The Role of the Social Network in Discovering Natural Helpers

Anonymous

Abstract—Pro-Social behavior is intended towards helping other people to build social support. Social networking sites like Twitter provide the perfect vantage point to observe such pro-social practices. Analyzing the interactions between users on Twitter will help in discovering pro-social behavior. Pro-social behavior can help us in finding where there is a booming quality of life or where group building activities are more needed to strengthen social coordination. In this paper, we discover pro-social behaviors on Twitter by doing natural language processing, keyword searching and analyzing the social network graph. Also, we build an encoder-decoder neural network model to identify Pro-Social behavior on the social network.

I. INTRODUCTION

Prosocial behavior benefits other people or society by donating, volunteering, helping, sharing, etc. There are many motivation for prosocial behavior. One is altruism, i.e., to help others with no expectation of compensation, and another is empathy, i.e., to place oneself in another person position. People can exhibit prosocial behaviors depending on the relation of the individual with his fellow citizens (the social network) [1]. The characteristic of the social network with the prosocial network is described in two contexts, formal and informal. "Formal" means contributing resources like donating blood through a nonprofit organization, while "informal" means helping another person via personal relationships [1]. One motivation for discovering informal prosocial behavior is to find where the quality of life better and where we need more community building to ensure the quality of life is better. The core idea of the paper is to analyze Twitter data and find informal prosocial behavior (can also be termed as the natural helper).

Natural helping is a robust strategy to make peace during violence by building trust among neighborhood and strengthening positive social norm [2]. Prosocial behavior promotes resilience and mental wellness [3]. Empathy reduces the risk of violence if neighborhood contains more natural helpers. Does the neighborhood location require less police presence where the natural helper reside, which we found them on the social network? Does natural helper which we identify from Twitter, have a real impact on the neighborhood?

In this paper, we perform a social network analysis with keyword searching (NLP) to find natural helpers. The previous work includes the finding of natural helpers using a predictor "thankfulness" among user and by building a central helper graph [3]. One of the drawbacks is they only consider thanking term to determine the natural helper, whereas there may be a wide range of different communities which have their specific way of expressing gratitude to others. In this paper, we examine different communities and their way of showing gratitude.

II. RELATED WORK

Dominijanni et al. [3] talked about discovering prosocial behavior on the social network by analyzing the user’s tweet. They used thankfulness, a predictor among people using Twitter to find natural helpers. Researcher hypothesize that natural helpers receive more gratitude from more people than the average users receive on twitter [3]. They study a social graph, where nodes are users and edges represent a message.
containing a thank you. Central helper is defined as the helper who had received a thank you message from more than one user. They conclude with the twitter used by the member of central helpers helps in building and maintaining relationship[3].

Homan et al. [4] talked about the pattern related to depression in a social network like TrevorSpace which is an online community dedicated to preventing suicide among lesbians, gay, bisexual, transgender and questioning youth. The researcher did social network analysis to detect the depression on the social network. To study the pattern related to depression, they study six different feature of the network: degree, triangle, clustering, core number, clustering coefficient and size of largest connected and biconnected component. They found some correlation between the feature of the network graph and the users of TrevorSpace showing a high depression. They run two different level of test for each of six different network feature. The result they found was consistent with previous research about the suicide prevention and have clinical implication [4].

Hagberg, Schult and Swart [5] talked about the networkX, a python library which is used to analyze social network and network algorithm. It includes data structures that can represent various graphs such as the directed graph, undirected graph, graph with a parallel edge and self-loops. The networkX provide many inbuilt properties such as shortest paths, betweenness centrality, degree distribution and clustering, etc. [5]. NetworkX with other existing application in python builds a robust social network analysis platform. We will be using networkX to create social network graph which will find the natural helpers on twitter.

Zaman, Hossain and Kautz [6] talked about how social media like Twitter can be used to improve the situations awareness in an emergency situation. The researchers aim to find useful information on Twitter for the users on a regular basis, by conducting an exhaustive analysis of tweet [6]. They did social network analysis to find the tweet that was about the emergency situation. They took the 911 emergency data and compared it with the user’s tweet to see if they could get any insight on the data. The result they found was surprisingly negative, the majority of the tweet in the same area where the emergency situation arose was not of the situation. They found very few percentages of users who created awareness by sharing the information about the situation.

Caughy, Leonard, Beron, and Murdoch [7] analyze the effect of neighborhood on child behaviors. The researchers examine the relationship between the child behavior and neighborhood physical, social condition and the geographic peers. They study the neighborhood with low income and ethnic minority that are located in the Southwest United States [7]. The researchers applied spatial analysis method to observe the effect of neighborhood on child behaviors and was able to find the distance surrounding the child house that affects child behavior. It would be interesting to find the correlation between natural helper in the same area as the child, i.e., the area with less natural helper might affect more on child behavior.

