International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 19 (2018) pp. 14257-14266

A Semantic Graph Based Micromodel to Predict Message Propagation for Twitter Users

[1]Prasanth G Rao, Venkatesha M, [2]Anita Kanavalli and [3]P Deepa Shenoy, [4]Venugopal K R
[1]Visvesvaraya Technological University(VTU), Belagavi, Karnataka-590 018, India.
[2]Ramaiah Institute of Technology (RIT), Bangalore, Karnataka-560 054, India.
[3]University Visvesvaraya College of Engineering (UVCE), Bangalore, Karnataka 560 001, India.
[4]Bangalore University, Bengaluru, Karnataka 560056, India.

Abstract

Twitter is the most popular social platform for broadcasting opinions, but the reach of tweets is often nondeterministic. Recent years has witnessed numerous agencies misusing the platform for entrapment techniques, psychological manipulation and fake news campaigns compelling Social Media firms to enforce stricter data protection policies with limited access as the norm. This paper presents a micro-prediction model for determining message propagation for a user, especially for the non-influential majority. Our framework uses Ego network and Named Entity Recognition in predicting message propagation. The work focuses on determining the possible users who would interact and their immediate reach. This is achieved by using Twitter API in a limited manner. We attempt to make a responsive prediction model; simple, stateless and scalable, capable of catering to parallel requests. The simulation predicts with an accuracy of 85% for data constituting 336768 connected users.

Index Terms—Twitter, Ego Network, Named Entity Recognition, Message propagation, Rule-based classifiers

INTRODUCTION

Social Media is an integral part of human lives with an enormous contribution as a platform for information exchange and people networking. From influencing personality attributes and trust factors, Social networks are known to define the very idea of social capital [1]. People with good reputation on networking sites stand a better probability to enriching opportunities, both of professional and personal nature. The impact of social web mining on Sociology, Governance, and Economics is noted and well acknowledged. Corporates invest considerable resources in devising a valuable and resourceful narrative for its current and prospective customers on social platforms enabling them to target for better profits and constructive market engagements. The recent example is the Ice Bucket Challenge. A simple act of dropping a bucket of ice water on a person with the idea to promote awareness about Amyotrophic Lateral Sclerosis(ALS) became a global phenomenon and managed to generate research funding worth millions of dollars. Psephologists are successfully predicting election results from sentiment analysis of the messages shared by the citizens.

Twitter is the most popular microblogging service with 330 million active users sharing 500 million *tweets* or messages each day. Table I details the recent usage statistics [2]. Each message can be 280 characters long and supports multimedia content. Apart from sharing opinions, users can tweet to any other public users using their address called a handle beginning with @ symbol. Users can endorse any tweets by sharing it (also known as retweeting) or liking a tweet using the favorite feature. User relations are formed by the means of *following* and *friends*. Let us say Jack is a Twitter user and Jill follows him. That makes Jill a friend of Jack. For Jack, Jill becomes a follower. Jill being Jack's follower will receive all his updates. Jack would receive Jill's tweets only if he chooses to follow her. Thereby it might be clear now that the social graph formed in case of Twitter is a directed graph.

Usage of Twitter in news media is a common sight. 83% of the world leaders have a Twitter account and actively engage with their followers on a daily basis. The participation of news media and government agencies gives the common citizen a medium to connect and converse. Despite the widespread usage, reality can be quite different. Communications today are largely one-sided. Upon comparing an average user to an influential user, differences begin to emerge. Further bots and rouge users acting as spammers, ideological extremists, etc. use the service for damaging intentions. The lack of effective and proactive surveillance systems allows such activities to go unnoticed. People in distress very often reach to Twitter for help however their tweets often go unnoticed. In a system with 40% of the data classified as pointless babble, it becomes impossible to determine the impacted users for the given message.

Table I. Twitter usage statistics

Attribute	Count
Total Users	1.3 billion
Active Users	550 million
Monthly active users	330 million
Verified users	293,027
Bots	23 million
Accounts without followers	391 million
Tweets per day	500 million
Average followers per user	707

This paper is organized as follows. Section II outlines the significant contributions made to the field thus far. Section III introduces the problem and describes the objectives of our work. Section IV details out the solution with insights into implementation. Section V presents the outcome of our execution. Concluding section VI briefly discusses the next steps.

RELATED WORK

A. Small World Phenomenon

Milgram [3] presents the notion of Small World Phenomenon, also known as the Six Degrees of Separation. Milgram created an experiment to determine the shortest paths of acquaintances required to reach each other. The experiment required delivering a letter addressed to a person in Boston routed via acquaintances. Milgram found the letter changed hands six times to reach the target resulting in two important discoveries. Firstly, it established the existence of a short path between otherwise unrelated people. Secondly, it showed people collaborating within their independent capabilities could deliver the letter outside their immediate social circle. The work went on to define many significant research streams.

