Predicting Depression via Social Media

Depression is a common illness worldwide, with more than 300 million people affected. Depression is different from usual mood fluctuations and short-lived emotional responses to challenges in everyday life. Especially when long-lasting and with moderate or severe intensity, depression may become a serious health condition. It can cause the affected person to suffer greatly and function poorly at work, at school and in the family [1].

Depression affects an estimated one in 15 adults (6.7%) in any given year. And one in six people (16.6%) will experience depression at some time in their life. Depression can strike at any time, but on average, first appears during the late teens to mid-20s. Women are more likely than men to experience depression. Some studies show that one-third of women will experience a major depressive episode in their lifetime [2].

At its worst, depression can lead to suicide. Close to 800 000 people die due to suicide every year. Suicide is the second leading cause of death in 15-29-year-olds [1].

Risk Factors for Depression

Depression can affect anyone—even a person who appears to live in relatively ideal circumstances.

Several factors can play a role in depression [1]:

- **Biochemistry:** Differences in certain chemicals in the brain may contribute to symptoms of depression.
- **Genetics:** Depression can run in families. For example, if one identical twin has depression, the other has a 70 percent chance of having the illness sometime in life.
- **Personality:** People with low self-esteem, who are easily overwhelmed by stress, or who are generally pessimistic appear to be more likely to experience depression.
- **Environmental factors:** Continuous exposure to violence, neglect, abuse or poverty may make some people more vulnerable to depression.

The National Ambulatory Medical Care Survey (NAMCS) found that the number of people diagnosed with depression has increased by 450% since 1987 [3, 4, 5].For every person who took an anti-depressant in 1987, there are now more than five.

But that increase is deceptive. Yes, there are more people getting diagnosed with depression, but there are three explanations for that.

1) Depression has become more common.

2) Anti-depressants have gotten better. Anti-depressants sometimes cause fatigue, heart arrhythmias or cognitive impairment, but these side-effects are rare. In the 1980s they were common. [6]

3) It is no longer taboo. According to the NAMCS survey, less than 20% of people with depression in 1987 sought treatment.

Treatment of Major Depressive Episode Among Adults

Figure 1 shows data on treatment received within the past year by U.S. adults aged 18 or older with major depressive episode. Treatment types include health professional only, medication only, and health professional and medication combined [7].

- An estimated 44% received combined care by a health professional and medication treatment.
- Treatment with medication alone was least common (6%).
- Approximately 37% of adults with major depressive episode did not receive treatment.

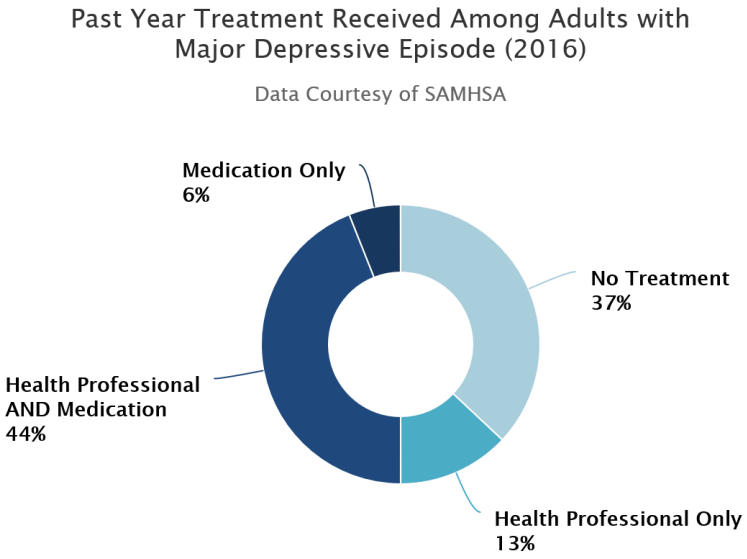


Figure 1

Treatment of Major Depressive Episode Among Adolescents

Figure 2 shows data on treatment received within the past year by U.S. adolescents aged 12-17 with major depressive episode in 2016. Treatment types included health professional only, medication only, and combined health professional and medication [7].

- An estimated 19% received care by a health professional alone, and another 19% received combined care by a health professional and medication treatment.
- Treatment with medication alone was least common (2%).
- Approximately 60% of adolescents with major depressive episode did not receive treatment.

