),
 ('run', 'VB')]
```

TFIDF

```
#coverting the comments into list
comment_list = data_stemming.tolist()
comment_list
```

```
 'ferrari sake championship gp clearli turn point rb ferrari merc',
 'ace first test proud',
 'seek probe udtapunjab leak point finger amarind aap',
 'wrap senseaboutmath th',
 'hey white peopl call peopl white race ident med',
 'might shown today regurgit talk point name call',
 'sometim rais brow rais bar golfstrengthandcondit strong felixfoisgolf',
 'greathonour careerconvo',
 'design innov learn space includ wateringhol cave mountaintop campfir h',
 'altright use amp insecur lure men whitesupremaci',
 'carri gun help take gun control stop black market terror get wors',
 'use power mind heal bodi altwaystoh healthi peac',
```

```
# Converting list into array

docs = np.array(comment_list)
```

```
[ ] import sklearn
    import sklearn.ensemble
    vectorizer = sklearn.feature_extraction.text.TfidfVectorizer(lowercase=False)
    train_vectors = vectorizer.fit_transform(docs)
    train_vectors

    # printing vectors
    print(train_vectors)
```

```
  (0, 23481)    0.23919104039858957
  (0, 14708)    0.22595905585564077
  (0, 7557)     0.32568517024157123
  (0, 24156)    0.33687034829612006
  (0, 7823)     0.7988916972430302
  (0, 9011)     0.18425456756777517
  (1, 10571)    0.3584147302540872
  (1, 7119)     0.3725600096100806
  (1, 20634)    0.3483784948316893
  (1, 29088)    0.32419698005329806
  (1, 30096)    0.3483784948316893
  (1, 19696)    0.26168867114052213
  (1, 4467)     0.2367878205382297
  (1, 28995)    0.1950713182941508
```

DATA VISUALIZATION

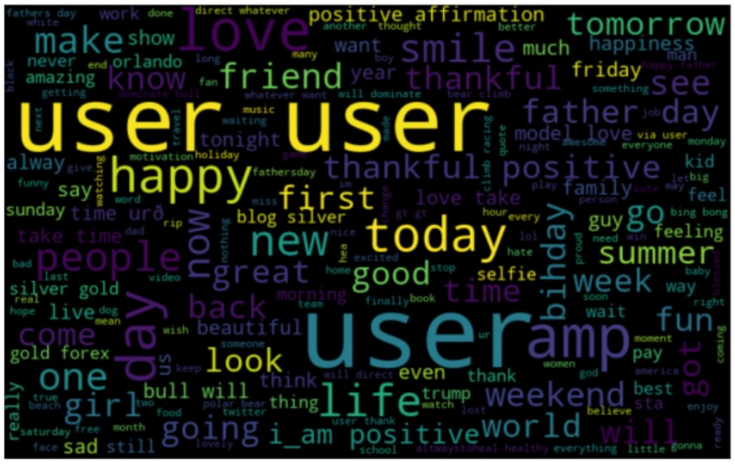
▾ Word Cloud of original data

```
# Generating Word Cloud

allWords = ' '.join([text for text in data['tweet']])

wordcloud = WordCloud(width=800, height=500, random_state=21, max_font_size=110).generate(allWords)

plt.figure(figsize=(10, 10))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.show()
```



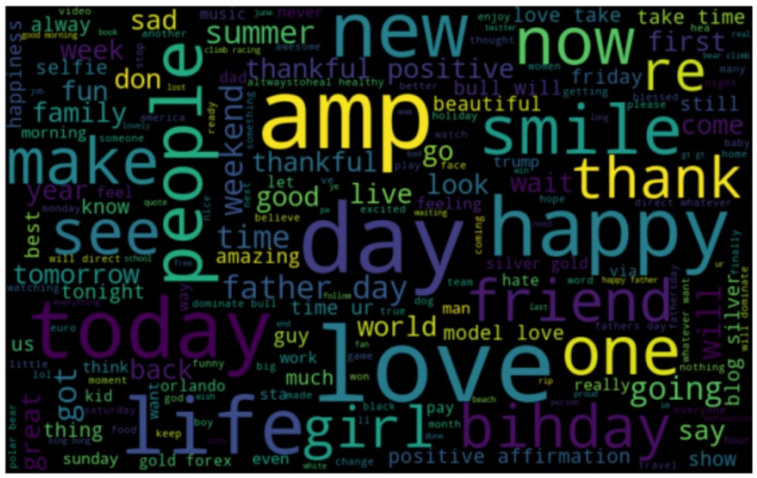
Word Cloud of cleaned data

```
# Generating Word Cloud

allWords = ' '.join([text for text in cleaned_text])

wordcloud = WordCloud(width=800, height=500, random_state=21, max_font_size=110).generate(allWords)

plt.figure(figsize=(10, 10))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.show()
```



Wordcloud after Stemming

```
# Generating Word Cloud

allWords = ' '.join([text for text in data_stemming])

wordcloud = WordCloud(width=800, height=500, random_state=21, max_font_size=110).generate(allWords)

plt.figure(figsize=(10, 10))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.show()
```

Used MLP Classifier for Deep Learning Approach

▾ Model building

```
# Splitting into 80% Training and 20% Testing data

import sklearn.model_selection as ms
train_data, test_data = ms.train_test_split(data, test_size=0.2, random_state=42, shuffle=True)

from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import TfidfTransformer, TfidfVectorizer
from sklearn.svm import SVC, NuSVC, LinearSVC
from sklearn.model_selection import GridSearchCV
from sklearn.neural_network import MLPClassifier


# Defining Parameters for the TFIDF Vectoriser and Model and then gridSearchCV for HyperParameter Tuning during training

#Linear_svm Model
params = {'tfidf__max_df': [0.9, 0.95],'tfidf__ngram_range': [(1,1), (1,2)]}

#Using Pipeline function to vectorise the tweets and then training the support vector classification machine on the vectors
pipeline = Pipeline([
    ("tfidf", TfidfVectorizer(sublinear_tf=True, stop_words='english')),

    # Used Multi-layer Perceptron classifier
    ("mlpC", MLPClassifier(hidden_layer_sizes=(2), activation='relu', solver='adam', alpha=0.0001, batch_size='auto', learning_rate='constant')),])
```

```
#Using GridSearchCV for HyperParameter Tuning
gs = GridSearchCV(pipeline, params, cv=12, verbose=2, n_jobs=-1)

#Training the model on the Cleaned training data
gs.fit(train_data['cleaned_tweet'], train_data['label'])
print(gs.best_estimator_)
print(gs.best_score_)

#Predict the Labels for the cleaned test data
predicted = gs.predict(test_data['cleaned_tweet'])

sum = 0
for p, y in zip(test_data.label, predicted):
  if p == y:
    sum = sum + 1
print('Accuracy: ', sum/len(predicted)*100)

#Final Score

from sklearn.metrics import f1_score, accuracy_score

#Printing the F1 Score and the Accuracy score
print('F1 Score : ', f1_score(test_data.label, predicted))
print('Accuracy : ',  accuracy_score(test_data.label, predicted))
```

F1 Score: 68.30%

Accuracy: 96.37%

YouTube Link: https://youtu.be/JOubUekH4T8