In the name of God

Department of Computer Engineering

Natural Language Processing
Final Phase Report *

Yeganeh Morshedzadeh

Student Number: 96521488

Spring 2021

Contents

List of Figures

List of Tables

Abstract

In this project, we tried to use Natural Language Processing to better understand Depression and Anxiety posts. The dataset is gathered from Reddit communities r/depression and r/Anxiety.

For this project, at first, we wrote a project proposal (Google Docs), and afterwards, in the first phase (Google Docs), we gathered data and made some exploratory data analysis.

In the final phase, we went deeper, and tried various NLP tasks, such as, computing Word2Vec, Tokenization, Parsing, and creating a language model based on our the dataset.

Part I

Word2Vec

Filename: 3_word2vec.ipynb

Code

For this part we have three Word2Vec models, named as dep_w2v_model, anx_w2v_model, and all_w2v_model. Moreover, with boolean parameters, load and save, the model will be saved and/or loaded in the my_word2vec function.

Table 1: Word2Vec vocabulary size

	label	vocab_size
0	depression	2054
1	anxiety	2175
2	all	3223

Results and Examples

To make the visualizations more relevant, we will look at the relationships between a query word (in **red**), its most similar words in the model (in **blue**), and other words from the vocabulary (in **green**)
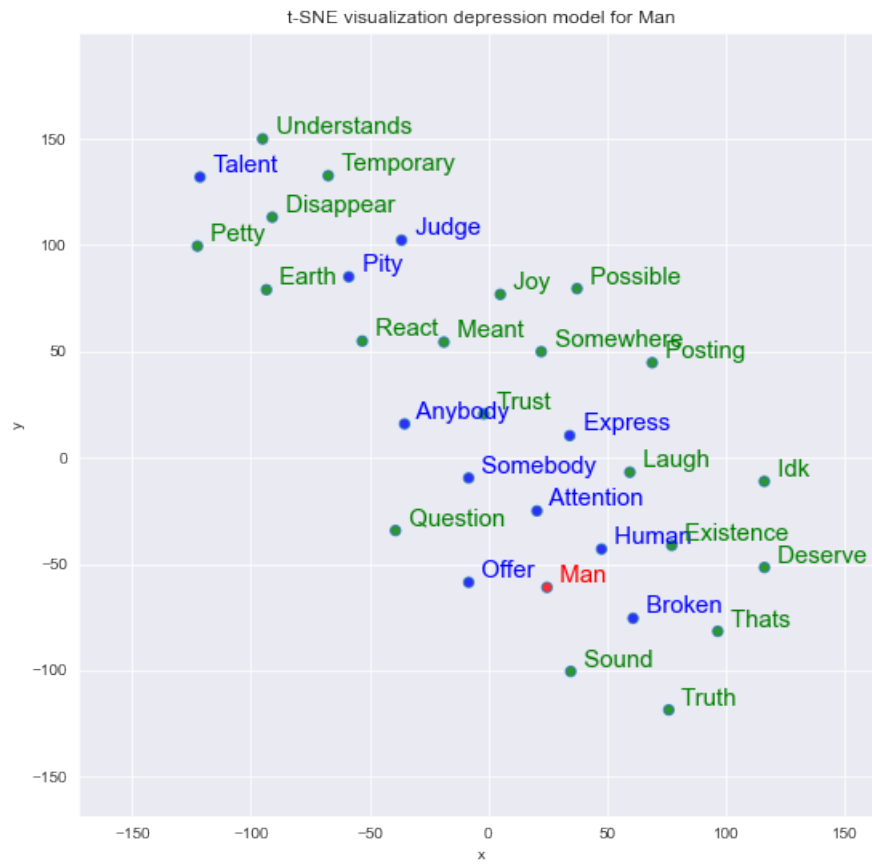
Figure 1: t-SNE visualization for man

Part II

Tokenization

Filename: 4_tokenization.ipynb

Code

In this part we have used KFold to split our data into train and test. Afterwards, we train SentencePiece model based on the data. Lastly, we compute <UNK> on our test dataset.

