Influence Network in Twitter

Deekshant Mamodia*
deekshant19119@iiitd.ac.in
Indraprastha Institute of Information Technology
Delhi, India

Jain Piyush Pradeep*
piyush19122@iiitd.ac.in
Indraprastha Institute of Information Technology
Delhi, India

ABSTRACT

Microblogging is becoming a popular communication tool for expressing the views and influence the people over a topic around the world. Here, twitter is used to describe the topology and information spreading dynamics of Online Social Networks. In twitter among million of users a small percentage of influencers present. They have followed by large number of active user that consumes and adding content to their tweet and propagate into the network. The research purpose of this project is to determine a grounded approach for measuring social networking potential and the influence of the alpha user of individual Twitter user in the twitter retweet network and follower-followee network . We have presented various models to find influencers in the network with the help of various features extracted and developed from the twitter network. Along with this a method was proposed to find the next possible retweet if we are given the information about the user and current tweets.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning algorithms*; • **Information systems** → *Collaborative and social computing systems and tools.*

KEYWORDS

datasets, neural networks, graph neural network, influencer, twitter network

ACM Reference Format:
Deekshant Mamodia and Jain Piyush Pradeep. 2020. Influence Network in Twitter. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PROBLEM STATEMENT

The aim of our project is to determine the topical influencers in the online social networks. The other goal is to determine the next influencer in the network based on the retweet data of the user and the topological similarity of the existing influencers.

*Both authors contributed equally to this research.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 MOTIVATION

In recent times, there has been tremendous growth in the number of users on a social media platform, a large amount of textual information is sent across the world via the social networking sites, every day. Twitter is one of the social networking sites where information is spreading by a small set of people, influencers, alpha users that evolve small micro-networks around them. These small micro-networks are generated by subscribing or following them. The tweets updated by the influencers are consumed and multiplied by their followers and propagate into the network. The major interaction method used to propagate their messages is called "retweeting". The retweeting means to publishing the tweets of followee or repeating the message of the followee. This mechanism generates the trending topics, and people use them to discuss and exchange ideas without the necessity of having any explicit relation.

In recent times, finding influencers on social media is becoming increasingly important. It is useful for tasks like promotion of product, behaviour identification, making trend of a topic, analysis of people views on a topic. The social media influencers are also help in making decision and spreading of the messages with a rapid speed. So our main goal is to identify these influencers in the influence network of Twitter.

3 DATASET DETAILS

The mechanism used by twitter users for communicating with each other is the follower and followee relationship. To generate this relationship we have generated the directed graph in which each follower has directed link to their followee. Twitter REST API was used to build our dataset. They provide programmatic access to read and write Twitter data.

For the network, we have collected the tweets, retweets, follower and followee and many more attributes on the topic "CAA". We have used tweepy cursor API to get messages and other attributes with query operator "CAA". Using this dataset we explicit show the relationship between follower and followee based on retweets between them and also the follower and followee network. This follower and followee relationship is also viewed as the message spreading network. The table 1 shows the features of the directed follower and followee network of the twitter. The table 2 shows the features of the retweet network of the twitter,

Features	Value
Nodes	159833
Edges	206645

Table 1: Follower Followee Network.

Along with the retweet information other data related to tweets were also extracted these features are namely follower Id, followee

Features	Value
Nodes	86494
Edges	205814
Average Degree	4.7590
Average in/out Degree	2.3795
Average clustering coefficient	0.01045
Average neighbour degree	1.1747

Table 2: Retweet Network.

Id, follower name, no of followers of user ,no of friends of user, statuses count,tweet id, tweet text,retweet count of tweet, tweet time, retweet time, likes, language, followee friend.

4 BASELINE IMPLEMENTATION

The baseline paper uses information spreading mechanism and user participation on twitter. The paper strongly based on the degree distribution of the network. They develop the directed retweet network based on a specific topic where each node represents the users and the edges are linked according to who retransmit (retweeted) whose messages. The graph is unweighted. They measure the popularity of the user from the perspective of followers by calculating the Kin/Kout ratio. The rich get richer agglomeration of phenomena was found in the network.[3]

They also calculate Rin which is defined as the amount of retransmission by others for a user.The greater the Rin means a larger number of others influenced by the user. Then they find the relationship between retransmissions and users in degree and out-degree and they separated popular users from unpopular users. It can be seen that popular users can gain a large number of retransmissions with a low level of activity. In twitter, millions of user play the middleware role in the network so each user adds some contribution to the propagation of the message and the influence of the alpha user. They measure the influence based on the temporal behaviour of the tweets. User enthusiasm can be defined as the average ratio of the time that a user spent responding to a message to the persisting time of the message spreading process. They proposed that a higher ranking in the queue leads to a greater possibility for a user to influence others.

