Detecting Depression in Social Media

Project Overview

Copied from MLND project proposal by Anne Bonner
July 7, 2019

Proposal

Domain Background

More than 300 million people suffer from depression and only a fraction receive adequate treatment. Depression is the leading cause of disability worldwide and nearly 800,000 people every year die due to suicide. Suicide is the second leading cause of death in 15-29-year-olds.[1] Diagnoses (and subsequent treatment) for depression are often delayed, imprecise, and/or missed entirely.

It doesn't have to be this way.

Social media provides an unprecedented opportunity to transform early depression intervention services, particularly in young adults. This project will work to capture and analyze patterns of social media activity associated with the onset and persistence of depressive symptoms. By building an algorithm that can analyze Tweets as exhibiting depressed features, it will be possible for individuals and their parents, caregivers, and medical professionals to analyze texts and social media for linguistic clues that signal deteriorating mental health far before traditional approaches currently do. Analyzing linguistic markers in social media posts allows for a low-profile assessment that can compliment traditional services and will allow a much earlier awareness of depressive signs than traditional approaches.

I am pursuing the concept that linguistic changes in Tweets may be used to construct statistical models that can detect and even predict early depression warning signs in

---

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

ways that can complement and extend traditional approaches to diagnosis.[2] Changes in language and activity have been consistently correlated with changes in activity on social media.[3] Emotional language and linguistic features used in social media posts have been proven to indicate feelings that characterize major depression.[4] People consistently use social media platforms like Twitter to share their thoughts and opinions in written form. Posts on Twitter are particularly useful for analysis because they are most commonly made in the course of daily activities and happenings, which provides a rich and consistent means for capturing behavioral attributes that are relevant to an individual's thinking, mood, socialization, communication, and activities. Because depression is an illness that so often requires the self-reporting of symptoms, social media posts provide a rich source of data and information that can be used to train an efficient model.

Problem Statement

Over the last few years, there has been growing interest in using social media as a tool for public health, ranging from identifying the spread of flu symptoms to building insights about diseases based on Twitter posts.[5] However, research on using social media for analyzing behavioral health disorders is still in its infancy. Park et al.,[6] found initial evidence that people post about their depression and even their treatment on social media. Eichstaedt, et al., found that Facebook language can predict depression in medical records.[7] De Choudhury et al.,[8] examined linguistic and emotional correlates for postnatal changes of new mothers and built a statistical model to predict extreme postnatal behavioral changes using only prenatal observations. Reece, et al., developed computational models to predict the emergence of Post-Traumatic Stress Disorder in

---

[2] https://www.nature.com/articles/s41598-017-12961-9

[3] http://www.munmund.net/pubs/icwsm_13.pdf

[4] https://arxiv.org/pdf/1804.07000.pdf

[5] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0012948

[6] https://pdfs.semanticscholar.org/8dd5/8913bd343f4ef23b8437b24e152d3270cdaf.pdf

[7] https://www.pnas.org/content/115/44/11203

[8] https://www.microsoft.com/en-us/research/publication/predicting-postpartum-changes-emotion-behavior-via-social-media/

Twitter users.[9] This highlights the potential of social media as a source of signals about likelihood of current or future episodes of depression.

With this project, I am working to expand the scope of social media-based mental health measures and use existing research that has proven the correlation between depression and specific linguistic features in order to build an algorithm that can signal early text-based warning signs of potential depression. [10][11] The final outcome of this project will be to analyze specific user Tweets in order to determine trends that point towards possible impending depression. By analyzing markers such as personal and possessive pronoun usage, overall sentiment, and use of relevant words associated with depression, it should be possible to create a model that can give an individual insight into his or her mental health and well being.

Datasets and Inputs

In order to build a depression detector, there are two kinds of tweets that will be needed for this project: random tweets that do not necessarily indicate depression and tweets that demonstrate that the user may have depression and/or depressive symptoms. The random tweets dataset can be found from the Sentiment140 dataset available on Kaggle. There are no publicly available datasets of tweets indicating depression, so these tweets will be retrieved using the Twitter scraping tool TWINT. The scraped Tweets will need to be manually checked and Tweets will be cleaned and processed by removing links, images, hashtags, mentions, emojis, stop words, and punctuations.

Because the nature of social media content poses serious challenges to applications of sentiment analysis, VADER will be utilized for the general sentiment analysis of Tweets. VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media for general sentiment analysis that is specifically attuned to sentiment in microblog-like contexts. It allows for not only the classification of sentiment, but also the associated sentiment intensity measures.[12] This is extremely

---

[9] https://www.nature.com/articles/s41598-017-12961-9

[10] https://www.groundai.com/project/utilizing-neural-networks-and-linguistic-metadata-for-early-detection-of-depression-indications-in-text-sequences/

[11] http://www.munmund.net/pubs/icwsm_13.pdf

[12] http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf

useful because Tweets often contain multiple sentiments. VADER doesn't require training data, but is constructed from a human-curated, valence-based, generalizable sentiment lexicon which is fast enough to be used with streaming data.

Solution Statement

While sentiment analysis is is well suited for determining whether or not text is positive, negative, or neutral, these categories do not tell you whether or not that text has characteristics typical of a person with depression. By training the model to look specifically at Tweets containing linguistic features characteristic of depression, it will be possible to train an algorithm to classify a Tweet as one that exhibits either "depressive" or "normal" features.

