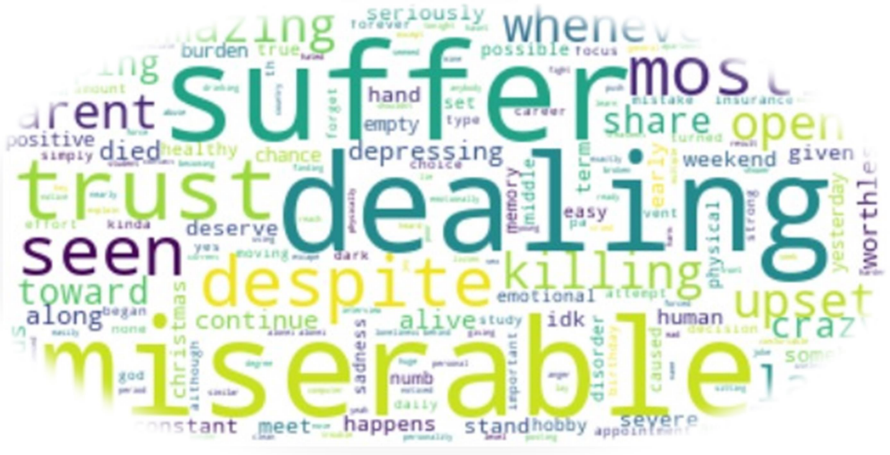
מכון טכנולוגי חולון
Holon Institute of Technology

Final project in Data Mining

Guild: Dr Jonathan Schler

Analysing Depression



Daniel Sabba , Omer Pesah

ABSTRACT

The topic of our project is depression.

Major depression constitutes a serious challenge in personal and public health, millions of people each year suffer from depression and The World Health Organization now ranks major depression as one of the most burdensome diseases in the world.

One of the ways to find those people who suffer from depression is from social networks, Tweeter, Facebook, Reddit and more, as we advance in technology, there are more and more tools from the fields of AI, NLP, and ML, many of these fields have many tools that does a great job in text analysis, and we will even use some in our project, they can predict whether an individual is depressed or not.

Searching the web for depressed people, we found communities of depressed people sharing their thoughts and emotions in forums.

In this paper, we analyze the "depressed" text; manipulating the data, extract features, categorize, and try to understand what are the attributes of "depressed" text, and how we can "predict" whether a text marked as depressed or not.

2. Experiment steps and design

2.1 The first stage of our project is to gather data

How we collected the data, and what kind of data we focused on getting.
The data we collected are posts from a forum of depression called BeyondBlue.
To understand what the attributes of depressed text are, we need to compare it with undepressed text, which we found on GitHub.

2.2 Basic feature extraction using text data

To gather insights from the cleaned data, we had to do some manipulations on the text, e.g. word count, char count, average word length, counting the stop words, the lexical diversity of the posts, the date that the post was written, etc.

2.3 Basic Text Pre-processing of text data

First, we have raw text, that is filled with tags, numerical values, punctuations, misspellings, common non-sensical txt (/n), upper letters, spam posts, links, Frequent words removal, Rare words removal, and a lot of stop words that adds noise to the context, we will need to remove them.

2.4 Advance Text Processing

We used several methods such as - Tokenization, stemming, lemmatization, n-grams, term frequency(tf), inverse document frequency(idf) and tf-idf, part of speech tagging, sentiment analysis, topic modeling, principal component analysis, and for predicting whether a text is labeled as depressed, we built a predictor using several methods e.g. support vector machine, multinomial Naïve Bayes, and KNearestNeighbors.

All those methods helped us build a model, gather more insights on our data, and adjusting it the best, so our model wont overfit the data.