III. METHODS

We use twitter data extracted from Data-sift containing roughly 6.6M Record from July 2013 to June 2014. The data include tweets from New York State[3]. The dataset contains information like tweet, userID, mentionUserID, location (latitude and longitude), City, Time, etc. The Github link for this project is as follow:

https://github.com/tripperroc/natural-helpers

A. Social Network Graph

We estimate natural helpers receive more expression of appreciation, from individuals, than the average Twitter user gets. We used natural language processing in addition to the social network analysis to determine the natural helpers. We use two set of keywords. Set 1 keyword to contain thanking predictor, identified by the scientist are:

Set 1 Keyword = {thank, thanks, thank you, ty}

For this paper, we consider a different community, i.e., youth, where they have a different way of expressing gratitude between users. For example, tweets like ‘blessed to have the best bestest friend @k_gilbs3’ or uses of emojis in tweets like ‘@josephvalvarez is my savior right now’. Set 2 keyword identified by the community partners are:

Set 2 Keyword = {blessed, bless up, major key, much love, props, respect, give me life, preach, real mvp, shoutout, OG, S/O, keep it real, is the bomb, is the man, is the shit, lit, 😂😂, 😂, 😂, 😂, 😂, 😂}

For the initial segment of the task, we use both sets of the keyword to build a social network graph, where the node represents a user and edge between the users represent the message containing keyword from either of the set. We infer the nodes with more number of incoming edges, uncovers the pro-social behavior. Also, they often act as social bridges between different communities and tend to recognize and utilize other helpers.

![Fig. 2. Keyword Searching Block Diagram](image)

For each pair of users (u,v), the edge is defined from user ‘u’ to user ‘v’ if a message from ‘u’ to ‘v’ contains keywords from either set. Using the social network graph, we describe various subgraph Reciprocal graph (R), Mutual Gratefulness Graph (M), Binary Pairs (BIN), Helpers Graph (H). We further divide the helper’s graph into Terminal Helpers (T) and Central Helper (CH) [3].
Reciprocal graph (R): The edge is present between two users if a user has mentioned the other user [3]. 
\[ R = \{ u,v \} \mid \min\{ \text{weight}(u,v), \text{weight}(v,u) \} > 0 \]

Mutual Gratefulness Graph (M): The edge is present between two users if a user has mention keyword from either set in the tweet to the other user [3]. 
\[ M = \{ u,v \} \mid \min\{ t - wt(u,v), t - wt(v,u) \} > 0 \]

Helpers Graph (H): The Helpers graph is an intersection of mutual gratefulness graph with the additional constraint of excluding pairs of individuals who thank each other, but no one else [3]. 
\[ H = M \cap \{ u,v \mid \min\{ \delta(u), \delta(v) \} > 1 \} \]

Binary Pairs (BIN): The Binary pair graph is an intersection of mutual gratefulness graph with the additional constraint of including pairs of individuals who thank each other, but no one else [3]. 
\[ \text{BIN} = M - H \]

Central Helper (CH): In central helper graph, the edge is present between two users if it belongs to helper graph and the indegree of a node is greater than 1 [3]. 
\[ CH = \{ u \in \text{helpers} \mid \deg(u) > 1 \} \]

Terminal Helpers (T): In terminal helper graph, the edge is present between two users if it belongs to helper graph and the indegree of a node is not greater than 1 [3]. 
\[ \text{TH} = H - CH \]

Apart from this graph we also construct a graph CR, which contains the nodes of R and additionally satisfy the graph constraints we place on CH [3].

**B. Encoder Decoder Model**

One of the key challenges in finding natural helpers on social network graph is to examine the keywords carefully for the correctness of the result. We must identify a way which is efficient in determining natural helpers from different communities. The idea is to build an encoder-decoder neural network model. Using this, we don’t have to worry about missing of keywords or the context of the tweet. However, One challenge is to get the data for training. There is two way that we can get the training data, one is to label the subset of the entire dataset manually, and another one is to crowdsource like mechanical Turk where a random user label the data for money. For our project, we label the data manually. The question we ask when labeling the data is “Is the tweet expressing thankfulness, gratefulness, affirmation, appreciation, or respect (y/n/m)?". If the answer is yes, there are many possible follow up questions that we are interested in like,

- Who is the object of thankfulness, gratefulness, affirmation, appreciation, or respect expressed in the tweet?
- Why is the tweeter expressing thankfulness, gratefulness, affirmation, appreciation, or respect?
- Is the tweeter expressing thankfulness, gratefulness, affirmation, appreciation, or respect for benefits personally received or was the target a third party (y/n/m)?
- Was the tweeter expressing thankfulness, gratefulness, affirmation, appreciation, or respect for online or offline actions (y,n,m)?

