Don't worry, I am not depressed... Or am I?

Luka Pavlović, Martin Sršen, Krunoslav Jurčić

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{luka.pavlovic, martin.srsen, krunoslav.jurcic}@fer.hr

Abstract

Language can provide powerful insight into personality, social or emotional status and also mental health. In this paper we researched which words contribute the most to classifying someone as depressive or someone as not depressive. On top of that, we tried adding several features that were extracted from the dataset to see if they might help make a better decision. Collection of data that we used is gathered from Reddit, a social media platform. It is contained from posts and comments from control group and from group of people diagnosed with depression. For our model we chose logistic regression with term frequency-inverse document frequency(TF-IDF) because it showed to have the best performance. One of the features that provided a slight performance boost is average time between two sequential posts or comments. Our model showed better F1 performance than other models that used the same dataset.

1. Introduction

Depression is a common mental disorder whose importance is often being overlooked. If not treated properly, it can lead to more serious problems, disability, psychotic episodes and unfortunately in some cases even to suicide. According to World Health Organization, more than 264 million people worldwide suffer from depression. As it is the case with the majority of mental illnesses, early detection of the problem can prove to be very useful in its prevention. With the development of technology, people are spending more and more time on the Internet, and their activity on various websites (e.g. social networks, blogs...) can tell a lot about them. Language is a powerful indicator of personality, social or emotional status, but also mental health. It will surprise no one to learn that those with symptoms of depression use an excessive amount of words conveying negative emotions, specifically negative adjectives and adverbs such as "lonely"or "sad". More interesting is the use of pronouns like "me", "myself" and "I" which are used much more by those with symptoms of depression which means they are generally more focused on themselves. Someone's social media account content can often provide us with valuable information about that person's emotional status. However, the current technology used to deal with depression issues is only reactive. For instance, some specific types of risks can be detected by tracking Internet users, but alerts are triggered when the victim makes his disorders explicit, or when the criminal or offending activities are actually happening. We believe there is a better way in detecting early detection by using natural language processing tools which can be used on everyday social network posts. The main idea of this paper is to present one method of predicting whether a person is having depressive thoughts and intentions using natural language processing. We tried to achieve this by extracting words that are most commonly used by Reddit users who are suspected to have depressive tendencies alongside other features that proved to be group specific. The dataset used in this paper was first presented by Losada and Crestani in 2016. The dataset, which is more detailed described in Section 3, consists of numerous Reddit posts and comments made by various users. In order to achieve that, we used a number of different approaches combined with different machine learning models, who are more accurately described in Section 4.

In this paper we make four main contributions:

1. we explore the influence of certain words or word phrases in detecting depression
2. we explore the influence of post/comment time frequency as a feature
3. we explore the influence of post/comment length as a feature
4. we explore the influence of sentiment in users posts as a feature

2. Related works

Even though depression is a well-known mental health issue, not a lot of datasets and papers issuing this distinctive natural language processing problem are present nowadays. This can be explained as a lack of general interest in the problem. Another reason is the fact that classification of such specific behavioural patterns can be quite challenging given the fact that the matter is rather subjective. Earlier mentioned paper by Losada and Crestani (2016) provided great insight regarding this topic. Their dataset is actually the first dataset for research on depression and language use. The details of this dataset, which was also used in this paper, are described in Section 3. For the evaluation of their model,a new evaluation metric was proposed by Losada and Crestani, called early risk detection error (ERDE) measure. This measure is based on the accuracy of the decisions and the delay in detecting positive cases.

Interesting paper which implemented sentiment analysis in order to assess whether the user has depression or not was done by Wang et al., where sentiment analysis method is proposed utilizing vocabulary and man-made rules to calculate the depression inclination of each micro-blog. Secondly, a depression detection model is constructed based on the proposed method and 10 features of depressed users derived from psychological research. We can see many researchers trying to develop models off of ever-growing social networks where a lot of text information is being placed. Using NLP with that kind of advantage to tackle growing depression problem might be sufficient in years to come to explore depression further and provide psychiatrists with new and insightful data.

Another great approach was proposed by Benton et al. (2017) about estimating suicide risk and mental health in a deep learning framework additionally included the effect of modelling gender, which has been shown to improve accuracy in tasks using social media text. The authors of this paper developed neural multi-task learning (MTL) models for 10 prediction tasks (suicide, seven mental health conditions including depression, neurotypicality, and gender). The results of their model showed that choosing the right set of auxiliary tasks for a given mental condition can yield a significantly better model, and that The MTL model dramatically improves for conditions with the smallest amount of data. Most importantly for our work, the results also showed that gender prediction does not follow the two previous points, but improves performance as an auxiliary task.