The second baseline was explaining the basic functions and communication possibilities on Twitter and their interpretation in the context of social influence. They analysed the Austria twitter users and identify the strength and weakness of existing grading concepts of rankings. They proposed the new metric Social networking potential (SNP) which determines the ability to influence a certain audience.[1]They calculate the three different performance indicator:

- Follower/Following Ratio (r_f) which compares the number of users who have subscribed to the updates of user A with the number of users that user A are following.
- Retweet Ratio (r_{RT}) detects how many of A's tweets imply a reaction from the audience.
- Interactor Ratio (r_i) must be determined. This is the number of individual users who retweet content or mention user A divided by the total amount of followers of user A.

When we focusing on the content-oriented interactions, retweet ratio must be looked carefully. The twitter is content-oriented but the follower and followee relationship should also be considered for the calculation process of the SNP. The two ratios are summed up and divided by 2 to calculate the social networking potential (SNP). A score of 100% means that all tweets of the user are acted upon and all followers interact with the user. The score may be higher than 100 % if a tweet is retweeted more than he publishes tweets himself.

5 METHODS

This section describe the methodology used for the identification of influencers in the twitter network.

5.1 Tweet Sentiment

The sentiment of the tweets can also play a significant role to define the characteristics of the user. This helps in finding the influential people in the network and finding the next possible retweet. Consider for our case major influential people were those who were in the support CAA. Based on this the tweets were marked as either 1 (Support CAA), 0 (Neutral), or -1 (Against CAA) using the sentiment of the tweet. For this major problem was that most of the tweets were in the regional languages which was a major problem so first the tweet text was converted to English language and than after that sentiment for tweets were analyzed. The number of nodes (tweet/retweet) with there tag are shown in table 3

Sentiment	Tag	Number of Nodes
Favours CAA	1	91487
Neutral	0	15931
Against CAA	-1	27729

Table 3: Number of twitted and there sentiment

5.2 Feature Generation

Feature of the tweets and users plays and important role in the task in prediction of whether user is an influencer or not. In these features some are generated from the twitter API and some features are developed by using the features of node(users) in the graph. The list of such features in detailed form are described below.

- **Tweet sentiment:** Sentiment of tweet related to topic plays a major role in finding the influential people or the next retweet. Sentiment of a user can be the sum of all the the tweet and retweet by a user related to a topic.
- **Tweet time:** Time at which the tweet was created also gives information about the user, for example, it can be combined with the retweet/likes at that some time to find the appeal of the tweet.
- **Retweet time:** Retweet time also plays an important role as it helps in finding how much interest does a user has in another user, for example, consider a person who is a die-hard fan of an actor than he/she will retweet it within few minutes or few second of the tweet.

- **Tweet likes:** The number of likes of a particular tweet can be used as a measure of how many peoples liked the viewpoint of the user.
- **Tweet retweet:** Similarly, as the tweet likes, no. of retweet can be used as measure that how many people agree to the user viewpoint and another thing is that increase in retweet increases the in-degree in the retweet network.
- **Follower count:** Follower count is one of the direct measures which can be used to measure the popularity of the user. Increased follower count generally increases the chance of a retweet.
- **Follower-Followee ratio:** The follower-Followee ratio is a much better measure as compared to follower count because some people try to increase their influence by following the people arbitrarily this increases there chances of following back this, in turn, increases the influence of the user although he may not be as influential as compared to other people.
- **Statuses count:** Follower statuses count (No of tweets) is used to find how active the user is on the twitter.
- **Activity measure:** It is simplest activity measure for measuring the different activity of the user. It is defined as

$$GeneralActivity(i) = T + L + RT \tag{1}$$

- **Topical Strength:** This feature is used to restrict the user attention to the tweets related to some specific topic. The formula for topical strength is defned as

$$TS(i) = \frac{T + RT + RP}{T} \tag{2}$$

- **Signal Strength:** It indicates how strong is the author's topical signal. It is defined as

$$SS(i) = \frac{T}{T + RT} \tag{3}$$

- **Follower rank:** It is the simpler popularity measure that uses follow-up relationship between users. It is a slight variation follower-followee ratio.The formula for follower rank is defined as

$$FollowerRank(i) = \frac{F1}{F1 + F2} \tag{4}$$

- **Popularity**: Popularity of an user is defined by its number of followers. It is defined as

$$Popularity(i) = 1 - e^{-\lambda.F1} \tag{5}$$

- **Degree centrality:** Degree centrality is the simplest centrality measure to compute.A node's degree is simply a count of how many social connections (i.e., edges) it has. The degree centrality for a node is simply its degree.

where T = number of tweets of a user
L = number of likes by the user
RT = number of retweets of a user
RP = number of replies
F1 = number of followers
F2 = number of followees

5.3 Retweet Detection

Retweet Detection in the tweeter network is an important task as it helps us to detect which tweet the user is going to retweet next related to the particular topic. For this, we have used the follower-followee retweet network related to a particular topic and along with that for a particular topic create one user-retweet network in which nodes will be user and tweets where all the outgoing edges are from user nodes and all the incoming edges are in tweet nodes. Toy example for the network is described in fig 1.

