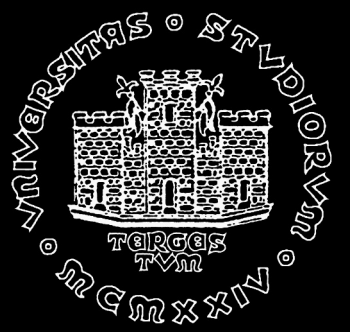
Sentiment Analysis

Sentiment analysis of tweets to find depression in people

Babaei Elham



Natural Language Processing
A.A. 2021/2022

Datasets

- 10314 tweets from Tweeter show if people are depressed

	Index	message to examine	label (depression result)
0	106	just had a real good moment. i missssssssss hi...	0
1	217	is reading manga http://plurk.com/p/mzp1e	0
2	220	@comeagainjen http://twitpic.com/2y2lx - http:...	0
3	288	@lapcat Need to send 'em to my accountant tomo...	0
4	540	ADD ME ON MYSPACE!!! myspace.com/LookThunder	0
...	...	...	...
10309	802309	No Depression by G Herbo is my mood from now o...	1
10310	802310	What do you do when depression succumbs the br...	1
10311	802311	Ketamine Nasal Spray Shows Promise Against Dep...	1
10312	802312	dont mistake a bad day with depression! everyo...	1
10313	802313	0	1

10314 rows × 3 columns

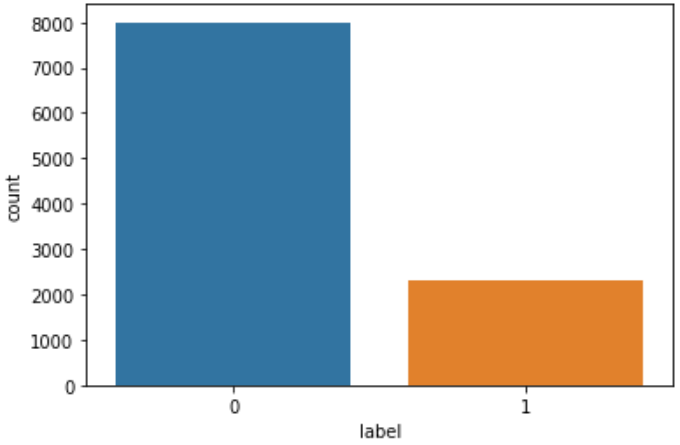
Datasets


Negative tweets- depression
0
100
200
300
400
500
0
200
400
600
800
1000


Positive tweets- Non-depression
0
100
200
300
400
500
0
200
400
600
800
1000

Datasets

Unbalanced dataset



Downsampling:

- 4614 Total Tweets
- 2314 Class 1
- 2300 Class 0

Preprocessing

- Lower casing
- Removal of Punctuations
- Removal of Stopwords
- Stemming
- Lemmatization
- Removal of emojis and symbols

Before:
@kathyIreland Hugs 4 kathy. Though my mom's in heaven, I know she's smiling at my accomplishments & all the people who keep me positive.

After:
hug kathi though mom heaven know she smile accomplish peopl keep posit

N-Grams

Uni-Grams

	word	tf	idf	tfidf
292	depress	2233	1.779664	337.807401
555	im	508	3.312862	116.632135
661	love	299	3.791252	90.107526
49	anxieti	338	3.652416	80.716458
638	like	324	3.729377	79.811210
274	day	277	3.872478	78.406171
1121	thank	186	4.229153	76.045954
472	good	251	3.977838	75.440302
609	know	233	4.053262	61.221955
1137	time	217	4.095621	60.750047

Bi-Grams

	word	tf	idf	tfidf
38	depress anxieti	129	4.565625	95.326352
4	anxieti depress	131	4.565625	93.825473
25	cure depress	52	5.462867	48.657695
140	gon na	65	5.306025	44.542782
219	mental health	55	5.481915	34.198251
152	great depress	35	5.877811	30.992298
81	depress nap	35	5.849640	29.934071
303	wan na	36	5.849640	26.805891
116	dont know	32	5.967423	21.430549
278	suffer depress	30	5.999172	21.377647

Tri-Grams

	word	tf	idf	tfidf
29	im gon na	21	6.342117	21.000000
18	depress tie kid	31	5.967423	13.863621
21	emot intellectu develop	31	5.967423	13.863621
31	kid emot intellectu	31	5.967423	13.863621
36	mom depress tie	31	5.967423	13.863621
47	tie kid emot	31	5.967423	13.863621
5	cannabi ea depress	17	6.542787	10.607038
39	puff cannabi ea	16	6.599946	10.166031
14	depress mental ill	8	7.235935	8.000000
27	happi mother day	8	7.235935	8.000000

Top Collocations

```
{'mental_health': 17.47636554621849,
 'intellectu_develop': 15.779131355932204,
 'emot_intellectu': 15.516145833333333,
 'tie_kid': 12.731196581196581,
 'puff_cannabi': 12.2825,
 'cant_wait': 8.170661157024794,
 'ist_saddepress': 8.0,
 'kid_emot': 7.638717948717948,
 'choicedepress_choicedepress': 6.320987654320987,
 'saddepress_ist': 5.359375,
 'feder_reserv': 4.5,
 'ice_cream': 4.454545454545454,
 'deepika_padukon': 3.5714285714285716,
 'harri_potter': 3.5714285714285716,
 'depress_anxieti': 3.117091131679991,
 'anxieti_depress': 3.0473081295362463,
 'reserve_caused': 3.0,
 'chemic_imbal': 2.9761904761904763,
 'mental_ill': 2.761036809089833,
 'kelli_clarkson': 2.25,
 'regularli_cut': 2.25,
 'panic_attack': 2.178649237472767,
 'gsk_catenin': 2.0,
 'modul_gsk': 2.0,
 'prospect_cohort': 2.0,
 'ronaldo_fanboy': 2.0,
 'vern_troyer': 2.0,
 'wreak_havoc': 2.0,
```