For this project, we only consider the base question, i.e., “Is the tweet expressing any gratitude?". We divide the output class into three labels.

- Yes(y): Tweet expressing thankfulness, gratefulness, affirmation, appreciation, or respect
- No(n): If it does not express any thankfulness, gratefulness, affirmation, appreciation, or respect
- Maybe(m): Cannot decide whether the tweet is of thankfulness or not.

Once we get the data to train, we build an encoder-decoder model. Encoder Decoder model, a recurrent neural network, is a sequence to sequence model used for predicting the sequence of words. Application of the encoder-decoder model is text translation, for instance, English to French, or text summarizing. It is also used for text classification. The traditional neural network requires creating of two separate model for a sequence to sequence problem, which then creates problem while merging them into one and generating the output sequence. The encoder-decoder model solves this problem with ease, but with one restriction, length of output cannot be greater than the input. This should not be the problem in our case as the model output is of only one length. The encoder-decoder LSTM model was developed to tackle the natural
language problem. The encoder and decoder model consist of two stages:

- Encoder: The Encoder reads the entire input sequence (in our case tweet of the user) and converts it into an internal representation, often a fixed length vector. We ignore the encoder output.
- Decoder: The Decoder takes input from the Encoder which is a fixed length vector and outputs the class (in our case y, n, or m).

![Encoder Decoder Model](image)

From the first table, we observe that the "@ message" and mean of "@ thanks" grow drastically. One interesting thing is the first table which consists of only thanking keyword has less "@ thanks" concerning the second table. In table III, we find the number of users is in both the graph, number of users that are in thanking graph and not in other graph and vice versa.

![Social Network Graph](image)

From the table I and II we observe that the "@ message" and mean of "@ thanks" grow drastically. One interesting thing is the first table which consists of only thanking keyword has less "@ thanks" concerning the second table. In table III we compare both the graphs, i.e., graph constructed from only set 1 keywords vs. graph built from set 1 and set 2 keywords. In table IV, We find the number of users that are in both the graph, number of users that are in thanking graph and not in other graph and vice versa.

![Experiment Results](image)
we observe that the common user acts as a bridge between different communities. Next stage is to validate the two code, one is the current coder, and another one is the older version of the code from UR. In Table V and Table VI, we contrast both the code, i.e., graph constructed from current code vs. graph constructed from older code. Table V shows the number of users different in both code and Table VI shows the different users from both the code.

### TABLE V. COMPARISON BETWEEN NUMBER OF USERS DIFFERENT IN OLDER CODE FILE AND CURRENT CODE

<table>
<thead>
<tr>
<th>Subgraph</th>
<th>Common User Users in older file only</th>
<th>User in current code only</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>13320 0 0</td>
<td>0 0</td>
</tr>
<tr>
<td>RXM</td>
<td>12320 9 0</td>
<td>9 0</td>
</tr>
<tr>
<td>M</td>
<td>876 0 9</td>
<td>0 9</td>
</tr>
<tr>
<td>H</td>
<td>358 0 3</td>
<td>3 3</td>
</tr>
<tr>
<td>BIN</td>
<td>516 2 8</td>
<td>2 8</td>
</tr>
<tr>
<td>TH</td>
<td>215 2 2</td>
<td>2 2</td>
</tr>
<tr>
<td>CH</td>
<td>141 0 3</td>
<td>0 3</td>
</tr>
</tbody>
</table>

### TABLE VI. COMPARISON BETWEEN DIFFERENCE OF USERS IN OLDER FILE AND OUR CURRENT CODE

<table>
<thead>
<tr>
<th>Subgraph</th>
<th>Users in older file only</th>
<th>Users in current code only</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>RXM</td>
<td>emiltsch, baazpollock, marilynten, kendallcolton, noelle_marie07, fritz_ellie, andi_culver, gymnasticasey1, tippingtmidia</td>
<td>NA</td>
</tr>
<tr>
<td>M</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>H</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>BIN</td>
<td>house_steph, hannwilliams34</td>
<td>NA</td>
</tr>
<tr>
<td>TH</td>
<td>Tookerann, <em>Torrie</em></td>
<td>NA</td>
</tr>
<tr>
<td>CH</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

To view the User Tweets that are different, you can go to: https://github.com/tripperroc/natural-helpers/blob/Rinkesh_branch/compare_two_csv.txt

Along with the result, we also build the Wordcloud for the central helper users using the set1 keywords to see the various keywords that user uses. We see some interesting result, where we don’t expect terminal helper to use the word like fuck and shit regularly. However, we did expect words like love.

