DETECTING SUICIDAL PATTERNS USING TWITTER DATA

Sandhya

May 10, 2018

1 Abstract

In this paper, I aimed to use twitter data to detect patterns of suicidal behavior. Twitter data was collected from 232 individuals marked with a presence of suicide or healthy status. The objective of this research is three-fold: First, to analyze the timely change in individual's speech as they get close to their committing suicide. Whether there are key differences in the patterns and behaviors among them and those who are marked healthy. Second, to build a predictive model to predict the risk of suicide in Twitter users. Lastly, to analyze and compare the network structure of user-follower relationship among the two groups. The results showed the tweets of individuals marked with suicide became less positive in the last 10 days before the suicide. The number of tweets was lower implying that individuals became less active. The network structure of healthy individuals are much dense and more connected than the ones marked as a suicide. The binary predictive model based on support vector machine can correctly build a tweet belonging to suicide group 50% the time and correctly predict the other label 98% of the time.

2 Introduction

There are contrasting point of views regarding the advancement in social media.[2] found a correlation between time spent scrolling through social media apps and negative body image feedback. While [3] claims that social media can help easily connect people with the world. Some would suggest that social media has provided them with opportunities to financially sustain their businesses while others would argue that it has made people less ambitious and more self destructive. While both approaches might hold true, research shows that social media can be helpful in predicting the signs of depression based on her social media existence[1]. The way people use social media can be an indicator of their mental health state. As humans, we use words and gestures to convey our emotions and thoughts, to tell stories, and to understand the world. Machine learning and Natural Language Processing is making it possible to extract meaningful patterns from it. Using this project, I want to learn and apply those techniques to analyze users' tweets and see whether there are patterns that can hint towards an individual being prone to suicide. The idea is that if we are able to identify those patterns, we can use them to mark the individuals who are at risk of suicide and take early measures to prevent it.

3 Data

For this study, I am using a twitter data containing 61k tweets from 300 individuals from 2012 to 2017. The tweet dataset consists of tweets from twitter ids with a label that hints towards the healthy or suicide status of an individual. Out of 61k tweets, 36k tweets are from twitter ids with 'inconclusive' status. Hence those tweets have been excluded from the dataset. This forms a dataset of about 22k tweets of 232 twitter ids marked with status: Likely, Maybe and No.

3.1 Data Joining and Cleaning

One of the challenges to work with tweet data is free form text. So, Data cleaning becomes an important step before anything. Each tweet may contain links, abbreviation, keywords, emoticons, punctuation and even the misspelled words that a human can understand but are not part of any dictionary. So, it is crucial to streamline and clean the text in each tweet. Each tweet is pre-processed by first stemming it, removing stop words, URLS, punctuation, emoticons and any keywords that may not be adding any meaning.

For the data joining, the health status column was appended to each tweet based on twitter user id. A sentiment column was also calculated and added to the overall dataset. Other than that some helper columns from the timestamp like dayofweek, monthofyear etc were also separately appended as columns to make analysis easier to perform.

3.2 Data Distribution

As shown in the table below, the data is highly skewed. 94% of all tweets belong to 'healthy' group while only 0.4% tweets belong to 'affected' group with total count of 87 counts which may not be represent a good sample of overall population and may produce biased results.

Labels	No. of twitter ids	No. of tweets
Likely	4 (1.7%)	87 (0.4%)
Maybe	41 (17.6%)	41 (17.6%)
No	187 (80.6%)	187 20698 (94%)

4 Results

4.1 Text Analysis

Figure 1 shows the most frequent words used by the people who died of suicide seven days before their last tweet comparing to the Figure 2 which shows the same analysis over last 30 days before the suicide. The analysis can be extended to any time period by just specifying the number of days in the code line. The reason to do this analysis is to see if there is change of speech overtime as they progress towards the last day. The plot seems to highlight some interesting patterns as I see it. In the last 7 days, the more noticeable words are vicious, apology, hairloss, thank, love etc. Comparing it to 30 day plot, the more noticeable words are bath, want, guy, gear, love etc. This might not be a quantifying difference between the two groups but it hints towards the difference in the tone and topic of tweets as approaching towards the last day.