Jon Kleinberg [4] presents a decentralized algorithmic interpretation of the small world phenomenon. Kleinberg's model for the small world phenomenon is a *k*-dimensional matrix of nearest-neighbors. The distance measure is defined between points in the matrix *x* and *y* as $d(x,y)^{-k}$ Equation 1 gives the Probability *p* of the shortest routing path.

$$p(x \leftrightarrow y) = \frac{d(x,y)^{-k}}{H_k(n)} \qquad (1)$$

$H_k(n)$ is a normalization constant.

Robert E Hiromoto [5] further explores the Kleinberg's model with random graphs in Neuroanatomical networks elaborating on the complexities and concerns around parallelism. He explores data communication schemes over different topologies to discover the algorithm succeed to avoid random uncertainties to a good extent. The concept of small world phenomenon is extensively applied to Social Networks. Facebook research [6] found the mean degree of separation to be at 3.57 for 1.59 billion active users. Masaru Watanabe et al. [7] deduce the mean degree of separation for twitter at 4.59. Noticeably both the numbers are much lesser than 6 signifying the strong interlinking among users.

B. User Graph

Much work has gone into the field of user characterization and discovering user networks. Hughes et al. [8] define the Big-Five personality predictors for social networking sites. The Big-Five consists of five broad personality traits, namely, Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness. Natural graphs, such as social networks, email graphs, or instant messaging patterns, have become pervasive through the Internet. These graphs are massive, often containing millions of nodes with billions of edges. Ahmed et al. [9] address the issue of factorizing such natural graphs using vertex partitioning algorithms. The algorithms are developed using distributed methods. The partitioned graphs have vertices labeled *owned* and *borrowed*. The *borrowed* vertices are shared between graphs and used for convergence and completeness. Ahmed et al. [10] extend this further with regional context. The framework detects regional contexts from regional models and language models and identifies the geographic locations for the information shared on microblogging services. Lin et al. [11] has done extensive work to address the problems of extracting and analyzing communities, but the factors that drive their formation are still not well understood. Papadopoulos et al. [12] have defined explicit and implicit communities and have discussed the strategies for Scalability. Roth et al. [13] suggest friends with implicit graph based on email exchanges. Graphs are constructed based on users addressed in email exchanges, together with weighting functions for edge priorities, implicit graphs are derived.

C. Ego Networks

Ego networks is a micro-graph outlining a person(*ego*) and his interactions with the other people(*alter*) in his neighborhood. Figure 1 shows the representation (also called Dunbar's circles) as put forward by Arnaboldi and his colleagues [14]. The innermost circle, *support clique*, symbolizes the strongest social relationships. Outer circles, represented with a progressing larger diameter or population with proportionally reducing intimacy, are respectively called *sympathy group*, *affinity group*, and *active network*. The work establishes the potential of Ego networks in learning the cognitive properties that define human relations in the real world. McAuley and Leskovec [15] use Ego networks to learn implicit social circles on Twitter. The circles are defined based on features such as hometowns, birthdays, colleagues, political affiliations, etc.

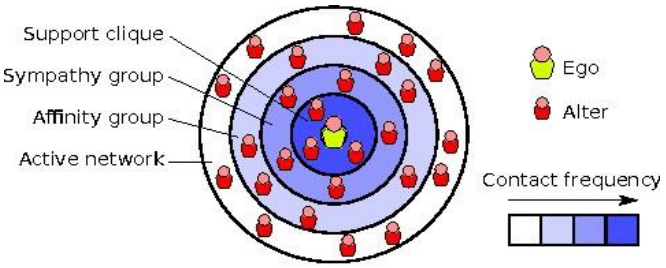


Figure 1. Ego Network Model

D. Semantic Analysis and Modeling

Given the nature of Twitter posts with the informal language, semantic analysis to find sentiment associated with tweet can help us deduce contexts about those posts. In [16], authors proposed experiments using semantic tools such as DBPedia, WordNet, and SentiWordNet for training classifiers for Twitter messages. Results shown in this paper are based on SVM model and Naive-Bayes approach for the different features extracted for a given Tweet. Here f-measure was obtained to classify tweets based on its sentiment positivity. Similarly, Ren et. al. in [17] used a neural network-based model for classifying tweets to its sentiments. Results are given for both balanced and unbalanced neural network taking considerations for local

features against context-based network. This again emphasizes the importance of context information to obtain better results for classification. In [18], we can see a prediction model developed based on Tweets. Idea is to predict box office revenues using the tweets based on the rate of similar posts. A linear regression model has been proposed here which enhances the prediction results when the extracted sentiments fed into the model.

Graph-based approaches are also tried and tested for developing recommender systems using Knowledge Graphs (KG) which were built leveraging the semantic knowledge sources. Ren et. al. [17] proposed a movie recommender system making use of KGs and built a probabilistic logic model called *ProPPR*. Proportionate density of dataset amongst the graph plays a major role. Karidi et al. [19] demonstrated a system to recommend both tweet and *followee* based on a KG build using Alchemy API Taxonomy service. User tweets are then sampled against the KG to find the

Topics of Interests (ToIs) and to build respective user profiles, i.e. a subset of KG and obtained using Steiner Tree. Both cases demonstrate the capabilities of graph-based approaches.