3. Targets and measurements

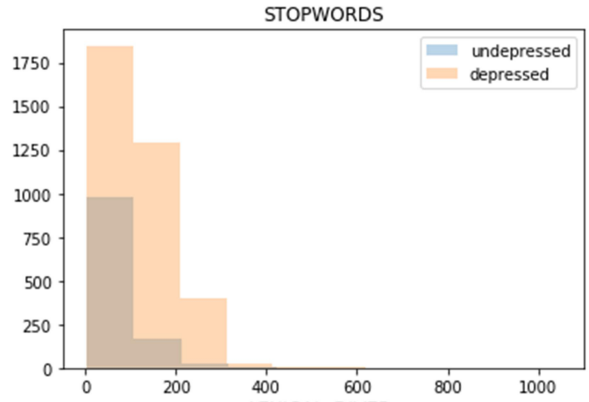
Our main target is to understand what makes depressed text marked as "depressed", what are the attributes that makes it unique, to see if there is a difference between depressed and undepressed, and if we can "predict" whether a text is tagged as depressed or not.
We will need to extract all the measurements we can from each dataset, and examine the difference and similarity between each measurement.
The measurements we look at, are detailed in sections 2.2 and 2.4.

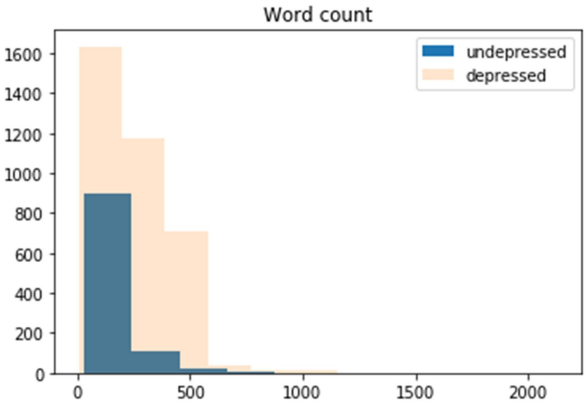
4. Experiments results and conclusions

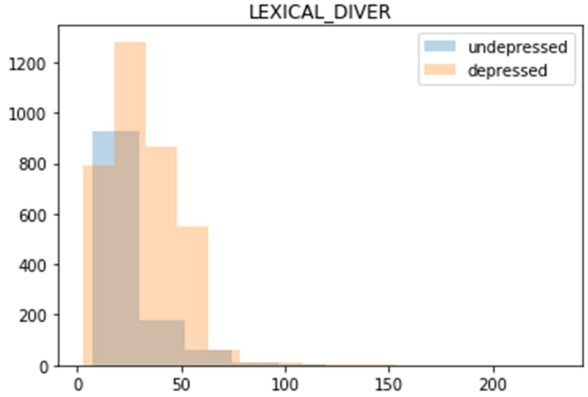
In the following section we will describe our finding and conclusions.

4.1 Basic feature extraction

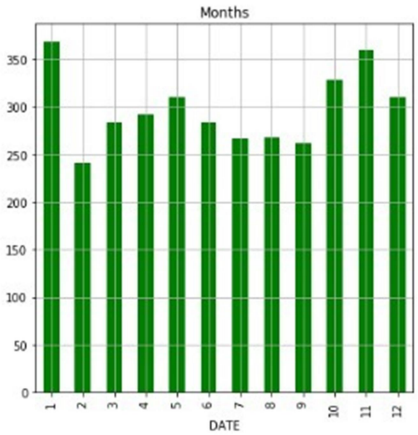
As we can see, there is a clear difference between depressed and undepressed text, looking at the lexical diversity, we can see that depressed people have larger vocabulary, additionally, they use more words, hence, got more stop words.