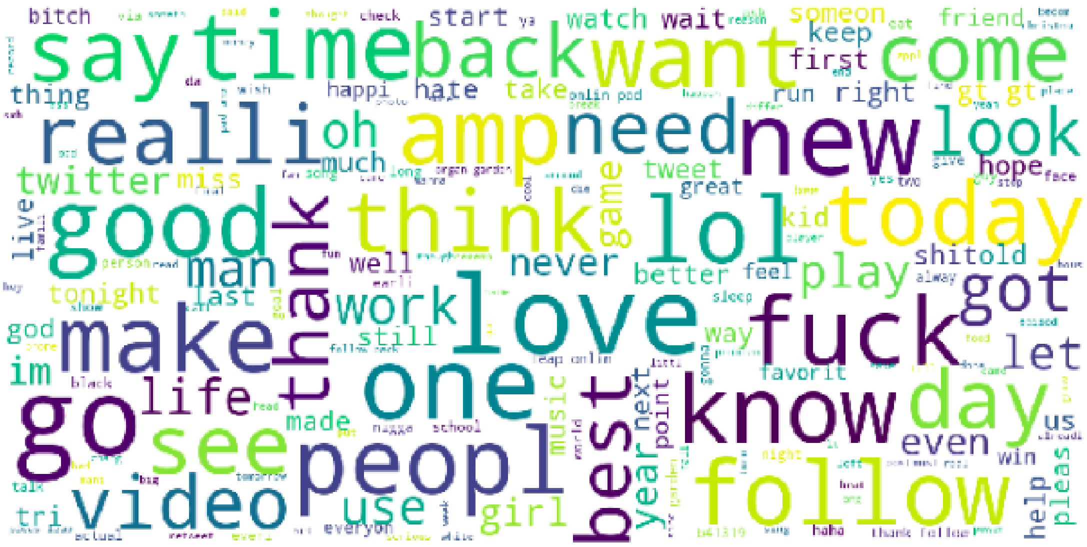

Figure 1: Frequent words in seven days before the death

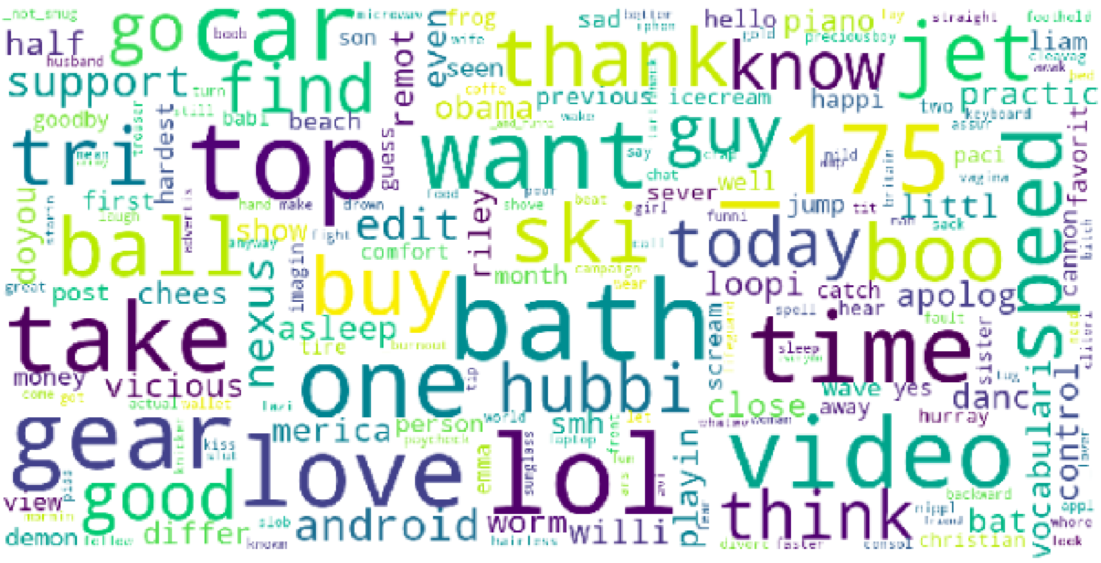

Figure 2: Frequent words in 30 days before the death

But looking only at the word frequency may not be a good idea because the total number of tweets (hence words) in those time period may vary hinting towards a biased result.

Figure 3 shows the average sentiment of tweets per over the last 15 days before the suicide compared to Figure 4 that shows the same analysis for healthy people. The reason to do this analysis to detect the overall emotion/sentiment of one day over the other. This might hint towards whether the pattern to be positive or negative on daily basis is same or different for suicidal vs healthy individuals.