E. Predicting retweets

Retweets can be predicted by analyzing statuses. Paper [20] handles streaming prediction of retweets using time-sensitive models. Authors establish predicting retweets are possible with 202 tweets predicted by two subjects. The first subject predicts with only the text of tweets with an accuracy of 76.2%. The second subject predicts with an accuracy of 73.8% tweets containing social circle information. Finally proposed unsupervised model predicts with an accuracy of 69.3% and supervised model with 82.7%. Chenhao Tan et al. [21] use Topic and Author-Controlled(TAC) pairs to predict retweets based on the wording used in the tweets.

The works discussed so far have focused on specialized problems with abundant data at disposal. However, the datasets used in the tests are not available in the public domain making comparisons with existing work and drawing parallels difficult. Concerns surrounding privacy violations, data misuse and accusations of content profiling for illegitimate causes are the reasons which made social data much less accessible. All social networking sites including Twitter enforce strict security measures and publishes limited data via highly monitored API implementations. Much research goes in identifying influential users in blogging [22], [23] and microblogging forums. Twitter has 239K verified users against 330 million accounts [24] i.e., less than 0.001% for all the users. With limited data on disposal, qualitative research in social media should be capable of determining actionable tweets and users among a heap of forwards, spams and other tweets employing light and stateless algorithms. Our work is an attempt in this direction. Therefore, we need micro prediction models which work with limited data while providing comparable results. We present one such micromodel using Ego networks and Named Entity Recognition to predict user retweets. The approach is expected to give us shorter computing times with similar results.

PROBLEM DEFINITION

Given a Twitter user, devise a micro-prediction model working with limited data calls to Twitter API. The objective of the model is as follows.

1) Plot possible reduced Ego network for message propagation
2) Discover topic interests and associated sentiment forUsers
3) Predict retweet probability for the user

The Twitter APIs used in the work are listed with their respective rate limits in Table II. The required data for prediction is fetched in a single window. The data required to validate our prediction is scanned from the next levels of followers. To optimize data calls, we restrict our analysis to the overall graph structure and observable user history. Features such #hashtags are removed from the present scope of prediction. Tweets in the English Language are only considered in the present work. The micromodel framework will be scalable and stateless to ensure the framework can run at any number of parallel instance with any volume of data.

Table II. Rate limits per 15 minutes window

Endpoint	Resource family	Requests / window (user auth)
GET followers/list	followers	15
GET users/lookup	users	900
GET statuses/lookup	statuses	900
OVERALL	ALL	900

METHODOLOGY

Figure 2 outlines the design of the micromodel prediction framework implemented. There are six components as seen.

- ***Twitter Graph Crawler***: Collects the user and friends information with breadth-first traversing. The approach ensures data is read as an Ego network with relevant neighborhood information. We read users till Level 2, with level 0 being the Ego user.
- ***User Status Analysis*:** Derive statistical measures for the tweets/statuses for each user in the graph. Compute average following and friends for followers, turnaround time per tweet and retweets per tweet. Follower information serves as inputs to the prediction model, and remaining data is used for output validation.
- ***Ego Network Reduction***: Determine users for which Ego network for level larger than 1 exist and generate weighted Ego network. Prune the connections based on turnaround time and retweet probability. The approach is retrospective.
- ***Named Entity Categorization***: Determines topic sets in each tweet using Named Entity Recognition.

- ***SentiScore Evaluation*:** For every Named Entity identified, a sentiment score is generated using SentiWordNet 3.0 [25] and average sentiment score is determined for each user.
- ***Message Propagation Prediction***: Enriched datasets consisting of Named Entities Annotations and Sentiment score is subjected to classifications algorithms for predicting retweets.