![WordCloud of the Central Helper](image)

**Fig. 5.** WordCloud of the Central Helper

### B. Encoder Decoder Model

We first label the 1000 tweet where 600 tweets contain set2 keywords, 200 containing set1 keywords and 200 are the random tweet. We divide 1000 tweet into two datasets, i.e., 800 train and 200 test data. The result had a low precision and low recall. One of the reasons is because of imbalance of classes. 'Yes' dominate the 'no' class and the 'maybe' is in deficient proportion. So we changed the 'maybe' to 'no' so that we balanced the class and reran the classifier. The result returned was a little better but was not convincing enough. So we decided to label more 1000 data in the same manner and rerun the classifier on the whole data set. We obtain a good recall and precision, but it was still not good enough to run the model on the entire twitter dataset. So we decide to label more 1000 data which is completely random to balance out the 'no' class and reran the model over entire 3000 data set. The result returned was biased, i.e., the model was predicting only one class ('y'). One of the reason is of data not structured properly. Table VII shows the metrics for 1000 and 2000 dataset.

### TABLE VII. RESULT FOR 1000 AND 2000 TWEETS

<table>
<thead>
<tr>
<th>Data</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>(800 train + 200 test data)</td>
<td>0.65</td>
<td>0.45</td>
<td>0.54</td>
</tr>
<tr>
<td>(1600 train + 400 test data)</td>
<td>0.78</td>
<td>0.65</td>
<td>0.71</td>
</tr>
</tbody>
</table>

There were a group of 4 people who label the same data to see the difference between people from different communities about how one look at the expression of gratitude. We use the Cohen-Kappa Score, which measures the inter-rater agreement between different user. We found some great result, except for the one user "Homan" who had different opinions than the other users. Table VIII shows the Cohen kappa score among the users.

### TABLE VIII. COHEN-KAPPA SCORE BETWEEN DIFFERENT USERS

<table>
<thead>
<tr>
<th>User 1</th>
<th>User 1</th>
<th>Cohen Kappa Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rinkesh</td>
<td>Amen</td>
<td>0.91</td>
</tr>
<tr>
<td>Rinkesh</td>
<td>Ann</td>
<td>0.90</td>
</tr>
<tr>
<td>Homan</td>
<td>Amen</td>
<td>0.55</td>
</tr>
<tr>
<td>Homan</td>
<td>Amen</td>
<td>0.56</td>
</tr>
<tr>
<td>Amen</td>
<td>Homan</td>
<td>0.57</td>
</tr>
</tbody>
</table>

One way we can validate our keywords result is to generate confusion matrix between the labeling vs. the keyword. One changed we made in the labeling is to change the 'maybe' class to 'no' or 'yes' as the 'maybe' class is not present in the keywords sections. The accuracy, precision, and recall were pretty good. Table IX and X shows the confusion matrix.

### TABLE IX. CONFUSION MATRIX (CHANGING MAYBE TO YES)

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Manual Labelling labels</th>
<th>yes</th>
<th>no</th>
<th>1156</th>
<th>340</th>
<th>1027</th>
</tr>
</thead>
</table>

### TABLE X. CONFUSION MATRIX (CHANGING MAYBE TO NO)

| Keyword | Manual Labelling labels | yes | no | 1234 | 385 | 982 |

### V. CONCLUSION

In this paper, we analyzed the two graph from set 1 and set 2 keywords. The result demonstrates the possibilities of
two communities which has a different way of expressing gratitude and the common users from both graphs as a bridge between them. We also observed the person from different communities has their specific way of seeing the appreciation in the tweet. The encoder-decoder model promised good results but not convincing enough to run it on an entire dataset. The confusion matrix of keywords vs. labeling done by the human showed proper validation of the keywords. However, we need to be more careful while examining the keyword. In the future, we plan to get more training data either by doing it manually or from crowdsourcing website like Mechanical Turk. We may also want to improve our encoder-decoder model by using attention mechanism or changing the dense layer. After getting a convincing result from an encoder-decoder model, we will use the model to classify the entire twitter dataset and use the same approach of the social network graph to determine the natural helpers. After that, we want to know whether the natural helper that we identified have any impact on their neighborhood.

VI. ACKNOWLEDGMENT

I would like to offer my gratitude to Prof. Christopher Homan for consistently being an excellent tutor and guiding me throughout my project. I would also want to thank Ann Marie White and Amen Ptah from the University of Rochester, Medical Center, for constantly supporting and guiding me. I would also like to thank Prof Bailey for preparing me for the milestone presentation and poster.

REFERENCES