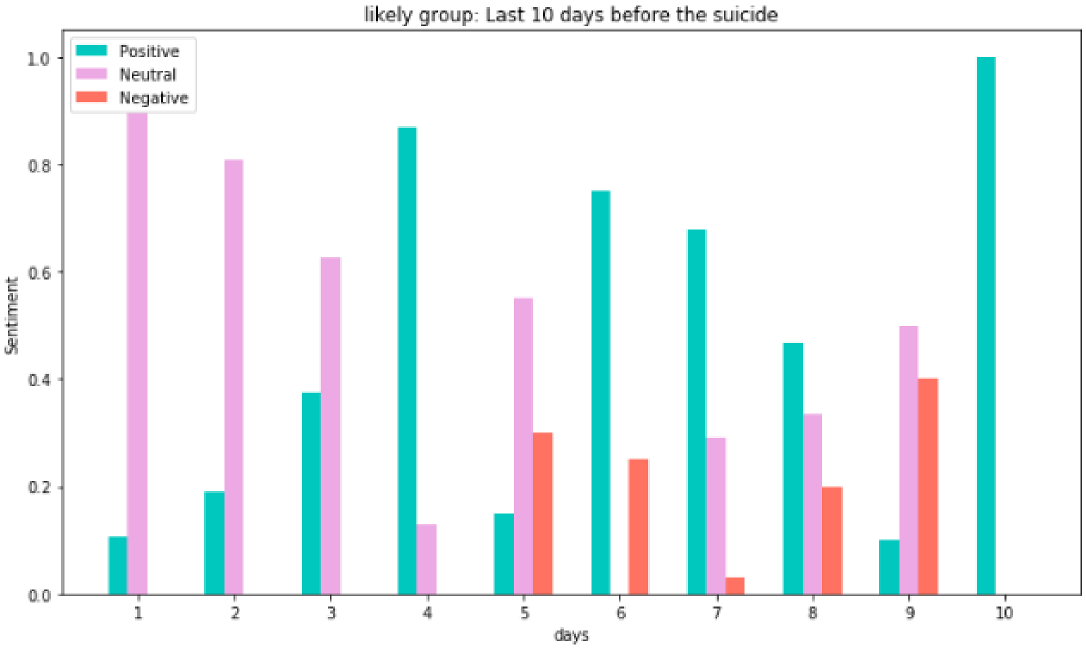

Figure 3: sentiment per day for last 10 days before the death

Figure 3 shows that people who are at suicidal risk may not express the negative speech but their tendency to be positive in the speech has lowered as approaching the last day. Surprisingly, the negative sentiment is missing altogether for last 4 days. The one reason could be the lower number of tweets implying that they become less active on social media.