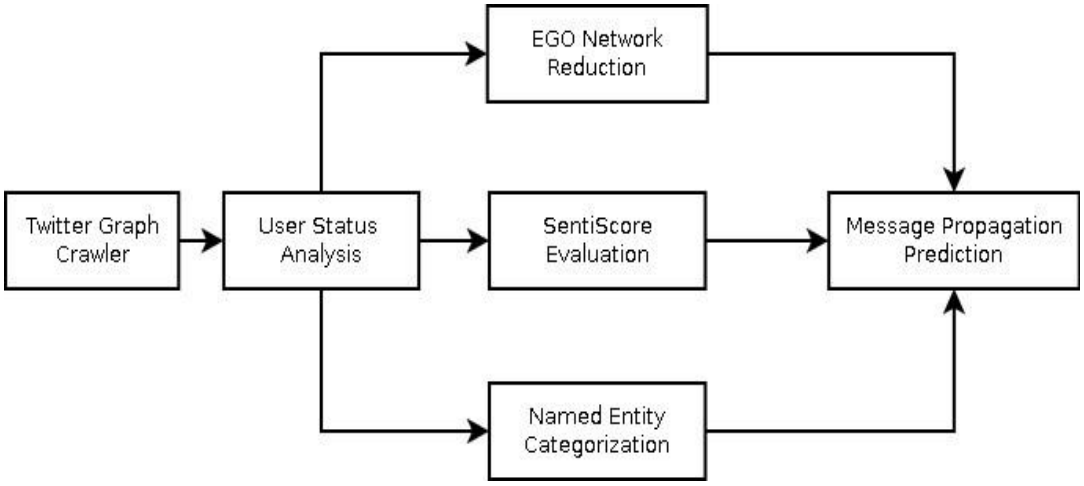


Figure 2. Architecture of Message Propagation Predictor

*A. **Data Sampling and Graph Definition***

The data is generated in a linear order to accommodate rule-based classifiers like Decision Trees and Decision Tables. Rule-based classifiers are lightweight algorithms having relatively lesser learning times and are quick to model making it a suitable candidate for our micromodel framework. The delay in processing times is attributed to the rate limits enforced by Twitter. Twitter API response provides the friends, followers, status count directly. Subsequent attributes such as the simple average of followers and friends are deduced by crawler along the follower references. Simple average, *x*, is defined by Equation 2.

$$\overline{x} = \frac{\sum_1^n x_i}{n} \tag{2}$$

Sampling size of statuses per user is set to the recent 10(= *n*) tweets. Average tweet rate of the user, *d*, is defined with the interval for create date between the tweets, as shown in Equation 3.

$$\overline{d} = \frac{\sum_1^{n-1} (d_{i+1} - di)}{n-1} \tag{3}$$

We define the Ego networks as a graph *G*(*V*,*E*), where *V* is the weighted node defined as *V* (*d*,*retweet*). *E* denotes *following* relations. The graph is a temporal graph with the timescale defined by *d*. At any given distinct d^0, the set of $V'(d > d^0)$ represented the probable users to respond. *V'*/*V* gives the probable audience in the interval.

*B. **Classifiers in Message Propagation Prediction***

We use J4.8 classifier (an improvement over C4.5 decision tree), Random Forest and Decision Table for predicting retweets. C4.5 [26] works towards minimizing Information

Entropy *H*(*T*) while maximizing Information Gain *IG*(*T*,*a*).

$$H(x) = -K \sum_{i=1}^{n} P(x_i) \log P(x_i) \tag{4}$$

$$IG(T,a) = H(T) - \sum_{v \in vals(a)} \frac{\{x \in T | x_a = v\}}{|T|} \cdot H(\{x \in T | x_a = v\}) \tag{5}$$

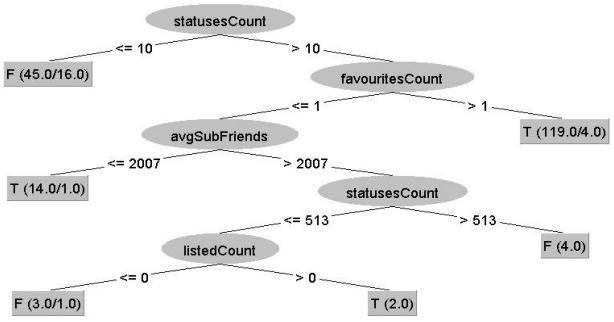
P(*x*) is the Probability function. *T* denotes the training set *T*(*x*,*c*) = (x_0,x_1,...,x_n,*c*) with *x* the attributes and *c* the class label.

Random Forest [27] or Random Decision Forest is an ensemble of decision trees predicting the outcomes by majority voting of the likely output. Random Forest algorithm is capable of both classification and regression.

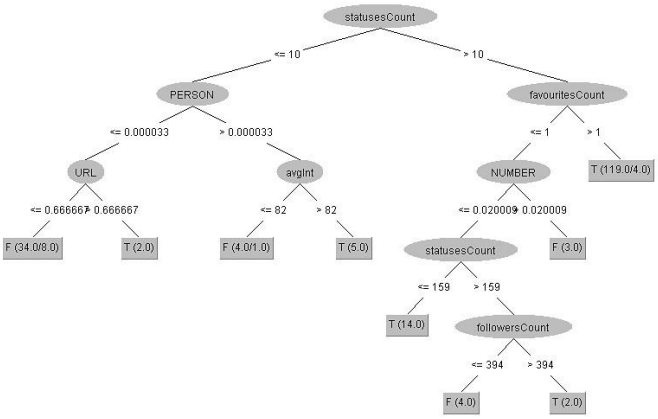
Decision Table Majority(DTM) [28] is a rule-based classifier which builds hypothesis based on an induction algorithm. DTM has two components. A set of features called *scheme* and a multiset of labeled instances called *body*. Given a target function *f* and a hypothesis class H, we define the optimal features to be the features used in a hypothesis *h* in H that has the highest future prediction accuracy with respect to *f*. Error of a hypothesis *h* using an independent test set τ is defined as *err*(*h*,τ) in Equation 6. c

$$\widehat{err}(h,\tau) = \frac{1}{\tau} \sum_{(x_i,y_i) \in \tau} L(h(x_i), y_i)) \tag{6}$$