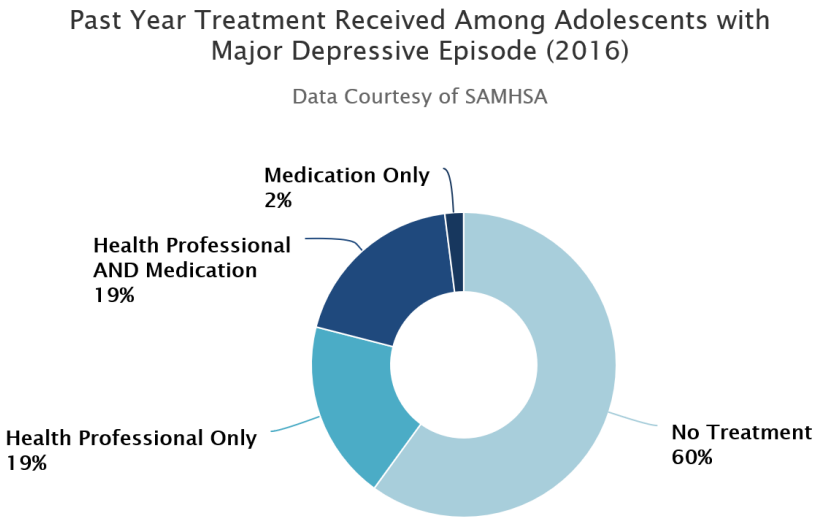


Figure 2

Therefore, preventing mental health problems can be helpful to improve human well-being. Early detection is a basis of the prevention of mental health problems [8]. Even if more and more people seek treatment, there are still several reasons such as lacking mental health knowledge, and stigmatizing attitudes towards mental patients, people with mental health problems are not motivated to seek professional help. More importantly, traditional methods for detecting individual mental health problems (e.g. self-report techniques and clinical diagnosis) cannot identify individual mental health status in real-time, which may lead to delayed reporting and can have negative impacts on personal mental health.

The development of Internet and information technology gives us an opportunity to find new method for detecting mental health problems. First, based on information technology, Internet behaviors can be collected and processed in a non-intrusive, accurate and efficient manner. Given that the relationship between Internet behaviors and psychological features (e.g. personality) has been confirmed in previous studies, which implies the possibility of detecting mental health problems through Internet behaviors analysis. Amichai-Hamburger and Ben-Artzi (2000) found that there exists relationship between Internet behaviors and personality. Gosling et al. (2011) collected digital records of human behaviors on social media and proved the accuracy of predicting personality by perceiving Internet behaviors. Furthermore, Kosinski and colleagues (2013) established computational models for predicting Facebook user's psychological profile and personal preference [8]. Wu et al. (2015) argued that computer-based personality judgments are more accurate than those made by humans.

In view of these advantages, some studies have been conducted to identify individual mental health problems based on Internet behaviors. Park et al. (2012) found that it is possible to identify one user's depressive emotion by analyzing his/her posts [8].

Our goal is to recognize depressed users from their web-behaviors. If this goal can be realized, the depressed persons can get effective treatments earlier. To reach this goal, we will build up a model by the aid of the **Neural Networks** to predict whether an user is depressed and the depressed ones' extent of depression. The tool to develop this method is **Python** and especially **Scikit-learn** Software, for the

machine learning methods, and the **Natural Language Toolkit** in order to process the xml files. We are going to perform a **classification** task, and the measure for evaluating our methods is the **Accuracy** of prediction and the **F measure**. Apart from the text analysis, some other data that we are going to take into account are:

1. The total posts of each user
2. The time of the posts, by classifying them into four categories based on the post's time, 0.00-18.00, 18.00-00.00, 00.00-8.00.

About the Sample

The challenge consists of sequentially processing pieces of evidence and detect early traces of depression as soon as possible. The task is mainly concerned about evaluating Text Mining solutions and, thus, it concentrates on texts written in Social Media. Texts should be processed in the order they were created. In this way, systems that effectively perform this task could be applied to sequentially monitor user interactions in blogs, social networks, or other types of online media. The test collection for this pilot task is the collection described in [Losada & Crestani 2016]. It is a collection of writings (posts or comments) from a set of Social Media users. There are two categories of users, depressed and non-depressed, and, for each user, the collection contains a sequence of writings (in chronological order). For each user, his collection of writings has been divided into 10 chunks. The first chunk contains the oldest 10% of the messages, the second chunk contains the second oldest 10%, and so forth.