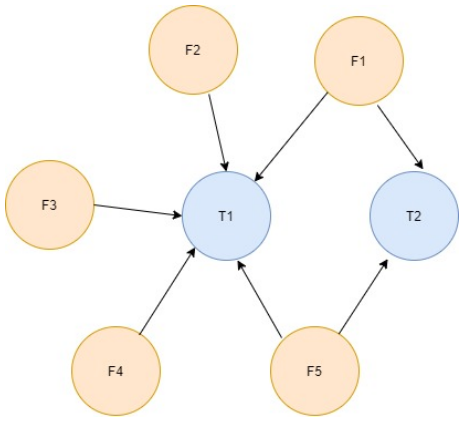

Figure 1: Tweet User network

For the task of the next possible retweet detection, some edges were removed from the retweet-user network. Now our task is to predict this which tweet will be retweeted next by the user. For these following steps need to be applied:

- Get all the previous tweets by the user from the retweet-user network. This is done by extracting neighbors for the user in this network.
- Get all the followee of the user from the follower-followee network. This is done by extracting neighbors for the user in this network.
- Now extract follower related data like follower count, friend count, likes, statuses count, etc.
- Extract data about the previous retweets by the user which includes language, sentiment, and the user id for the tweet.
- Now, iterate over all the possible next tweets (whose time is less than retweet time) and find the possibility using following features of tweet and user - most important is whether the tweet is done by one the followee and time of the tweet, other important features are languages and sentiment of the tweet and other not so important features includes follower count, statuses count, likes retweets, etc.

This does not give much information about the influencer in the network but helps to find the influence at the user level for a particular topic.

5.4 Influencer using Neural Network

In the above section, we have generated the features using the features generated by Twitter API. Using these features we will predict which users are influencers and which users are not influencers. So

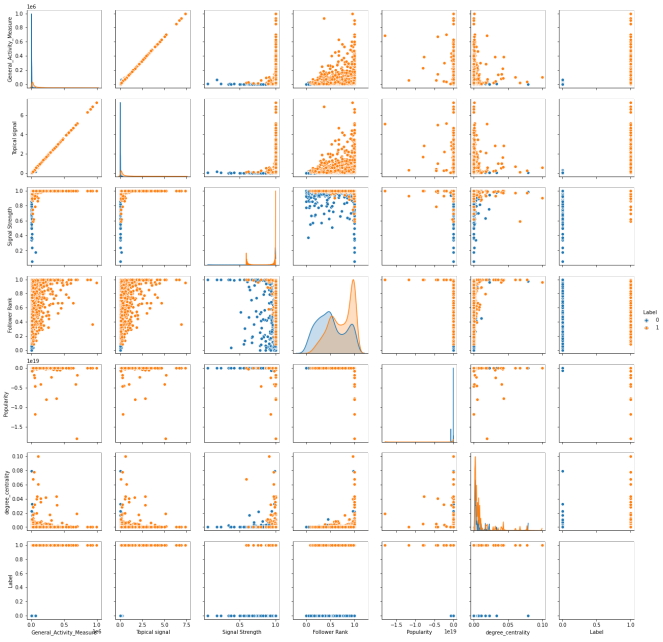

Figure 2: Correlation between different features of Neural Network

we use the non-linear model because it has a high capacity than linear models. So we choose the Neural Network for the task to find the influencers in the graph.

First, we selected those features which best describe the follower-followee relationship, user activity, and also the popularity of the user. For this, as we were lacking follower information for all the user we subsampled the graph to only those nodes which we had the followers information after this we had 3356 nodes in the graph. Now make data of unique followees with features like General Activity, Topical Strength, Popularity, Signal Strength, Follower Rank. Our approach is new so there is no ground truth as such. So to make balanced ground truth labels we put some restrictions on each feature for the user to be followers like the number of followers should be greater than 8000 and signal strength should be greater than 0.50, the topical signal should be greater than 0.05.