Results and Examples

Part III
Parsing

In this part, we used Stanza, which is a a Python NLP Package, and a collection of accurate and efficient tools for the linguistic analysis of many human languages. Starting from raw text to syntactic analysis and entity recognition, Stanza brings state-of-the-art NLP models to languages of your choosing.

More specifically, we used their Online Demo to create a manual .CoNLL file based on our dataset. Later, we can use Universal Dependencies CoNLL viewer to automatically generate parse tree from .CoNLL file.

Results and Examples

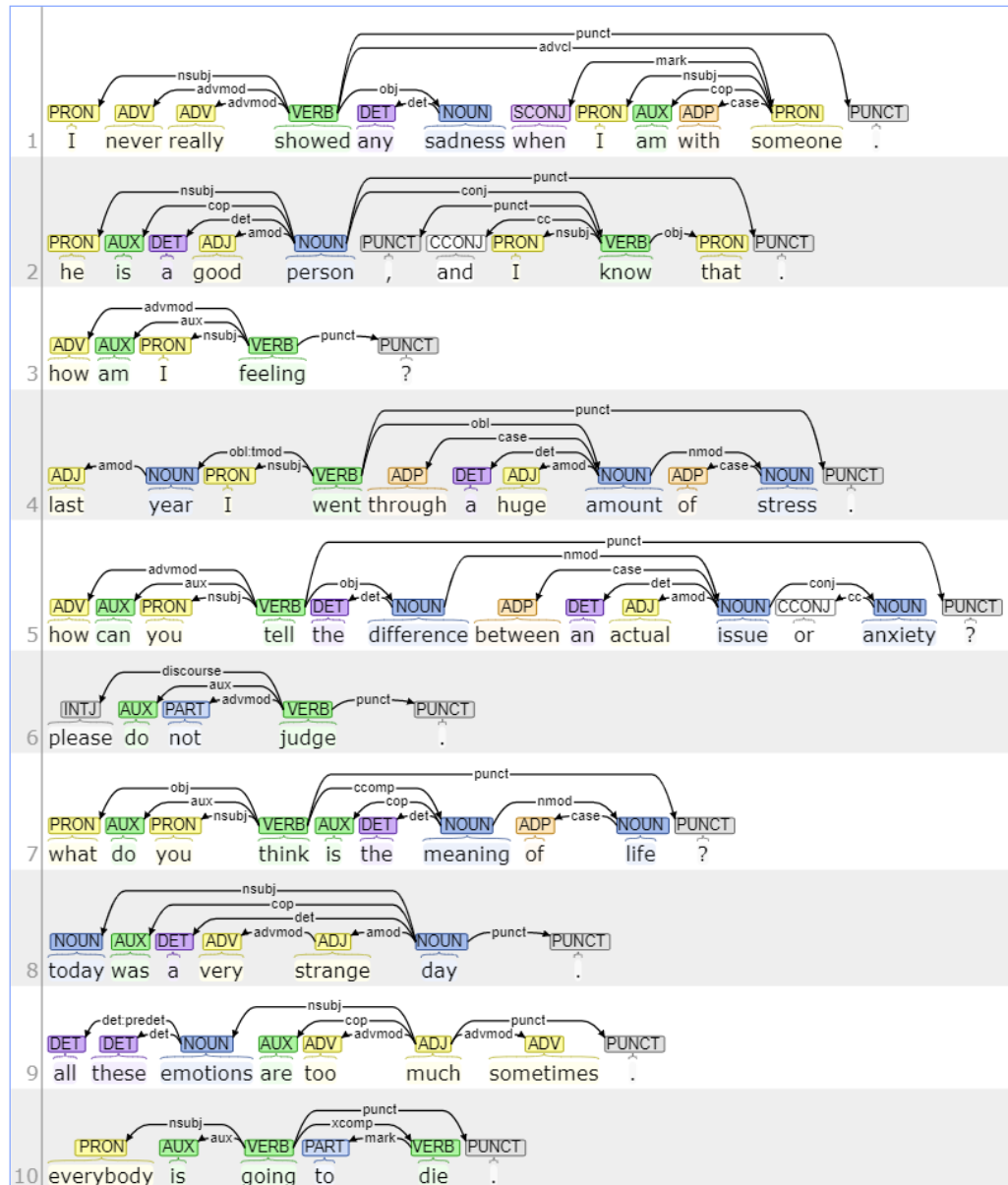The reported Unlabeled Attachment Score (UAS) on our test file was 86.05.