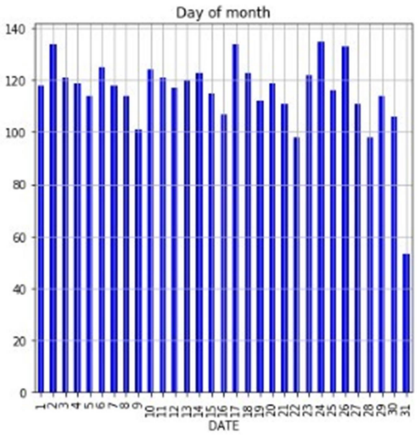


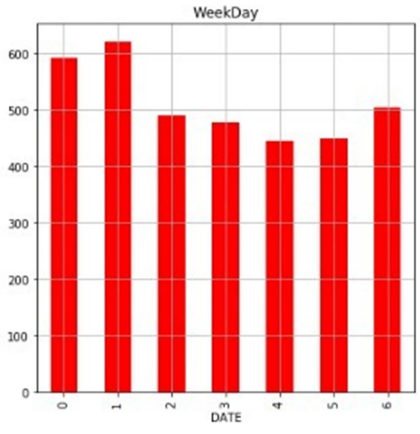


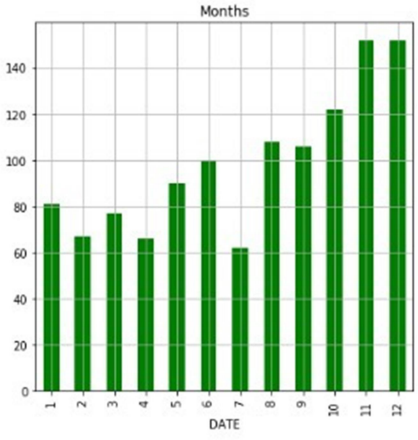


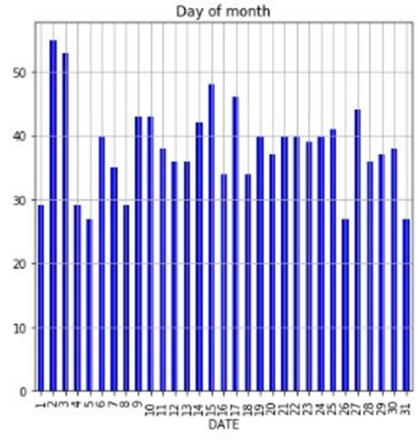
4.2 Analyzing the date

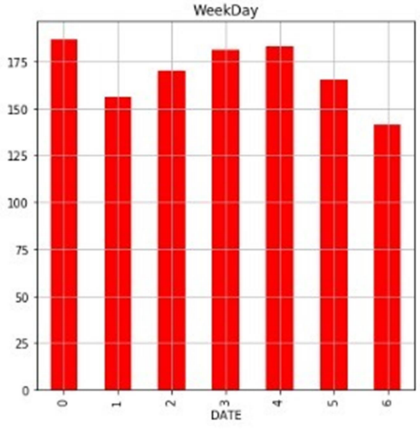












Top row- depressed posts, bottom row- neutral posts.
In the graphs above, we noticed some interesting info about the time the posts were written, looking at the month, we noticed that depressed people write more during the end of the year and the first month, and during the year have "ups and downs", while undepressed writers activity is weak during the entire year, and gets stronger in the end of the year.

By looking at the weekday, where Sunday is 6, depressed people are more active in the start of the week (sun, mon, tue), while in undepressed posts, we see that the activity is high at the end of the week, and Sunday.