Table 1 reports the main statistics of the train and test collections. Both collections are unbalanced (more non-depression cases than depression cases). The number of subjects is not very high, but each subject has a long history of writings (on average, we have hundreds of messages from each subject). Furthermore, the mean range of dates from the first to the last submission is quite wide (more than 500 days). Such wide chronology permits to study the evolution of the language from the oldest piece of evidence to the most recent one.

	Train		Test	
	Depressed	*Control*	*Depressed*	*Control*
Num. subjects	83	403	52	349
Num. submissions (posts & comments)	30,851	264,172	18,706	217,665
Avg num. of submissions per subject	371.7	655.5	359.7	623.7
Avg num. of days from first to last submission	572.7	626.6	608.31	623.2
Avg num. words per submission	27.6	21.3	26.9	22.5

Table 1

Except from the data presented above the folder also contains the training data:

- risk_golden_truth.txt: this file contains the ground truth (one line per subject). The code 1 means that the subject is a risk case of depression, while 0 means that the subject is a non-risk case.

- positive_examples_anonymous_chunks: this folder, which stores all the posts of the risk cases, contains 10 subfolders. Each subfolder corresponds with one chunk. Chunk 1 contains the oldest writings of all

users (first 10% of submitted posts or comments), chunk 2 contains the second oldest writings, and so forth. The name of the files follows de convention: <subjectname>_<chunknumber>.xml.

- negative_examples_anonymous_chunks: this folder, which stores all the posts of the non-risk cases, contains 10 subfolders. Each subfolder corresponds with one chunk. Chunk 1 contains the oldest writings of all users (first 10% of submitted posts or comments), chunk 2 contains the second oldest writings, and so forth. The name of the files follows de convention: <subjectname>_<chunknumber>.xml.

- writings-per-subject-all-train: This txt contains the number of posts each users has made.

- scripts evaluation

We are going to apply the methods mentioned above to both testing and training sample.

Analyzing the subject's posts

For every file we saved the user's ID and we analyzed each xml file with the following steps:

a. Tokenization
b. Removing the punctuation
c. Stemming
d. Remove stopwords
e. Bag of words for every post
f. Sum the BOWs of every post
g. Save the ID and the BagOfWords in a row of a file with total results.

The steps above have been followed for every xml file and then we concentrated all the data. We searched from the IDs the data that belongs to the same users and we created the user's total Bag Of Words. In addition to this, for every post we saved its time and we classified them to 3 categories (night 0.00-8.00 , working hours 8.00-16.00 , evening 16.00-00.00).

The next step is to create the collection's dictionary from the training data and then we found the term frequency for each user. In order to decrease the length of the dictionary we removed the words that appeared only once and twice to the training dataset. The dictionary contains 28914 words. So the first feature we will use is the **term frequency for each subject**, which gives us a table with the following dimensions for the training and for the testing dataset **486*28914** and **401*28914**. This is the **first dataset** that we will use to the Neural Network.

For the second dataset we added some features which are the following:

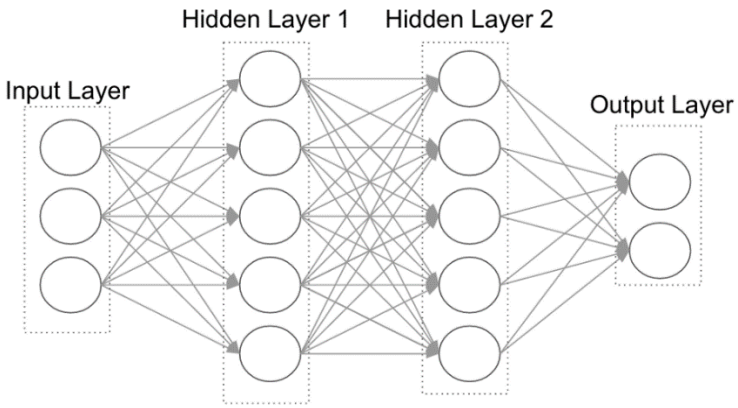
- The **total posts per user** divided by the average posts per user from the training data.
- The percentage of the posts that are made at night, at the working hours and at the evening for each user.