This project will utilize the Sentiment140 dataset as well a a dataset containing Tweets indicative of depression, meaning that the algorithm will be built and trained not only with the basic building blocks of sentiment analysis, but also with specific linguistic features of depression. By training the model on not only the Sentiment140 dataset, but also on Tweets specifically labelled with terms specific to depression, the algorithm will learn to identify not only negative sentiment, but specifically depressive terminology and features in order to classify Text and Tweets as either "depressive" or "normal."

Once the dataset has been built, the model can be trained. The model will be analyzed for accuracy and a classification report will be run to determine its precision and recall scores.

Benchmark Model

Because of the nature of mental illness and its subjectivity, it makes the most sense to use a binary classification model. For this project, I'll use Keras to build a LSTM with a convolutional neural network to determine whether or not Tweets appear to signal a depressed state of mind. The accuracy of the model will be evaluated and compared to a binary classification baseline model using logistic regression. The model will be analyzed for accuracy and a classification report will be run to determine precision and recall scores. To evaluate the effectiveness of this model, it will be run against another binary classification model using logistic regression. That model will be trained with the same training data and number of epochs.

Evaluation Metrics

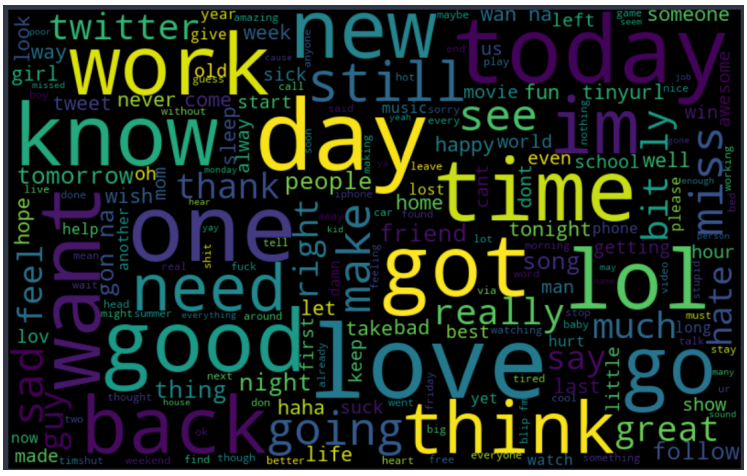
The accuracy of the model will be evaluated and compared to a binary classification baseline model using logistic regression. The models will be analyzed for accuracy and a classification report will be run to determine precision and recall scores.

Project Design

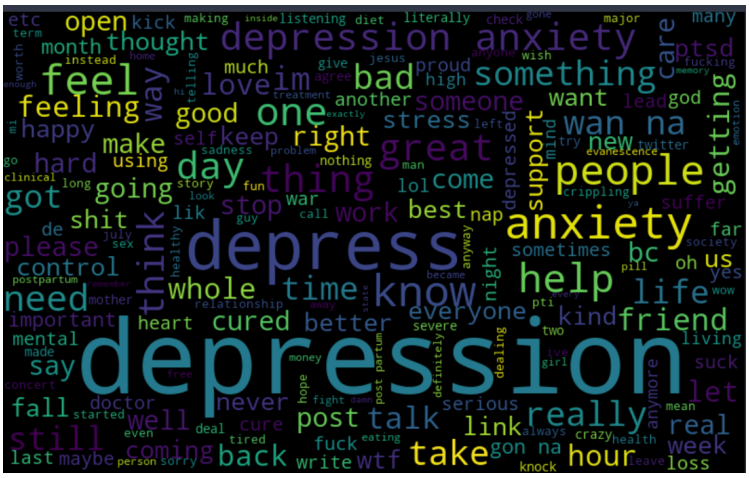
In order to build a depression detector, there are two kinds of tweets that will be needed: random tweets that do not indicate depression and tweets that show the user may have depression. A random tweets dataset can be found from the Sentiment140 dataset available on Kaggle, although it's worth exploring the Kaggle twitter_sentiment dataset to see if I can obtain more accurate results with Tweets that have been preprocessed to contain binary sentiment analysis results. There are no publicly available datasets of depressive tweets, so tweets indicating depression will be retrieved using the Twitter scraping tool TWINT. The scraped Tweets will need to be manually checked and all Tweets will need to be cleaned and processed by removing links, images, hashtags, mentions, emojis, stop words, and punctuations. NLTK will be utilized for stemming, lemmatization, and tokenization. VADER can be used as a tool to determine the sentiment of the cleaned Tweets.

Once the Tweets have been cleaned, it's easy to see the difference between the two datasets by creating a word cloud with the cleaned Tweets. With only an abbreviated TWINT Twitter scraping, the differences between the two datasets are clear:

Random Tweet Word Cloud:



Depressive Tweet Word Cloud:



The two datasets will be combined in order to build a single dataset that will allow the model to quickly learn to identify Tweets with depressive linguistic markers. The datasets will be split into training and testing sets, combined, and shuffled. Once the data are ready, a model will be built. The model will take in an input and output a single number representing the probability that the tweet indicates depression. The model will take in each input sentence, replace it with its embeddings, then run the new embedding vector through a convolutional layer. The convolutional layer will take advantage of that in order to learn some structure from the sequential data which will be passed into a standard LSTM layer. The output of the LSTM layer will be fed into a standard dense model for prediction.

Results can be summarized and compared against a benchmark logistic regression model in order to evaluate the effectiveness of the model. Once the trained model has been optimized, it can be used to analyze Tweets (and potentially texts and other social media posts) for early linguistic signs of depression, allowing a quick, effective, low-cost, and low-profile means of early awareness and intervention.