4.3 Bi-grams

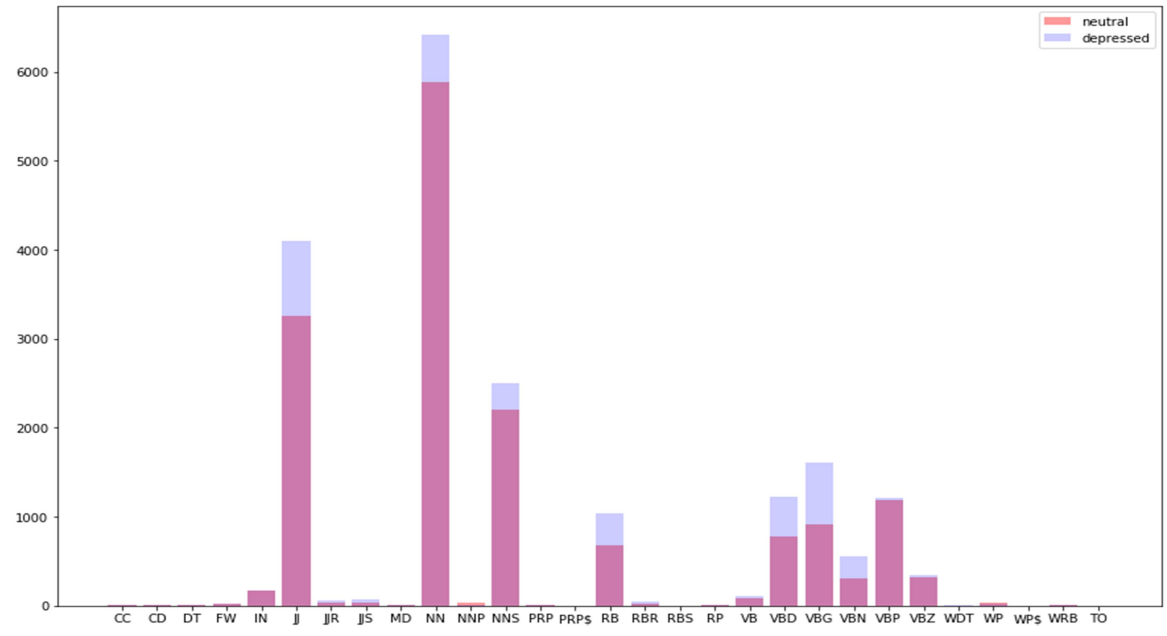
By looking at Bi-grams, we could "sense" the text and get the main idea, we had to look manually for bi-grams that contribute our understanding, and the strong bi-grams were not very helpful.

Adding bi-grams to our 3d projection, and topic modeling, we noticed that it groups the posts (after dimensional reduction), and made the clustering worse, we decided not to use this feature lda and pca.

403	suicide,or
460	hard,to
379	to,feel
327	my,depression
409	the,past

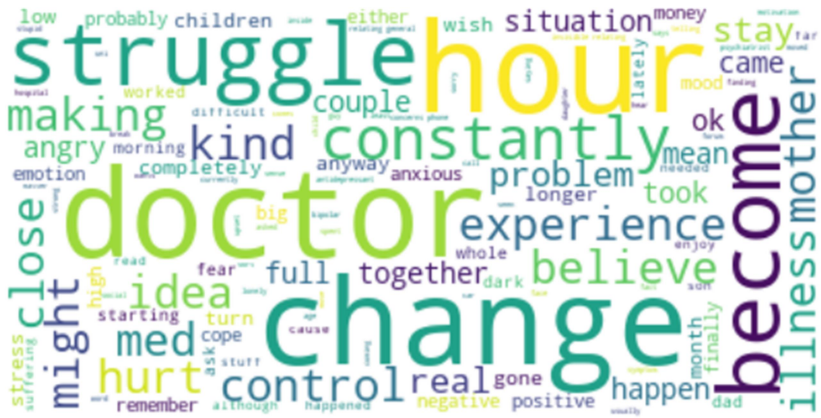
5514	I,have
4374	I,am
3101	and,I
2923	I,feel
2606	I,was
2439	I,don't
2115	to,be

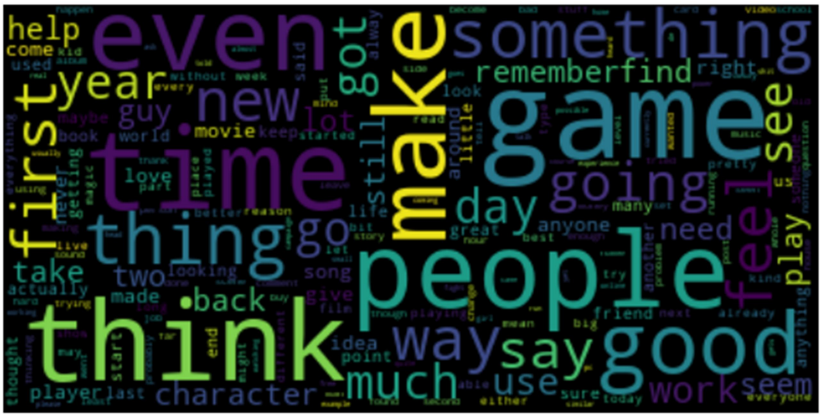
4.4 Part of speech tagging



POS tagging is a measurement that helped us understand the morphologic and syntactic properties.
In the graph above, we see the use of part of speech within the depressed and undepressed posts, writers use a lot of NN (nouns) and JJ (adjectives), as well as VB (verbs), and we couldn't find a distinct difference between the texts.