- Finally the data about whether the user is diagnosed with depression or not in the reality. This information is the output of our Neural Network.

  We added the data we mentioned above to the table with the **term frequency for each subject**, and the table has the following dimensions for the training and for the testing dataset **486*28918** and **401*28918**.

Neural Network

We developed a feedforward Neural Network with 2 hidden layers, with 15 neurons the first and 11 the second one. A **feedforward neural network** is an artificial neural network wherein connections between the nodes do *not* form a cycle. The feedforward neural network was the first and simplest type of artificial neural network devised. In this network, the information moves in only one direction, forward, from the input nodes, through the hidden nodes and to the output nodes. There are no cycles or loops in the network. We used the **Scikit-Learn** library to implement the NN.



Feedforward neural network with 2 hidden layers

Evaluating Measures

True Positives (TP) - These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes.

True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. False positives and false negatives, these values occur when your actual class contradicts with the predicted class.

False Positives (FP) – When actual class is no and predicted class is yes.

False Negatives (FN) – When actual class is yes but predicted class in no.

Accuracy - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$Precision = \frac{tp}{tp + fp}$$

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

$$Recall = \frac{tp}{tp + fn}$$

F1 score - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. The F1 score can also be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Micro avg: Calculate metrics globally by counting the total true positives, false negatives and false positives.
Macro avg: Calculate metrics for each label, and find their unweighted mean. This does not take label imbalance into account.
Weighted avg: Calculate metrics for each label, and find their average weighted by support (the number of true instances for each label). This alters 'macro avg' to account for label imbalance; it can result in an F-score that is not between precision and recall.

Dataset with features from NLP

As we mentioned previously, the inputs of the first NN, are the **term frequency** for each user and the output is the depression information for each subject. For the training data the input is a 486*28914 table and the output is a 486*1 table. For the testing data the input is a 401*28914 table and the output is a 401*1 table

Accuracy = 0.8853

	Predicted 0	Predicted 1
Actual 0	342 tn	7 fp
Actual 1	39 fn	13 tp

0 = non-depressed

1 = depressed

Classification Report

	Precision	Recall	F1 Score	Support
0	0.90	0.98	0.94	349
1	0.65	0.25	0.36	52
micro avg	0.89	0.89	0.89	401
macro avg	0.77	0.61	0.65	401
weighted avg	0.87	0.89	0.86	401

From the results above, we can assume that there is an imbalance to the classes. The micro average is biased to the bigger classes, as a result, its values are bigger than the other average values. The macro average is the mean value between the two classes and the weighted average takes into account the size of the class, so as a result the macro average is the smaller one and the weighted average is close to the micro average.

The precision, recall and F1 Score values from the class of the non-depressed are really positive; and that's expected because in this class belongs the 87% of the total testing data. These values, for the class of depressed, are not as high as the previous class, but the average values are high enough to say that our model works well.

Even if the accuracy score is 88.53%, the other measures are more important than the accuracy because our sample is an imbalanced one.

Dataset with features from NLP and feature related to time and total posts

As we mentioned previously, the inputs of the second NN, are **the term frequency** for each user, **the time of posting** and **the total posts** per subject and the output is the depression information for each subject. For the training data the input is a 486*28918 table and the output is a 486*1 table. For the testing data the input is a 401*28918 table and the output is a 401*1 table

Accuracy = 0.9152

	Predicted 0	Predicted 1
Actual 0	340	9
Actual 1	25	27

Classification Report

	Precision	Recall	F1 Score	Support
0	0.93	0.97	0.95	349
1	0.75	0.52	0.61	52
micro avg	0. 92	0.92	0.92	401
macro avg	0.84	0.75	0.78	401
weighted avg	0.91	0.92	0.91	401

From the results above, we make the same observations with the dataset with the less features. The accuracy score is 91.52%.

The dataset with the extra features gives us better results to all the evaluation measures. Not only the accuracy is better, but also the total fmeasure, precision and recall. The fmeasure and the recall of the class of depressed, has doubled its performance. All the average values, micro, macro and weighted have better results than the dataset with the NLP data.