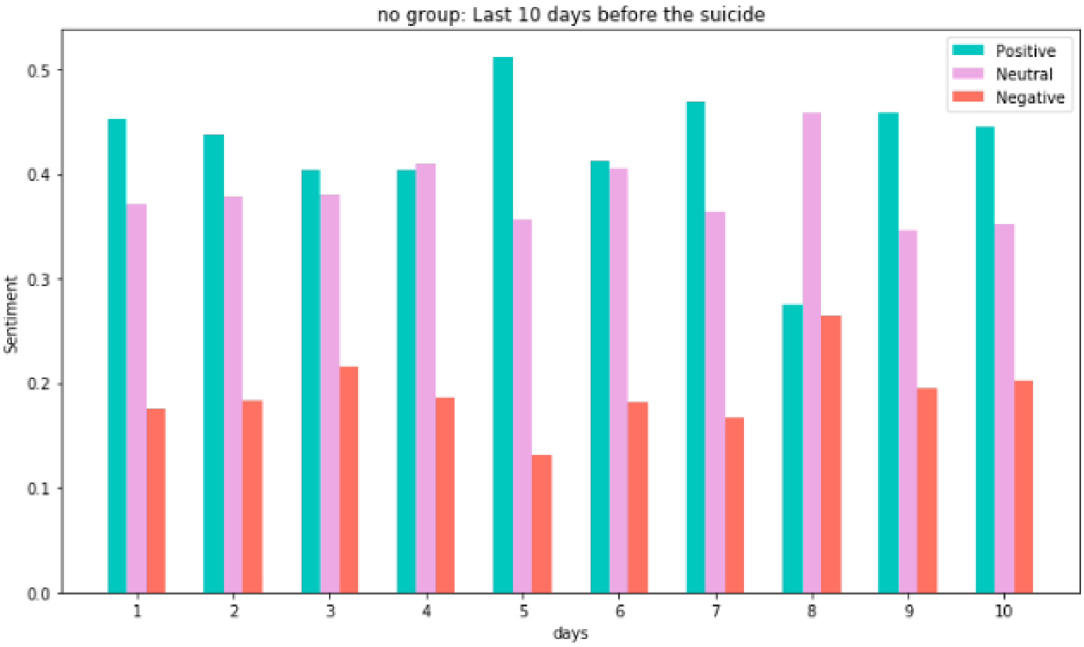

Figure 4: sentiment per day for last 10 days for healthy group

Figure 4 shows the average sentiment per day for healthy individuals. The results seems to be as expected. The sentiment seems to hint towards randomness or due to chance which it should be. Mood might be one factor behind different sentiments for each day but not following an ongoing pattern as it seems to be with the individuals with suicide status.

4.2 Statistical Learning

4.3 Feature Extraction

In order to use text in machine learning algorithms, it is required to convert them to a numerical feature vectors. Because the size of the data is not too big, I used count vectorizer to convert tweet text to numeric representation. This function counts the appearance of the words in each tweet. Once I got a vector, I applied tf-idf to normalize the word frequencies. This is to scale down the more frequent words in corpus that might be less informative overall. After these transformations, my data is sparse matrix with only tweet text information. So, I appended three columns to the sparse matrix namely dayofweek, sentiment and retweet_count to get data ready for building model.

When comparing the model performance, baseline model provides a point of reference. A baseline model can be a model that simply predicts the majority category class. Since this problem is already a imbalance class problem , I am treating Naive Bayes Multinomial as baseline model because of its high preference to majority class label.Naive bayes could only predict the true label for 'suicide' class 1% of the time (precision= 0.01). In order to have a better model, I considered support vector machine (SGDclassifier) which is widely regarded as one of the best text classification algorithms. I chose SVM also because the hyper-parameters are tunable and easier to play around with. I split the dataset into train and test set with 80% and 20% of the data. I used k-fold cross validation on the training set to unbiased evaluation of a model.

With default parameters, SGDclassifier performance was comparable to that of naive bayes. That is apparent because the data is imbalance. In order to avoid that problem, I used SVM to apply different weight to class labels. After tuning the weight parameter by running model on different values, 1:14 weight seemed to give the best result with precision with f1score of 0.5.

4.4 Feature Selection

The dataset at this point is high dimensional with 17195 features with 25k observations. I used Principle Component Analysis PCA to reduce the dimensions and chose top 2000 features to be included in the model. The number 2000 is chosen after building model on different no. of threshold feature.

In order to set class weight, I observed how the model performance (f1 score) is reacting to class_weight as shown in figure 5, which seems to be 1:13 weight ratio.