Models- Logistic Regression

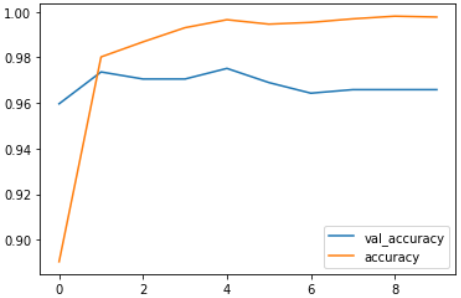
- Train: 3217 (70%)
- Test: 276 (20%)
- Validation: 110 (10%)

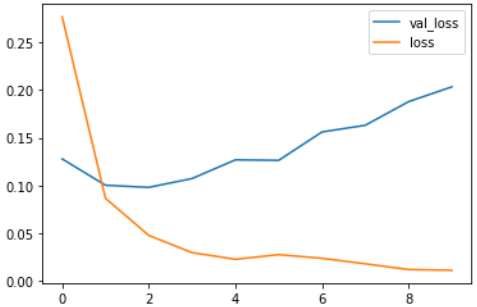
- TfidfVectorizer
- N-Grams N=1 to 2
- The size of observation*features matrix 3217*1479
- The word “depress” is a common feature among 1512 Tweets
- The coefficient for the word “depress” is 19.16
- The coefficient for the word “good” is -1.21
- Running time is few seconds

Models- BiLSTM

- Input layer with max_length equal to 110 (the 90% percentile)
- Embedding layer
- BiLSTM layer
- Dropout layer
- Output layer
- Sigmoid layer as the activation function
- #epochs 10
- Batch size 16
- Running time 20 mins

```
Model: "model_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
word_IDs (InputLayer)        [(None, 110)]             0

embeddings (Embedding)       (None, 110, 128)          128000

Bi-LSTM (Bidirectional)      (None, 512)               788480

dropout (Dropout)            (None, 512)               0

output (Dense)               (None, 1)                 513

sigmoid (Activation)         (None, 1)                 0

=================================================================
Total params: 916,993
Trainable params: 916,993
Non-trainable params: 0
```





Models- BERT

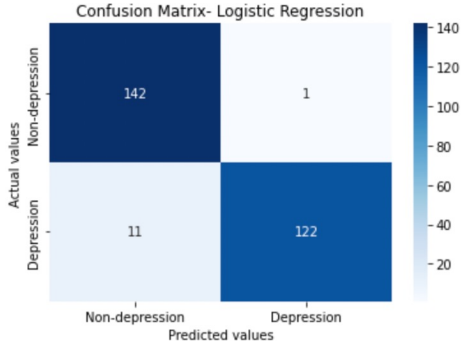
- “bert-base-uncased” tokenizer
- Learning rate 2e-5
- #epochs 5
- Batch size 16
- Running time 5 hours!

Models Evaluation

Logistic R

	precision	recall	f1-score	support
0	0.93	0.99	0.96	143
1	0.99	0.92	0.95	133
accuracy			0.96	276
macro avg	0.96	0.96	0.96	276
weighted avg	0.96	0.96	0.96	276

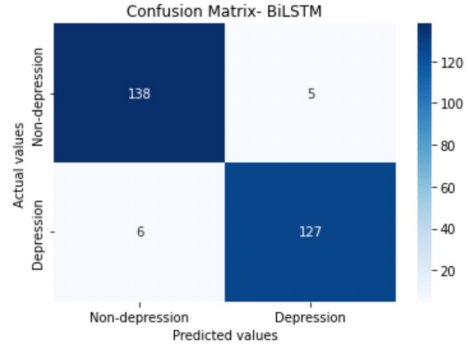
AUC= 0.95515011304485



BiLSTM

	precision	recall	f1-score	support
0	0.96	0.97	0.96	143
1	0.96	0.95	0.96	133
accuracy			0.96	276
macro avg	0.96	0.96	0.96	276
weighted avg	0.96	0.96	0.96	276

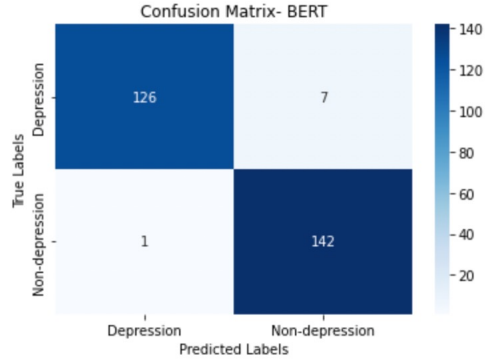
AUC= 0.9758399495241601



BERT

	precision	recall	f1-score	support
1	0.9921	0.9474	0.9692	133
0	0.9530	0.9930	0.9726	143
accuracy			0.9710	276
macro avg	0.9726	0.9702	0.9709	276
weighted avg	0.9719	0.9710	0.9710	276

AUC= 0.9701877070298123



Thank you for your attention!