Figure 2: Universal Dependencies for 10 sentences

1. I never really showed any sadness when I am with someone.

2. he is a good person, and I know that.

3. how am I feeling?

4. last year I went through a huge amount of stress.

5. how can you tell the difference between an actual issue or anxiety?

6. please do not judge.

7. what do you think is the meaning of life?

8. today was a very strange day.

9. all these emotions are too much sometimes.

10. everybody is going to die.

Part IV

Language Model

Filename: 5_language-model.ipynb

Code

In this part we have used KFold to split our data into train and test. Afterwards, we train SentencePiece model based on the data. Lastly, we compute <UNK> on our test dataset.

Results and Examples

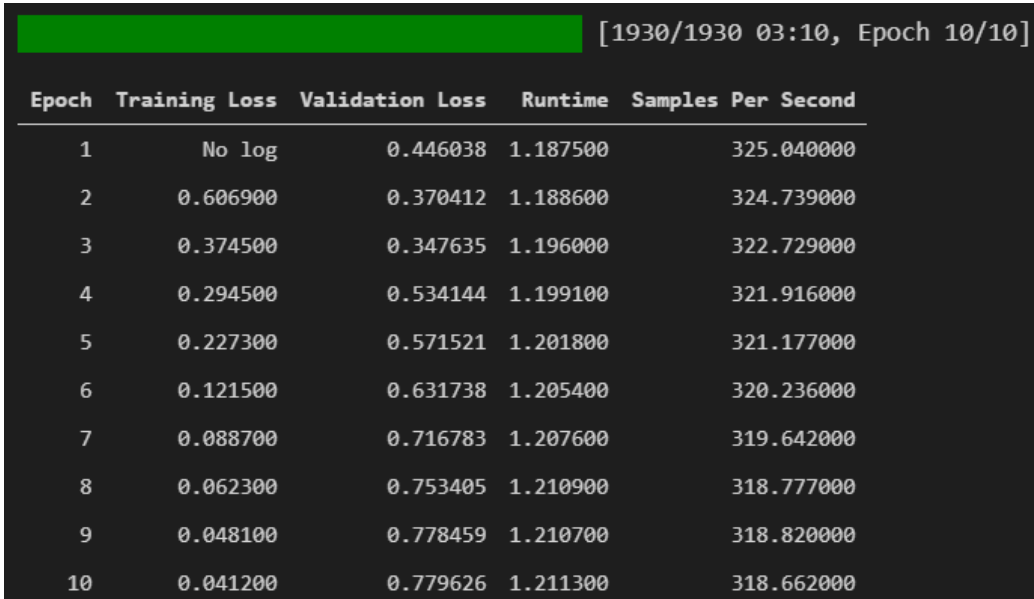|	seed_text	generated	is_anxiety	n_seq_words
0	firmly believe social anxiety sole reason deal...	enjoy water mate major gave center shower deal...	0	20
1	much ha happened last year covid ha definitely...	away feel like shallowest blandest person thin...	0	20
2	think everyone understands fear missing fomo u...	neighbor passionate therapy deeper word trigge...	1	20
3	took first dose lexapro today doctor advised s...	anxiety fear thought way something wa talking ...	1	20

Figure 3: Examples for simple language model

Part V
Fine Tuning

Classification

Filename: 6_finetune_classification.ipynb

Code

In this part we have used KFold to split our data into train and test. Afterwards, we train SentencePiece model based on the data. Lastly, we compute <UNK> on our test dataset.