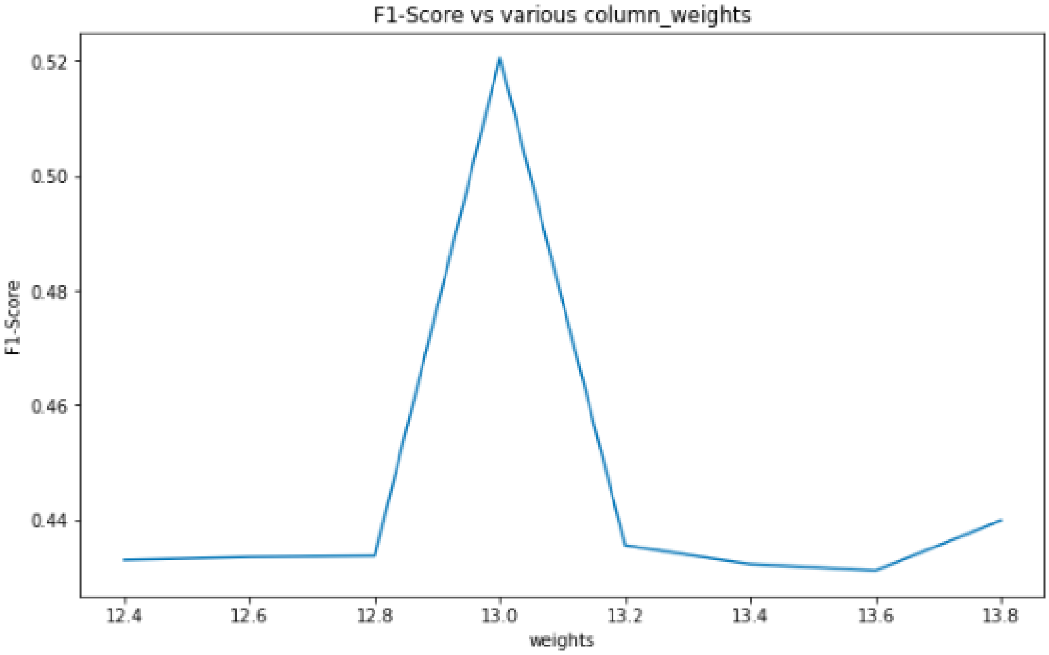

Figure 5: F1 score as column weight increased

The model trained on subset features from PCA seems to correctly predict the cases of 'No' suicide 98% of the time and correctly predict the 'Yes' cases only 48% of the time, giving total f1 score of 0.52 as shown in figure 6. These results are comparable to the model trained on all 17000 features.

Figure 6 shows the ROC curve which is a plot of the true positive rate against the false positive rate for the different predicted values. It is the measure of model performance. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test.So, ideally we want the orange curve as close to 1 on y-axis which is not the case for the trained SVM model.So, there is a need to get more data, come up with ways to handle imbalance class problem and perhaps try to apply different weights to different features in the model.

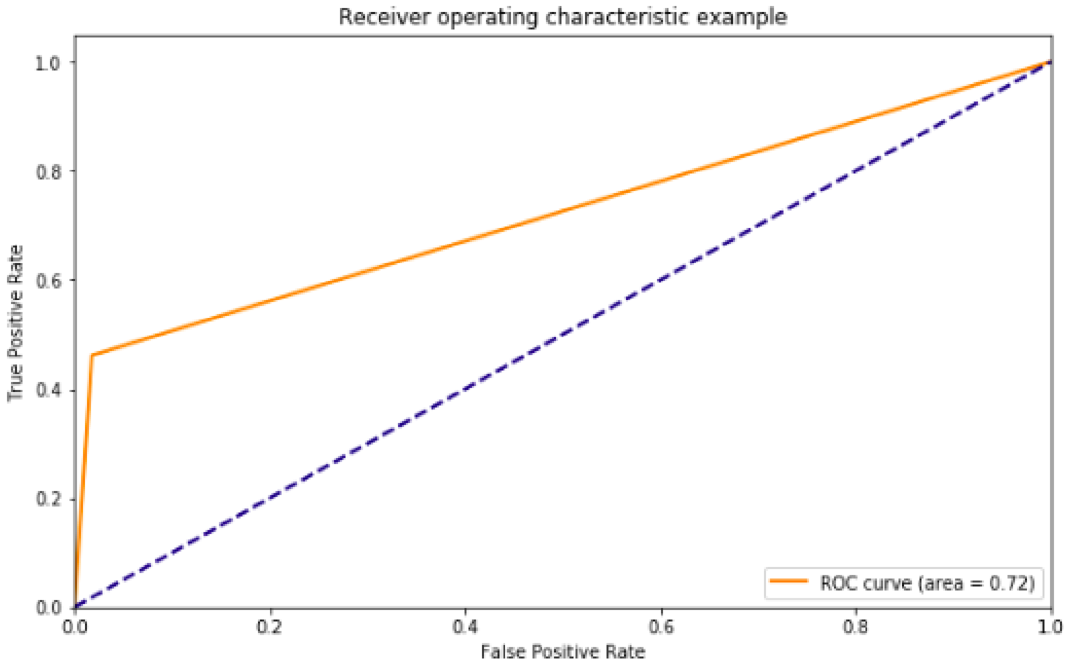

Figure 6: ROC curve

Figure 7 shows the confusion matrix that compares the predicted and test label accuracy in terms of Precision and Recall. The dark blue is true negative that is probability, the model correctly predicting the ’no’ label and bottom value on diagonal marked as 0.46 is true positive rate that is the probability, a model correctly predicted ’yes’ label. Ideally, we want these two probabilities to be 1. So, it is apparent that there is need to improve on the True Positive Rate.