3. Dataset

The dataset used in this paper contains Reddit posts and comments from 892 different users. 137 users have been diagnosed with depression, and the rest are a control group. The maximum number of data for each user is limited to 1000 posts and 1000 comments which is Reddit API limit, and both the posts and the comments are sorted by chronological order, which is important given the task of early depression detection. The collection was created as a sequence of XML files, one file per user. Users are labeled as depressive only if they explicitly stated that they were diagnosed with depression and undepressed users are also from depression related subreddits but are not depressed, i.e. someone around them is suffering from depression so they want to explore it . Each submission is represented with:

- ID NUMBER
- TITLE
- DATE
- TEXT

The train-test split of the dataset consists of 486 train subjects, 83 of which are positive, and 406 test subjects , out of which 54 are positive subjects.

3.1. Preprocessing

Cleaning data before using it to generate features for our model is very important, especially for Reddit data where we have several separated important groups of data. We first concatenated last N titles and texts for each user, where N determines how many posts our model requires to work correctly. Another type of data we extracted was time between each two of the sequential posts or comments for a certain user. After we grouped our data we then removed links from the text, lowercase everything and created “bag of words” representations of our sequences. Finally we applied WordNetLemmatizer on each word to create each word lemma.

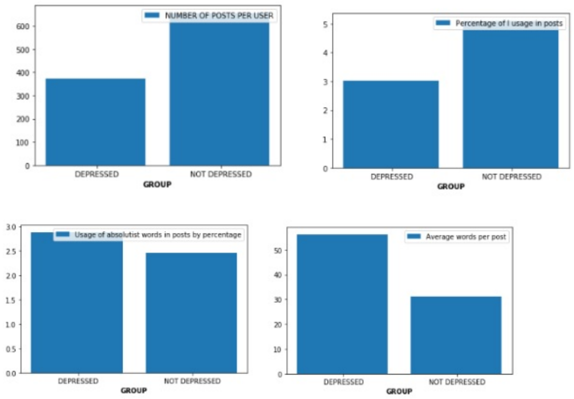

Figure 1: Data analysis on possibly interesting features mentioned in psychiatric literature

4. Problem

Studying depression related literature we found few claims that are supposed to characterise people suffering from depression and their interaction on social media. Claims we investigated are usage of ’I ..’, absolutist words, negative sentiment in tweets, number of posts per user and time between posts. There is also one important observation considering depressed people on social media and that is time of posting, but those claims were already investigated in a paper by Banović, Fatorić and Rakovac from 2019, and proved not important enough, at least not on this dataset. Here you can see different characteristics and their behaviour in our data.

Analysis of our data showed that some assumptions really do show a solid difference between depressed and not depressed groups. Absolutist words (absolutely, totally, whole, . . .) have no big difference between the groups, while average words per post and number of posts per user show significant difference.

4.1. Model

We implemented several models that take into account the factors we mentioned earlier. Firstly we vectorize words using term frequency-inverse document frequency(TF-IDF), which is basically a statistic that determines how important a certain word is to a document in some collection or corpus. The TF-IDF value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general. For the task of classification we used a Logistic Regression classifier which proved to have excellent results on various text classification tasks. Those models were evaluated against baseline models. As baseline we used two different naive approaches:

- Random guesser : simple model which guesses whether a user has depression or not randomly. Each guess has the same probability of being chosen
- Stratified random guesser : another simple model with one small improvement; guess is not entirely random.

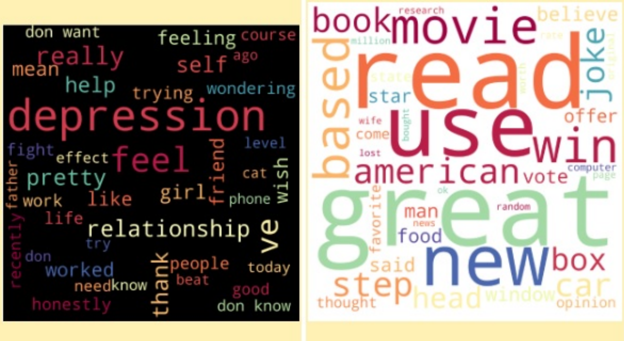

Figure 2: Unigrams that contribute most to depression classification (black background) and unigrams that contribute most to non depression classification (white background)

> It uses a percentage of depressed people in the train set and with that chance guesses whether the user will be depressed

Assumptions that proved insightful for our data and the features that we used to try improving model performance were: number of posts, sentiment of tweets, average time passed between posts and average words per post. Models were optimized following standard practices. We used grid search to fine tune models and TF-IDF vectorizer parameters. Parameters that showed to impact final results the most are the fact that we used both unigram and bigram words and that we ignored terms that have a document frequency strictly lower than the given threshold , the value which is in literature often called cut-off.

5. Results

Using this kind of model shows significant improvement over the F1 scores on the first 10, 100 and 500 posts over the model presented by Losada et al. This model also outperforms F1 scores across all identical data subsets by Banović, Fatorić and Rakovac. Although data analysis showed promising insights, the only feature that slightly improved base LR + TF-IDF model performance was average time passed between posts, while other additional features only created unnecessary noise that decreased models accuracy on the test set.