Now normalized the data and add make the labels of the data categorical. Divide the data for training (2349 users) and testing (1007 users). Now first add the input feature vector of size 8 with relu as the activation function. We have used some dropouts after each layer of the multilayer layer perceptron model which helps in reducing the overfitting of the model. Similarly, repeat the steps for the second and third layers. At last, evaluate the model with loss function binary cross-entropy and using adam optimizer.

5.5 Influencer using Graph neural network

For the task of finding the influencer in the network graph neural network performs better all the above-mentioned approaches because the graph neural network is able to capture the structural information better than any other previously defined measures. For this, as we were lacking follower information for all the user we subsampled the graph to only those nodes which we had the followers information after this we had around 3600 nodes in the graph. Here we are using semi-supervised learning as these graph data are generally large and graph learning seems to work fine for small annotated nodes as well. We have annotated around 200 nodes based on there features out of which some were taken as training data and others were taken as testing data.

Create a features vector as the one-hot encoding of the nodes this helps in finding the nearby nodes, and at the end of it append the features related to nodes such as the No of friends, Follower-Followee ratio, statuses count, retweet count, etc. as they are feature related to the nodes. If we have N nodes than the input feature matrix will be of size $N \times 6$, using the features of nearby nodes we generate current node representation as the summation of the features. Let the node intermediate representation be $\hat{u}$ and feature representation of nodes be u then,

$$\hat{u} = \sum_{v \in neighbour(u)} v$$

If we have A as the adjacency matrix and the feature matrix is $U = N \times 6$ than the output after this layer can be written as,

$$output = A \times U$$

The output will be multiplied with the weight matrix W and is passed through a relu function. So after this layer, the output of this layer $L1$ will be,

$$L^1 = R(A \times U \times W^1)$$

Similarly repeat this step with the second layer but here instead the feature matrix will be the one we generate from the first layer and the activation function will be softmax. So, after this the final output L2 will be,

$$L^2 = R(A \times L^1 \times W^2)$$

Weight are learn over the iteration with negative log loss function and adam is used as the optimizer function.

6 RESULTS

In this section, we are discussing the different results of all the models that are discussed above. Here ANN, MLP, and random forest are trained on different features and GCN is trained on different features. After the training of models on the training dataset, the model performance is evaluated on the testing dataset by using various metrics. Table 4 shows the performance of the models which are trying to classify the twitter user is an influencer or not.

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	90.52	90.05	90.18	90.13
ANN	94.44	94.04	93.90	93.98
MLP	96.34	96.12	96.20	96.25
GCN	93.33	96.22	81.81	86.92

Table 4: Different Models Performance Analysis

As show in the figure 4 Among Random Forest, ANN, and MLP models,MLP performs with a very high f1-score.Here GCN model is

trained on the different datasets with features consist of structural properties and twitter follower-followee relationship features.The GCN model F1-score is 86.92 which is quite good.