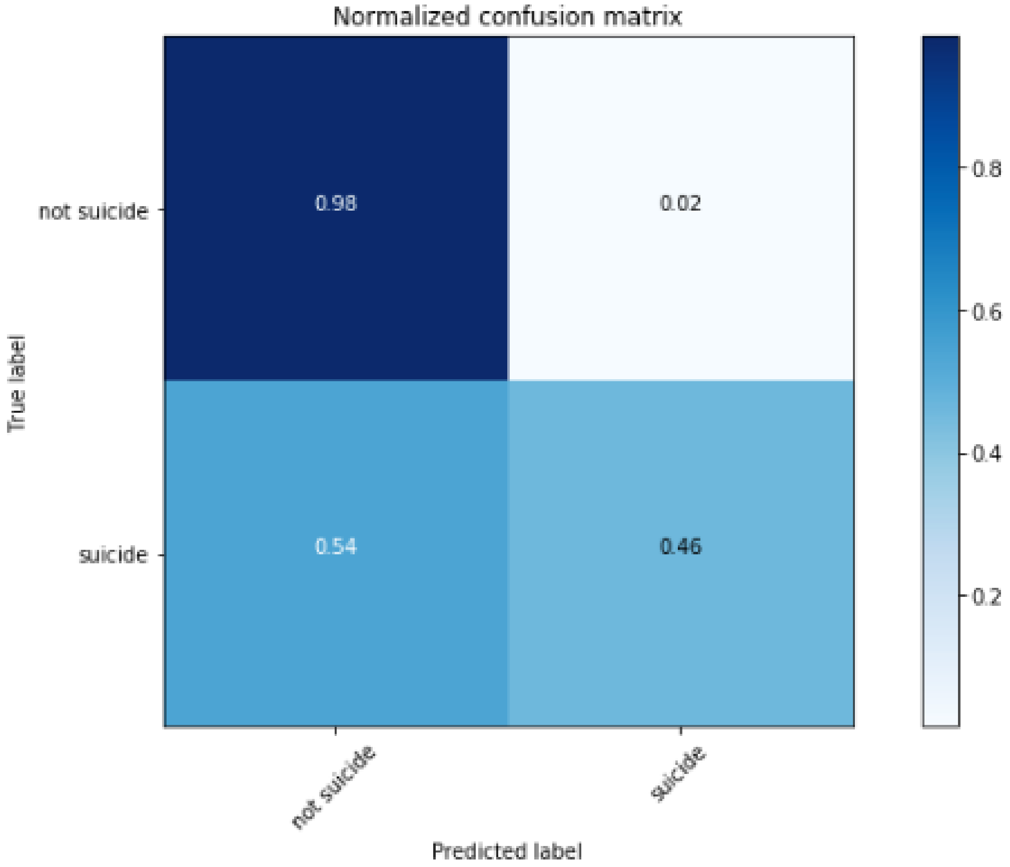

Figure 7: confusion matrix between true and false label

4.5 Network Analysis

In this part, the goal is to analyze and compare the network structures of affected and healthy group. Whether there is a visible patterns in the way these groups form their social network which, then, can be used to detect the patterns of suicide.

The network is a directed graph where each twitter ID is a node and a link represents a 'following' connection. A link from A to B would mean A follows B.[A—¿B]. Due to twitter rate limit and time constraint, I scrapped the data for first degree network of nodes in the data using Twitter API.

As indicated in table 1, the data is imbalanced, there are only 1% of twitter ids with likely status, 17% with maybe and 80% with no. So, the network will be heavily dominated by one group. In order to balance that out in a network structure, 15 nodes from each group were randomly sampled to be included in the network graph. This was done to get a clear picture of how network of each group is formed i-e affected and healthy.