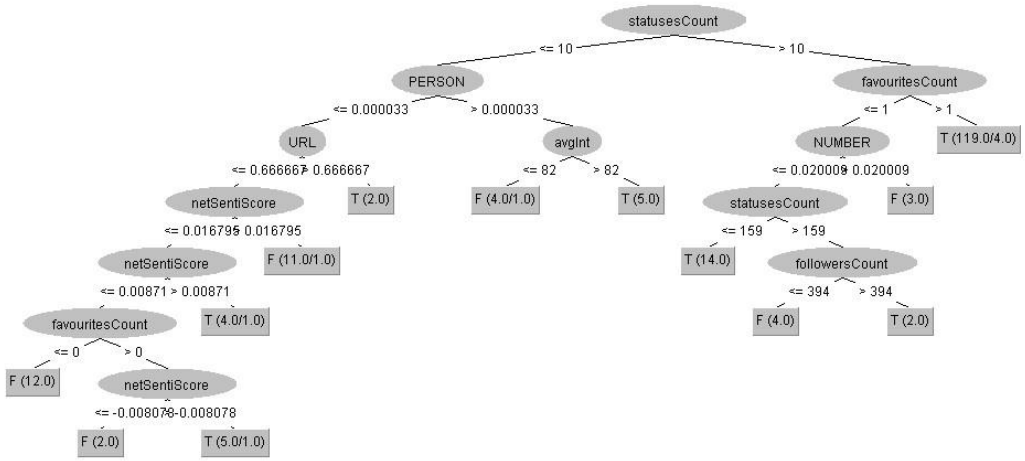
where *L* is a loss function. *h*(*x*) is zero-one loss function, i.e., zero if *h*(*x*) = *y* and one otherwise. The approximate accuracy is defined as 1 − *err*(*h*,τ). The objective of DTM is to determine an optimal feature subset, A^*, for a given hypothesis space H and target function *f* such that there exists a hypothesis *h* in H using only features in A^* and having the lowest possible error with respect to the target function *f*.



(a) *Dataset : GD*



(b) *Dataset : GDNE*



(c) *Dataset : GDNESS*

Figure 3. J4.8 Decision Tree Output for Datasets

*C. **Semantic Analysis***

Opinion Mining or Sentiment Mining involves the use of Natural Language Processing(NLP) in estimating subjective parameters or opinion about the given context. The authors of SentiWordNet [29] outline three broad distinctions necessary to deterministically tag an opinion to a text.

1) ***Determining Subjective-Objective polarity*** where distinction is made between fact(Objective) or an opinion(Subjective).

2) ***Determining Positive-Negative polarity*** where distinction is made if *Subjective* word presents a *Positive* or a *Negative* opinion.

3) ***Determining the strength of Positive-Negative polarity*** where a measure for the degree of *positivity* or *negativity* of the opinion is calculated.

Named Entity Recognition(NER) is a domain in NLP for categorizing Nouns or Named Entities into predefined classes such as people, location, time etc. Classification scheme helps us better tag the datasets in our subsequent analysis.

A total of 336768 connected users were analyzed, 243 users had Ego network available till level 2. To avoid human intervention, we use level 2 nodes to determine the expected classifier output.

RESULT

Dataset generation is implemented in Java threads and Mongo DB for data staging. Weka [30] is used to implement classifier. Graph analysis is done using Gephi [31]. NER is done using Stanford CoreNLP toolkit [32].

A. Dataset

The data required for our tests was downloaded using a custom

graph crawler polling the Twitter API. The crawler executed for 5 days collecting approximately 40GB data comprising of users and tweets. Dataset details are outlined in Table IV. To optimize the API invocations, the below list of assumptions are made.

- Users with more than 5000 followers are considered local influencer and excluded from further probing.
- Users following over 5000 accounts are excluded as their likelihood to respond to tweets are less.
- Accounts marked private are excluded.

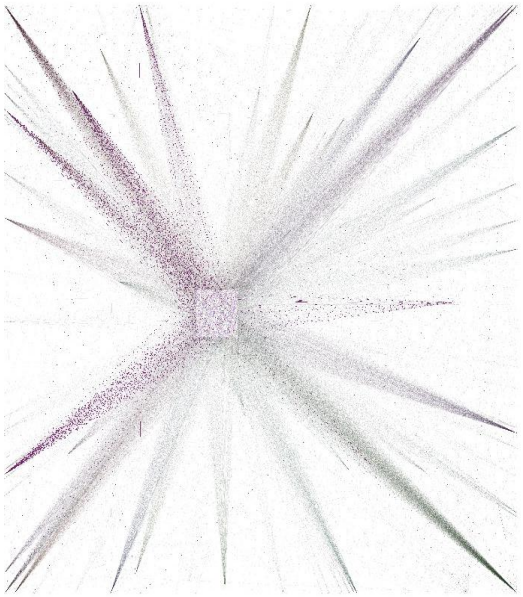
TABLE III. Annotations spread in *GDNE* and *GDNESS*