Epoch	Training Loss	Validation Loss	Runtime	Samples Per Second
1	No log	0.446038	1.187500	325.040000
2	0.606900	0.370412	1.188600	324.739000
3	0.374500	0.347635	1.196000	322.729000
4	0.294500	0.534144	1.199100	321.916000
5	0.227300	0.571521	1.201800	321.177000
6	0.121500	0.631738	1.205400	320.236000
7	0.088700	0.716783	1.207600	319.642000
8	0.062300	0.753405	1.210900	318.777000
9	0.048100	0.778459	1.210700	318.820000
10	0.041200	0.779626	1.211300	318.662000

[1930/1930 03:10, Epoch 10/10]

Figure 4: Training bert classification language model

Figure 5: Evaluating bert classification language model

Results and Examples

Figure 6: Examples for bert classification model

Language Model

Filename: 7_finetune_language-model.ipynb

Code

In this part we have used KFold to split our data into train and test. Afterwards, we train SentencePiece model based on the data. Lastly, we compute <UNK> on our test dataset.

Results and Examples

	seed_text	is_anxiety	generated_depression	generated_anxiety
0	talking family member face face friend gave bad ne	0	talking family member face face friend gave bad nek think bad one see world face bad get worse every time friend find fault find family friend give wrong friend feel bad day get worse want someone think really like world like see world like see past see world like	talking family member face face friend gave bad nexie irlfax prescription day ago ago bought caffeine drink caffeine med last night thought wa really bad ever tried caffeine drink really help feel good still got depressed since first thought wa really bad wa much way
1	stuck life want succeed taking much losing sad ene	0	stuck life want succeed taking much losing sad enecember last 2nd year lost love hope friend wa im going get better i dont remember last month wa looking im thinking maybe one day would start asking thing one day try get therapy everyday one night want	stuck life want succeed taking much losing sad eneapnity got good time anxious make anxiety worse day would get bad experience bad experience even worse go see doctor day think something anxiety make mistake stop going see doctor try trying try every day try find
2	turned kind long post tl dr recently depression an	1	turned kind long post tl dr recently depression an average day last 5 minute yesterday started dating would rather never get married got really lonely got divorced 2 year ago 4 year ago 2 year ago decided to return get married 2 year ago 1 year	turned kind long post tl dr recently depression an 11 year old went doctor wanted 2 month ago 3 month ago 4 month ago also went 3 month ago back last year ha diagnosed anxiety disorder later diagnosis anxiety disorder anxiety disorder schizophrenia diagnosis
3	issue forever whenever something bad happens make	1	issue forever whenever something bad happens make much worse time wa time every time time like want want feel like need something else could live without always felt like really bad time want become useless one time think like wrong something	issue forever whenever something bad happens make feel alone feel alone never want to deal honestly feel alone feel like getting together like never felt alone feel like really hard feel like need help someone know anyone else always scared really

Figure 7: Examples for distilgpt2 language model

References

[1] https://towardsdatascience.com/
    goodbye-world-4cc844197d51

[2] https://colab.research.google.com/github/google/
    sentencepiece/blob/master/python/sentencepiece_
    python_module_example.ipynb#scrollTo=ee9W6wGnVteW

[3] https://gmihaila.github.io/tutorial_notebooks/
    gpt2_finetune_classification/

[4] https://colab.research.google.com/github/
    philschmid/fine-tune-GPT-2/blob/master/Fine_
    tune_a_non_English_GPT_2_Model_with_Huggingface.
    ipynb#scrollTo=hKBSyNLgqF9K

[5] https://github.com/huggingface/notebooks/blob/
    master/examples/language_modeling.ipynb

[6] https://www.kaggle.com/pierremegret/
    gensim-word2vec-tutorial

[7] https://machinelearningmastery.com/
    how-to-develop-a-word-level-neural-language-model-in-keras/

[8] https://huggingface.co/transformers/training.html

[9] https://www.kaggle.com/achintyatripathi/
    gensim-word2vec-usage-with-t-sne-plot

[10] https://colab.research.google.com/github/
     borisdayma/huggingtweets/blob/master/
     huggingtweets-demo.ipynb

[11] https://huggingface.co/transformers/custom_
     datasets.html

[12] https://colab.research.google.com/github/google/
sentencepiece/blob/master/python/sentencepiece_
python_module_example.ipynb#scrollTo=-k5KbVgiYae-