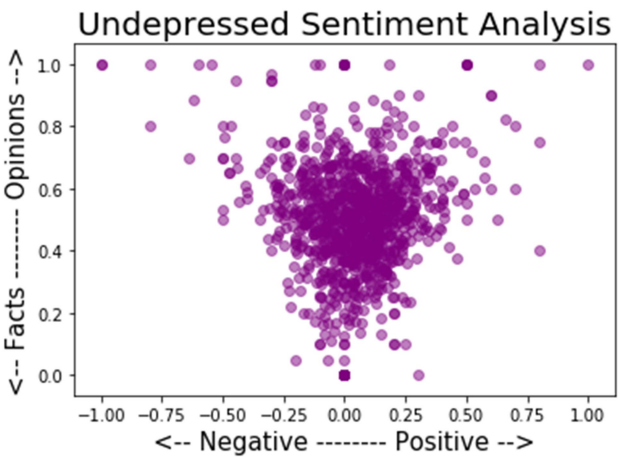
4.5 Word cloud

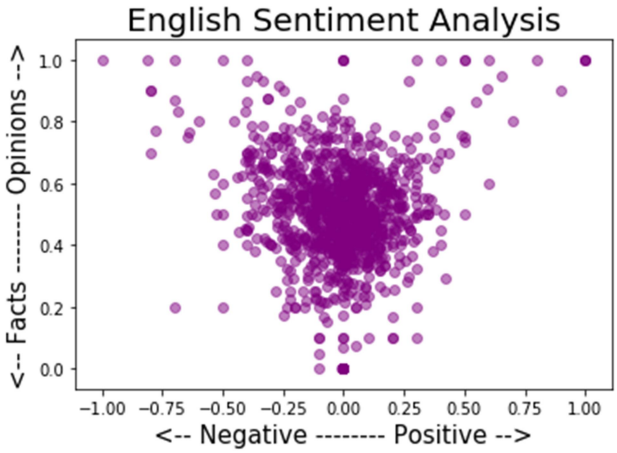




In this section we took the most common words and we tried to understand the general subject.
We can see that there are different subjects, but the general topic is the struggle dealing with depression, while the undepressed cloud (the black one) doesn’t have a specific subject.

4.6 Sentiment Analysis





The scatter plots above, shows us the polarity and subjectivity of the depressed and undepressed posts, we see that the English (depressed) posts are tend to be more negative then undepressed, and less positive. here are the differences –

Depressed – Polarity (over 0.25) = 71 (positive)

Polarity (under -0.25) = 115 (negative)

Subjectivity (over 0.6) = 279 (opinion)

Subjectivity (under 0.4) = 255 (fact)

Undepressed – Polarity (over 0.25) = 110 (positive)

Polarity (under -0.25) = 56 (negative)

Subjectivity (over 0.6) = 274 (opinion)

Subjectivity (under 0.4) = 256 (fact)

5. Topic Modeling and Principal Component Analysis

In order to cluster the data by topics, we used several methods of feature extraction (count vector, tf-idf vector), decomposition (Latent Dirichlet Allocation), and dimension reduction (PCA).