Annotation	Min	Max	Standard Deviation	Mean
LOCATION	0	0.50000	0.04972	0.00672
NATIONALITY	0	0.16667	0.01478	0.00196
NUMBER	0	1.00000	0.19572	0.07259
IDEOLOGY	0	0.00000	0.00000	0.00000
MONEY	0	0.17341	0.01496	0.00179
PERSON	0	1.00000	0.13125	0.03750
SET	0	0.02042	0.00223	0.00026
MISC	0	0.33333	0.03233	0.00466
TIME	0	0.00000	0.00000	0.00000
ORDINAL	0	0.50000	0.06420	0.01174
EMAIL	0	0.00000	0.00000	0.00000
CAUSE OF DEATH	0	0.10000	0.00894	0.00128
URL	0	1.00000	0.16107	0.05858
O	0	1.00000	0.27727	0.13745
STATE OR PROVINCE	0	0.02041	0.00241	0.00031
ORGANIZATION	0	0.33333	0.04604	0.01448
DATE	0	0.50000	0.07952	0.02461
CITY	0	0.10000	0.00975	0.00136
COUNTRY	0	0.50000	0.04945	0.00727
RELIGION	0	0.02083	0.00161	0.00013
PERCENT	0	0.02042	0.00166	0.00017
TITLE	0	0.16667	0.01980	0.00469
CRIMINAL CHARGE	0	0.02000	0.00158	0.00013
DURATION	0	0.50000	0.06235	0.01311

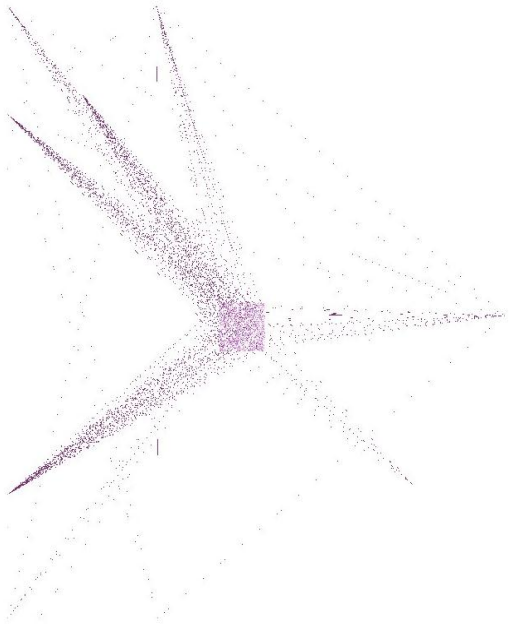
Upon constructing the Ego networks, we are left with 243 users having relations with about 33600 users and 395504 connections. For the simplicity of our analysis, we have organized the datasets in three configurations, each with increasing contextual information.

- ***Graph Data(GD)*:** Dataset constitutes of the statistical observed averages of user and their respective followers along the EGO Network. Description of attributes is detailed in Table V.
- ***Graph Data + NamedEntity(GDNE)*:** Dataset *GD* with the observed probability of occurrence each NER annotation in the tweets. Observed probability of occurrences value ranges between 0 and 1. Table III outlines the NER annotations discovered with their distribution statistics.

• ***Graph Data + Named Entity + Sentiment Score (GDNESS): Dataset GDNE:*** with average sentiment score attribute,*netSentiScore*, of tweets computed as the simple average defined in Equation 2. Sentiment score is computed for every Named Entity tagged.



(a) Ego Network for user prasanthgrao. Level 2. 24635 users



(b) Ego Network after pruning. 2137 users

Figure 4. Ego Network for user *@prasanthgrao*

Our prior work [33] predicts retweets for GD dataset using Decision tree algorithms J4.8 and Random Forest. We further the previous implementation with use of DTM with semantic data. The results will be analyzed for each of the dataset combinations in the next subsections.

Table IV. Input Twitter Data

Entity	Total Instances
Users(V)	336768
Relations(E)	395504
Statuses	264298

B. Ego Networks

Figure 4 shows the Ego networks for user *@prasanthgrao*. Figure 4a shows all the users following *@prasanthgrao* till level 2. The distance of any node from center is inversely proportional to the *retweet*. Figure 4b shows pruning outcome for $d = 1day$. Total users are seen reduced to 10%. The number of edges or *following* reduces to 1.86%, which also denotes the probability of retweet. The result confirms our assumption that Ego network of depth 1 provides the necessary information for our prediction model.

C. Retweet Prediction

J4.8 decision tree outcome for each dataset is shown in figure 3. For dataset *GD*, apart from the *statusesCount* and *favouritesCount*, we see the attribute *avgSubFriends* in decision tree signifying the importance of connected followers in getting retweets. With *GDNE*, we begin to see the importance of topics in determining the retweets. Tweets about *PERSON* and *NUMBER* are more decisive in determining the retweet probability. Moving to dataset *GDNESS*, sentiment starts to take precedence as we see the *netSentiScore* occurring three times in the entire decision tree as seen in figure 3c. Further, we observe sentiment playing a role in case pf *PERSON*(Subjective) while being absent with *NUMBERS*(Objective). The observation, we believe, is a reflection of the real-world behavior.