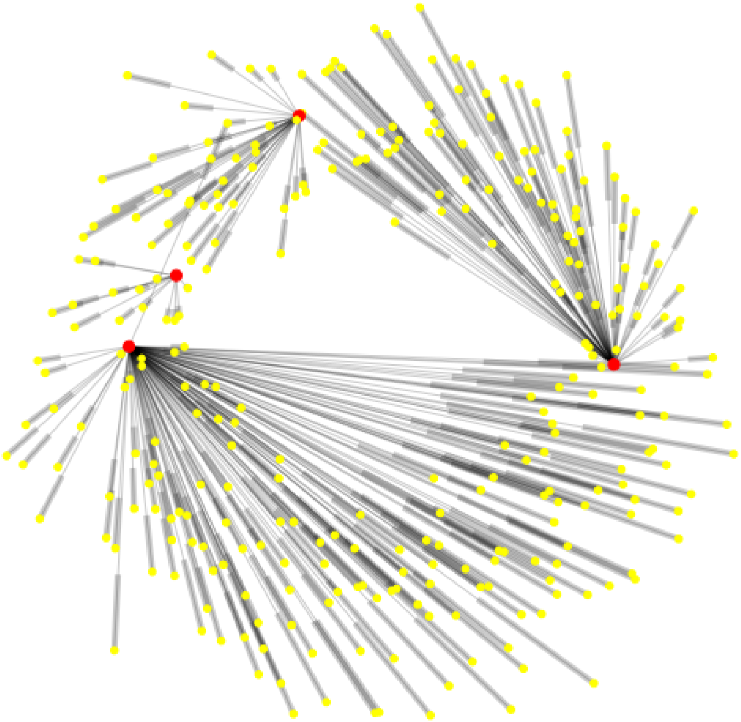

Figure 8: Network structure of Affected group

Figure 8 shows the network structure of affected group. The red colored nodes are the twitter ids of affected individuals and nodes in yellow are the people who they follow. Overall, it seems that this group is inclusive and not very dense, implying that individual who are affected do not form huge networks, but it is hard to say because there are only 4 nodes. This network might not be good representative of overall network structure of affected group.

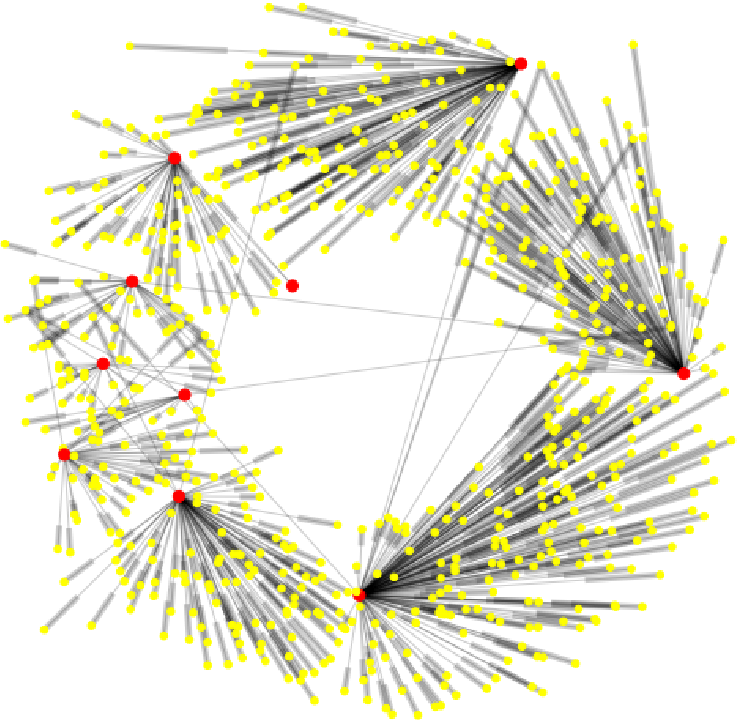

Figure 9: Network structure of Healthy group

Figure 9 shows the network structure of healthy group. These are the nodes randomly chosen from this group for plotting purposes. Overall it seems that that individuals in this group are more active and their network structure is much more dense. Because it is first degree network structure, all the healthy nodes seem to act as a hub in themselves. Moreover, it is interesting to see that this each sub network in this network seem to have links which are altogether missing in the figure 8 where each sub network seem to be isolated.

A more insightful approach would be take few twitter ids from each group and recursively get friend of friend-list to get multiple degree network structure. This would give the more rich results in terms of communities, degree distribution and hubs in the network.

5 Discussion

The study shows some of the differentiating factors like word usage, reduction in positive tone and network structure that may be helpful in detecting people who are at risk of suicide than those of healthy individuals. However, it seems that more data would be helpful in reaching any solid conclusion about the overall population of affected group. Moreover, advanced and useful NLP techniques like word-shift, emotion detection and topic detection can be used to further improve the results.

References

[1] Forecasting the onset and course of mental illness with Twitter data
https://www.nature.com/articles/s41598-017-12961-9

[2] Looking At This Social Media Platform For Just 30 Minutes Can Affect Your Body Image
https://www.refinery29.com/2017/09/171414/social-media-body-image

[3] Depression disclosures by college students on a Social Networking Site
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3110617/