Figure 2 Data projection using PCA

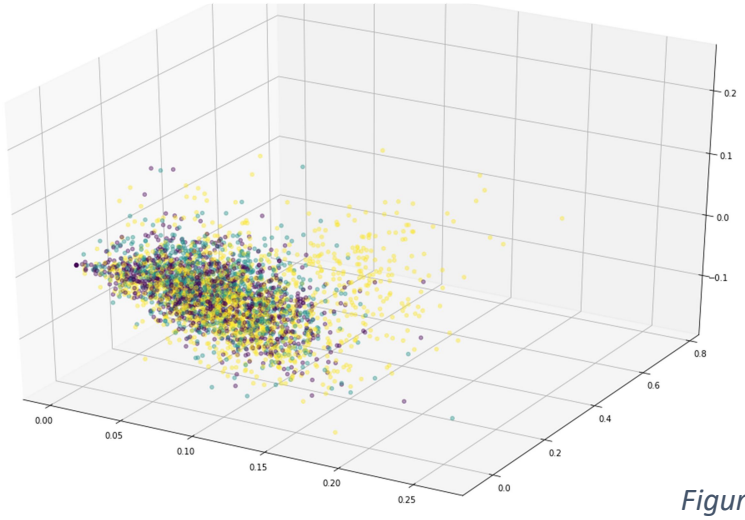


Figure 1 Devision of topics

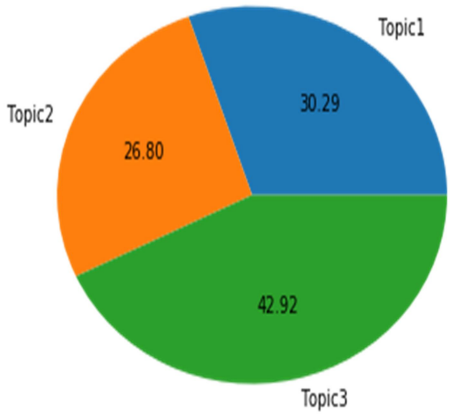


Figure 3 How the model selected the topics

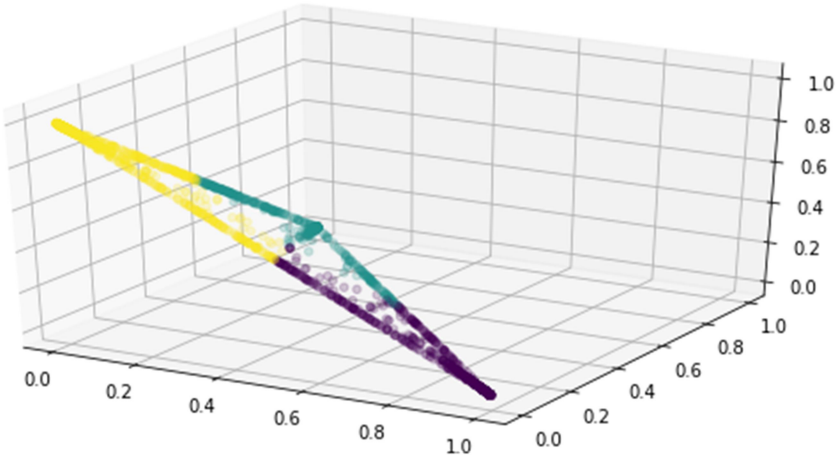


Table 1 Words from each topic

Topic1	Topic2	Topic3
usually	exercise	invisible
stuck	question	concerns
decided	major	encouraged
emotions	lack	moderators
matter	constant	relating
wondering	heart	offline
coming	body	clinicallytrained
upset	whether	black
worry	emotional	stupid
half	asked	dog

In figure 1 we see the distribution of the topics, which is pretty fair.
In figure 2 we see that there is a proximity inside the topics themselves.
In figure 3 we see how the model classified the posts, it looks like a triangle that each corner divided to a topic, the distribution seems pretty fair.
In figure 4 we see 10 words from each topic, seems like topic 1 is about feelings and emotions, topic 2 is about consulting and getting help, and topic 3 is more of a general topic relating to thoughts.

In conclusion, after examining all these features, we wanted to see if we can predict whether a post marked as depressed or not, we built a predictor using cross validate with several models (svc, Multinomial Naïve Bayes, KnearestNeighbors), we found that the results were great, we got the highest score with a 99% success rate, from this paper we see, that fighting with depression can be much easier for authorities if they will work with tools such as data mining and machine learning, we think the next stage to improve our paper is to examine real life habits, add more data, and try different language.