Table VI captures the classifier performance for *GD* dataset. Overall prediction accuracy with J4.8 is 80.7% and success rate for predicting actual retweets is 86.4%. The algorithm does not particularly do well for negative cases with an accuracy of 60%. Random Forest fares better with 82.7% but is over-trained to identify only positive retweets. For users with no retweets, the algorithm has accuracy of only 47.5%. Decision Table with 81% accuracy offers a balanced prediction model.

With NER, we see Decision Table significantly outperforms the other algorithms at 85% accuracy against 79% and 80.3% of J4.8 and Random Forest respectively. The results for both *GDNE* and *GDNESS* datasets are comparable as seen Table VII and Table VIII respectively.

CONCLUSION AND FUTURE WORK

The present work predicts the probability of a user getting retweets with an accuracy of 85% and identifies the possible paths for the interactions with limited data. The viability of micromodel is established with this work. We have only considered observable statistical characteristics. The restrictions of Twitter APIs which allow only a maximum of 900 requests every 15 minutes presents the required dataset for validation with a substantial time lag. Twitter does provide streaming APIs for real-time entries however; the API is not customizable and the response is a randomly sampled output of the current activity. The necessity, therefore, is of a heuristic self-learning solution centered around Fuzzy or Rough set approach can help us make a real-time streaming based micro-prediction model working with a highly limited data source.

Table V. Dataset Description

Attribute	Data Type	Description
screenName	String	Twitter handle/account
followersCount	long	Accounts following user
friendsCount	long	Accounts followed by user
statusesCount	long	Tweets published by user
favouritesCount	long	Total favorites
listedCount	long	Number of lists the user features in
tweetInternal	long	Average interval between tweets
avgSubFollowers	long	Average of accounts following user's followers
avgSubFriends	long	Average of accounts followed by user's followers
hasRetweets	boolean	Expected output

Table VI. Detailed Accuracy for Classifiers. *Dataset : GD*

Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
J4.8	0.864	0.400	0.888	0.864	0.876	0.448	0.643	0.816	T
	0.600	0.136	0.545	0.600	0.571	0.448	0.643	0.366	F
	0.807	0.344	0.815	0.807	0.811	0.448	0.643	0.720	Overall
Random Forest	0.918	0.525	0.865	0.918	0.891	0.434	0.856	0.949	T
	0.475	0.082	0.613	0.475	0.535	0.434	0.856	0.561	F
	0.824	0.430	0.811	0.824	0.815	0.434	0.856	0.866	Overall
Decision Table	0.891	0.450	0.879	0.891	0.885	0.450	0.803	0.919	T
	0.550	0.109	0.579	0.550	0.564	0.450	0.803	0.469	F
	0.818	0.377	0.815	0.818	0.816	0.450	0.803	0.822	Overall

Table VII. Detailed Accuracy for Classifiers. *Dataset : GDNE*

Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.871	0.450	0.877	0.871	0.874	0.417	0.698	0.843	T
J4.8	0.550	0.129	0.537	0.550	0.543	0.417	0.698	0.479	F
	0.802	0.381	0.804	0.802	0.803	0.417	0.698	0.765	Overall
	0.918	0.550	0.860	0.918	0.888	0.412	0.866	0.949	T
Random Forest	0.450	0.082	0.600	0.450	0.514	0.412	0.866	0.604	F
	0.818	0.450	0.804	0.818	0.808	0.412	0.866	0.875	Overall
	0.932	0.450	0.884	0.932	0.907	0.525	0.811	0.919	T
Decision Table	0.550	0.068	0.688	0.550	0.611	0.525	0.811	0.586	F
	0.850	0.368	0.842	0.850	0.844	0.525	0.811	0.848	Overall

Table VIII. Detailed Accuracy for Classifiers. *Dataset : GDNESS*

Algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.864	0.450	0.876	0.864	0.870	0.407	0.693	0.840	T
J4.8	0.550	0.136	0.524	0.550	0.537	0.407	0.693	0.471	F
	0.797	0.383	0.801	0.797	0.799	0.407	0.693	0.761	Overall
	0.905	0.575	0.853	0.905	0.878	0.364	0.853	0.946	T
Random Forest	0.425	0.095	0.548	0.425	0.479	0.364	0.853	0.563	F
	0.802	0.472	0.787	0.802	0.793	0.364	0.853	0.864	Overall
	0.932	0.450	0.884	0.932	0.907	0.525	0.811	0.919	T
Decision Table	0.550	0.068	0.688	0.550	0.611	0.525	0.811	0.586	F
	0.850	0.368	0.842	0.850	0.844	0.525	0.811	0.848	Overall

REFERENCES

[1] S. Valenzuela, N. Park, and K. F. Kee, "Is there social capital in a social network site? facebook use and college students' life satisfaction, trust, and participation," *Journal of computer-mediated communication*, vol. 14, no. 4, pp. 875–901, 2009.

[2] K. Smith. (2017) 44 incredible and interesting twitter statistics.[Online]. Available: https://www.brandwatch.com/blog/44-twitter-stats/

[3] J. Travers and S. Milgram, "The small world problem," *Psychology Today*, vol. 1, no. 1, pp. 61–67, 1967.