Given the fact that our model outperforms both mentioned above, it's important to question what can we learn from it, and can we spot any pattern in detecting depression. Using the Logistic Regression model gives us the ability to extract words that have the most impact in determining which group a certain user belongs to. We've made word cloud out of those words to help visualize what model recognized as words that contribute classification the most. Larger the word, larger the contribution.

Solely on this unigram word clouds we can see that there are words that we would intuitively assign to depressed people so it's interesting to see how the model sorts words for both classes. We can also look up bigram words to see how well they reflect our intuition. We can confirm from this data that those with some kind of depression are much more focused on themselves than on others.

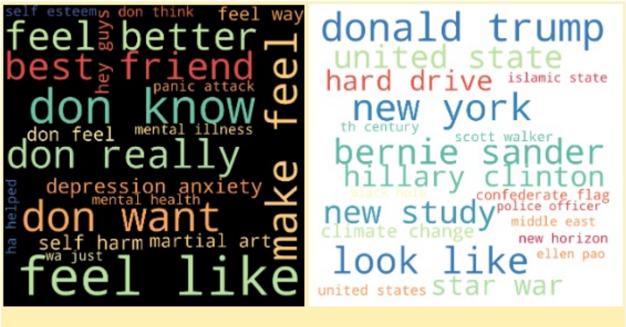

Figure 3: Bigrams that contribute most to depression classification (black background) and bigrams that contribute most to non depression classification (white background)

6. Future Work

The biggest problem while creating this model was the size and content of the dataset. Increasing the dataset is a costly process but doing so would improve model performance. Time of posts, that was the only thing added alongside users' posts, helped us create a better model so it's safe to say that additional details on users could also benefit researchers. Work presented by Benton et al. in modelling gender as an additional feature could also prove useful. We might try using word embeddings averaged across all words as input to the logistic regression model or using it directly with Long Short Term Memory Networks which is the main ingredient of most state of art models since it can make use of long sequential input data and its ordering by avoiding gradient vanishing and making use of recurrent connections.

7. Conclusion

Depression is mental disorder that, if not treated properly and early enough, could lead to more serious problems or even suicide. This kind of study could prove to be useful for psychiatrists, since a full history of posts can contain valuable insights to help better assess the patients. We experimented with the base model of the Logistic Regression model with TF-IDF features alongside other features that we thought could improve depression detection. Features we used were sentiment of posts, number of posts, length of posts, average time between posts where only average time between posts improved base model on certain data subsets. With that model we achieved a F1 score significantly higher than one presented in Losada et al. Our main goal was to find words or groups of words that contribute to detecting depression presented in section 5. Our work could be further improved by using deep learning models, including additional features that were not investigated in this paper such as gender of user or experimenting with different word embeddings.

8. References

[1] Mohammed Al-Mosaiwi and Tom Johnstone. In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. Clinical Psychological Science, 2018. [2] W. Bucci and N.

Table 1: F1 results from various methods

Model	F1 @ 5	F1 @ 10	F1 @ 20	F1 @ 50	F1 @ 100	F1 @ 200	F1 @ 500
RANDOM F1	0.1959	0.1959	0.1959	0.1959	0.1959	0.1959	0.1959
STRATIFIED F1	0.2060	0.2060	0.2060	0.2060	0.2060	0.2060	0.2060
LR TF-IDF	0.4375	0.5454	0.5903	0.64	0.6984	0.6871	0.6555
LR TF-IDF + SENT	0.3510	0.4984	0.5451	0.5981	0.6505	0.6395	0.6374
LR TF-IDF + POST LEN	0.3529	0.5012	0.5643	0.6334	0.6719	0.6570	0.6386
LR TF-IDF + SENT + POST LEN	0.3587	0.5043	0.5741	0.6195	0.6519	0.6471	0.6357
LR TF-IDF AVG DIFF BETWEEN POSTS	0.4375	0.5172	0.5669	0.6451	0.7086	0.6818	0.6722

Freedman. The language of depression. Bulletin of the Menninger Clinic, 1981 [3] David E. Losada and Fabio Crestani. A Test Collection for Research on Depression and Language Use. [4] Adrian Benton, Margaret Mitchell and Dirk Hovy. Multitask Learning for Mental Health Conditions with Limited Social Media Data. [5] Luka Banović, Valentina Fatorić and Daniel Rakovac. How Soon Can We Detect Depression? [6] Donik Vršnak, Mate Paulinović and Vjeko Kužina. Early Depression Detection Using Temporal and Sentiment Features. A. Genkin, D. Lewis, and D. Madigan. Large-scale bayesian logistic regression for text categorization. Technometrics, 49(3):291–304, 2007.