In conclusion, we can assume that the time of posting and the total posts of each user are important features for predicting weather a person is depressed or not.

Comparing our results with the results of e-risk 2017

In this section we are going to compare our results with the results of the the competition e-risk 2017, from which we got our data. Two of the 8 teams that participated, also used some type of Neural Networks. The Group from the University of Applied Sciences and Arts Dortmund (FHDO) submitted results obtained from five different models. These models employ linguistic meta information extracted from the users' texts. This team considered classifiers based on Bag of Words, Paragraph Vector, Latent Semantic Analysis (LSA), and Recurrent Neural Networks using Long Short Term Memory (LSTM). This team has considered a wide range of features: readability features, LIWC features (e.g., statistics on use of pronouns or verb tense), hand-crafted features (for example, specific terms related to antidepressants or diagnosis), and has put in practice sophisticated approaches based on LSTM, neural networks, LSA, and vectorial representation of texts [9].

The team from the School of Information of the University of Arizona, leveraged external information beyond the available training set. This included a preexisting depression lexicon and concepts extracted from the Unified Medical Language System (UMLS). With these features, they employed sequential–recurrent neural networks– and non-sequential –support vector machines– models for prediction. They also used ensemble methods to leverage the best of each model [9].

As we can see, by comparing our results with the results of the contestants,

- Our **F measure** (macro average which is our lowest) is 0.78 and their best is 0.64 (Institution FH Dortmund, Germany)
- Our **Precision** is 0.84 and their best is 0.69 (Institution FH Dortmund, Germany)

- Our **Recall** is 0.75 and their best is 0.92 (Institution U. Arizona, USA)

The best results of these measures doesn't belong to the same team or method, so our results are generally better than those of the contestants.

Institution Files	F1	P	R
GPLA	0.35	0.22	0.75
GPLB	0.3	0.18	0.83
GPLC	0.46	0.42	0.5
GPLD	0.47	0.39	0.6
FHDOA	**0.64**	0.61	0.67
FHDOB	0.55	**0.69**	0.46
FHDOC	0.56	0.57	0.56
FHDOD	0.57	0.63	0.52
FHDOE	0.6	0.51	0.73
UArizonaA	0.4	0.31	0.58
UArizonaB	0.3	0.33	0.27
UArizonaC	0.34	0.21	**0.92**
UArizonaD	0.45	0.32	0.79
UArizonaE	0.45	0.34	0.63
LyRA	0.14	0.11	0.19
LyRB	0.16	0.11	0.29
LyRC	0.16	0.12	0.25
LyRD	0.15	0.13	0.17
LyRE	0.08	0.11	0.06
UNSLA	0.59	0.48	0.79
UQAMA	0.53	0.48	0.6
UQAMB	0.48	0.49	0.46
UQAMC	0.42	0.5	0.37
UQAMD	0.38	0.64	0.27
UQAME	0.39	0.45	0.35
CHEPEA	0.48	0.38	0.65
CHEPEB	0.47	0.37	0.63
CHEPEC	0.46	0.37	0.63
CHEPED	0.45	0.36	0.62
NLPISA	0.15	0.12	0.21

References

[1] http://www.who.int/news-room/fact-sheets/detail/depression

[2] https://www.psychiatry.org/patients-families/depression/what-is-depression

[3] National Trends in the Outpatient Treatment of Depression, Mark Olfson, MD, MPH, Steven C. Marcus, PhD, Benjamin Druss, MD, Lynn Elinson, PhD, Terri Tanielian, MA, Harold Alan Pincus, MD ,2002

[4] National Trends in the Treatment for Depression From 1998 to 2007, Steven C. Marcus, PhD; Mark Olfson, MD

[5] http://www.multpl.com/united-states-population/table

[6] http://slatestarcodex.com/2014/07/07/ssris-much-more-than-you-wanted-to-know/

[7] https://www.nimh.nih.gov/health/statistics/major-depression.shtml

[8] Predicting Depression from Internet Behaviors by Time-Frequency Features, Changye Zhu ,Baobin Li , Ang Li , Tingshao Zhu ,2016

[9] CLEF 2017 eRisk Overview: Early Risk Prediction on the Internet: Experimental Foundations, David E. Losada, Fabio Crestani, and Javier,2017