[4] J. Kleinberg, "The small-world phenomenon: An algorithmic perspective," in *Proceedings of the 32nd Annual ACM symposium on Theory of computing*. ACM, 2000, pp. 163–170.

[5] R. E. Hiromoto, "Parallelism and complexity of a small-world network model," *International Journal of Computing*, vol. 15, no. 2, pp. 72–83, 2016.

[6] S. Bhagat, M. Burke, C. Diuk, I. O. Filiz, and S. Edunov.(2016) Three and a half degrees of separation. [Online]. Available: https://research.fb.com/three-and-a-half-degrees-of-separation/

[7] M. Watanabe and T. Suzumura, "How social network is evolving? a preliminary study on billion-scale twitter network," in *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 2013, pp. 531–534.

[8] D. J. Hughes, M. Rowe, M. Batey, and A. Lee, "A tale of two sites: Twitter vs. facebook and the personality predictors of social media usage," *Computers in Human Behavior*, vol. 28, no. 2, pp. 561–569, 2012.

[9] A. Ahmed, N. Shervashidze, S. Narayanamurthy, V. Josifovski, and A. J. Smola, "Distributed large-scale natural graph factorization," in *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 2013, pp. 37–48.

[10] A. Ahmed, L. Hong, and A. J. Smola, "Hierarchical geographical modeling of user locations from social media posts," in *Proceedings of the 22nd International conference on World Wide Web*. ACM, 2013, pp. 25–36.

[11] Y.-R. Lin, J. Sun, H. Sundaram, A. Kelliher, P. Castro, and R. Konuru, "Community discovery via metagraph factorization," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 5, no. 3, p. 17, 2011.

[12] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos, "Community detection in social media," *Data Mining and Knowledge Discovery*, vol. 24, no. 3, pp. 515–554, 2012.

[13] M. Roth, A. Ben-David, D. Deutscher, G. Flysher, I. Horn, A. Leichtberg, N. Leiser, Y. Matias, and R. Merom, "Suggesting friends using the implicit social graph," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2010, pp. 233–242.

[14] V. Arnaboldi, M. Conti, A. Passarella, and F. Pezzoni, "Ego networks in twitter: an experimental analysis," in *IEEE International Conference on Computer Communications Workshops*. IEEE, 2013, pp. 229–234.

[15] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 539–547.

[16] H. Hamdan, F. Bechet, and P. Bellot, "Experiments with dbpedia,´ wordnet and sentiwordnet as resources for sentiment analysis in microblogging," in *SemEval@NAACL-HLT*, 2013.

[17] Y. Ren, Y. Zhang, M. Zhang, and D. Ji, "Context-sensitive twitter sentiment classification using neural network," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI'16, 2016, pp. 215–221.

[18] S. Asur and B. A. Huberman, "Predicting the future with social media," in *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 1, 2010, pp. 492–499.

[19] D. Pla Karidi, Y. Stavrakas, and Y. Vassiliou, "Tweet and followee personalized recommendations based on knowledge graphs," in *Journal of Ambient Intelligence and Humanized Computing*, 04 2017.

[20] S. Petrovic, M. Osborne, and V. Lavrenko, "RT to win! predicting message propagation in twitter." *ICWSM*, vol. 11, pp. 586–589, 2011.

[21] C. Tan, L. Lee, and B. Pang, "The effect of wording on message propagation: Topic-and author-controlled natural experiments on twitter," *arXiv preprint arXiv:1405.1438*, 2014.

[22] G. Vasanthakumar, R. Priyanka, K. V. Raj, S. Bhavani, B. A. Rani,P. D. Shenoy, and K. Venugopal, "PTMIB: Profiling top most influential blogger using content based data mining approach," in *International Conference on Data Science and Engineering (ICDSE)*. IEEE, 2016, pp. 1–6.

[23] G. Vasanthakumar, P. D. Shenoy, and K. Venugopal, "PFU: Profiling forum users in online social networks, a knowledge driven data mining approach," in *IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*. IEEE, 2015, pp. 57–60.

[24] Twitter. (2018) Twitter verified. [Online]. Available: https://twitter.com/verified/following

[25] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." in *Lrec*, vol. 10, no. 2010, 2010, pp. 2200–2204.

[26] J. R. Quinlan, "C4. 5: programs for machine learning," 2014.

[27] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[28] R. Kohavi, "The power of decision tables," in *European conference on machine learning*. Springer, 1995, pp. 174–189.

[29] E. Cambria, R. Speer, C. Havasi, and A. Hussain, "Senticnet: A publicly available semantic resource for opinion mining." in *AAAI fall symposium: commonsense knowledge*, vol. 10, no. 0, 2010.

[30] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P.

Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[31] M. Bastian, S. Heymann, M. Jacomy *et al.*, "Gephi: an open source software for exploring and manipulating networks," *ICWSM*, vol. 8, pp. 361–362, 2009.

[32] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55–60.

[33] P. G. Rao, Venkatesha, A. Kanavalli, P. D. Shenoy, and K. Venugopal, "A micromodel to predict message propagation for twitter users," in *Proceedings of the International Conference on Data Science and Engineering*. IEEE, 2018.