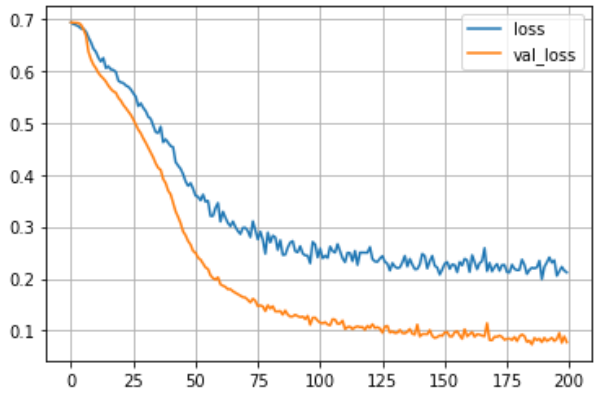

Figure 3: Epoch vs loss MLP

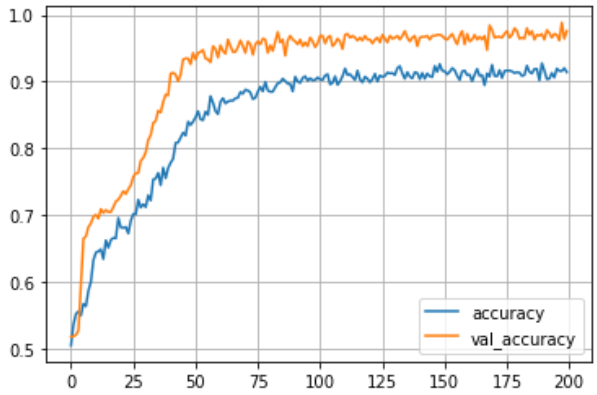

Figure 4: Epoch vs accuracy MLP

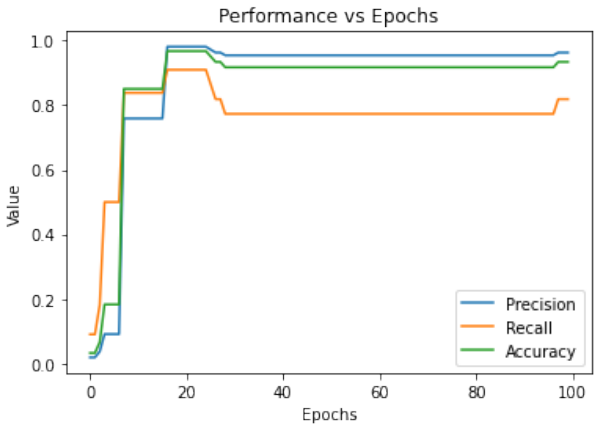

Figure 5: Epochs vs Peformance for GCN

In the figure 6 the confusion matrix for the influencer and non-influencer is shown. The main objective of our GCN model is to learn correctly the non-influencer which means not to classify the non-influencer as a influencer but vice-versa can be acceptable.This is clearly achieved as it is shown in the figure 6.

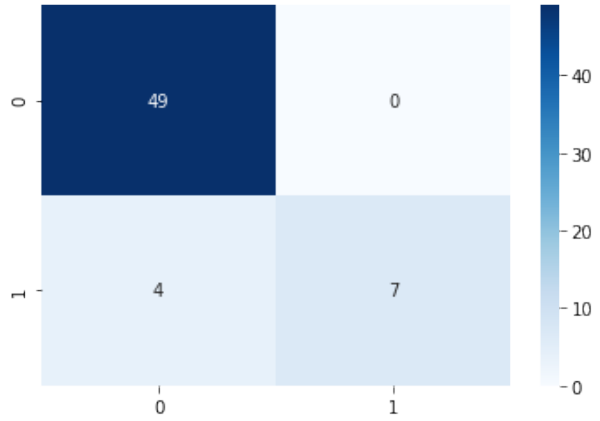

Figure 6: Confusion matrix for GCN

For the retweet prediction model accuracy of the model in predicting the exact next retweet is 72.35% and accuracy of the model decreased with the decreased in the training sample. Accuracy vs training sample graph is shown in figure 7. From this we can infer that this model performs well when the user data is large which is the case with most of the real life network.

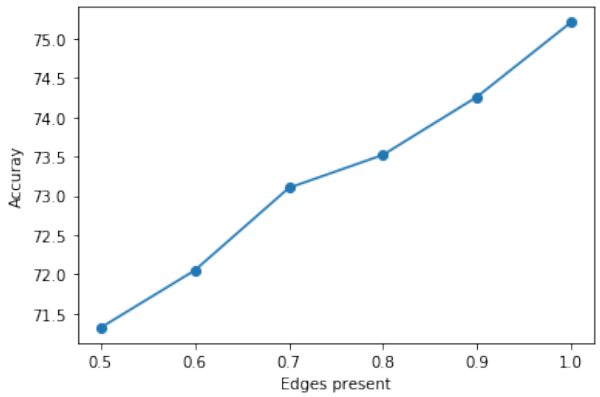

Figure 7: Test sample vs Accuracy

7 CONCLUSIONS AND FUTURE WORK

It is clear from the results shown in the table 4 that MLP works better than other models. But this is not the case despite that GCN has a low accuracy it performs better than other models in finding the influencer in the real-life network as it can capture the structure information in a much better way.

Models proposed above face class imbalance problem as the number of influencers are far less as compared to other users. This

was the reason due to which recall for influencer people was less. This can be improved by collecting more data or by using some sampling techniques.

Two features that are missing but could significantly improve the accuracy are checking whether the user is verified or not and second by acquiring follower information for all the nodes. Graph Sage and Graph attention networks can also be tried in the future to find the influencer in the graph.

REFERENCES

[1] Isabel Anger and Christian Kittl. 2011. Measuring influence on Twitter. *I-KNOW*, 31. https://doi.org/10.1145/2024288.2024326

[2] Fabián Riquelme and Pablo González-Cantergiani. 2016. Measuring user influence on Twitter: A survey. *Journal of Information Processing and Management* 52 (04 2016). https://doi.org/10.1016/j.ipm.2016.04.003

[3] Xin Zhang, Ding-Ding Han, Ruiqi Yang, and Ziqiao Zhang. 2017. Users' participation and social influence during information spreading on Twitter. *PLOS ONE* 12 (09 2017), e0183290. https://doi.org/10.1371/journal.pone.0183290