User Connectivity and Event Popularity Based Re-Tweet Prediction in Social Networks

Yadala Sucharitha, Y Vijayalata, V Kamakshi Prasad

Abstract: In recent times, social network services speedup the information proliferation among user groups, leaving the customary media such as newspaper, TV, discussion, web journals, and online interfaces far behind. Different messages are spread rapidly and broadly by re-tweeting in micro-blogs. Foreseeing re-tweet behavior is incredibly challenging because of different reasons. Existing forecasting models basically overlook sociological information and they don't acquire complete benefit of these emerging factors, influencing the performance of anticipation. In addition, the poorness of re-tweet data also seriously upsets the performance of these approaches. In this article, we take Sina micro-blog for instance, intending to anticipate the probable quantity of re-tweets of an original tweet as per the time series dispersion of its top n re-tweets. So as to deal with the above issue, we present the idea of a tweet life-cycle, which is essentially calculated by three parameters called the reaction-time, content-significance, interim-time circulation, and afterward the given time series dispersion arch of its top n re-tweets is fitted by a two-stage function, in order to foresee the quantity of its re-tweets in specific time period. The stages in the function are partitioned by the life-cycle of the original tweet and various functions are utilized in the two stages. We have assessed our methodology on real-world data-set; moreover contrast our outcomes with baseline methods. Our examinations prove that the proposed methodology can precisely anticipate the quantity of future re-tweets for a particular tweet.

Keywords: Re-tweet prediction, social media, micro-blogs.

I. INTRODUCTION

Online micro-blogs, for example Twitter& Face-book have turned into hugely well-liked in modern time. These services are system composition framework created by communication between clients. The issue of evaluating the mechanisms core the phenomenon of variety of micro blog messages is of extraordinary incentive for some activities, for example, publicity and promotional marketing, affecting and advancing, early observing and crisis reaction [1]. The propagation of information in micro blogs has brought phenomenal enhancement and has accelerated inter-personal communication. Re-tweet mechanism gives an approach to enable social clients to hold the most recent news and help undertakings to do advertising on social-media stage [2]. Along these lines, it is of extraordinary practical importance to examine and investigate the re-tweet practices for improving the data spread and user involvement in micro blogs. Among the some micro blogs, Twitter is one of the best in proliferating information continuously, and the propagation adequacy of a tweet is associated to the number of times the message has been re-tweeted [3]. Re-tweet forecast is a basic and critical task in micro-blogs as it might impact the procedure of information discussion.

The anticipation of message dissemination is one of the key difficulties in understanding the practices of micro blogs. In this research, we study that challenge with regards to the Twitter micro blog. Specifically, our main focus is to foresee the proliferation behavior of any specific tweet within a time of 30 days. This is caught by estimating and anticipating the quantity of re-tweets [4]. To design the re-tweeting actions, we utilize the datasets crept by the WISE-2012-Challenge from Sina-Weibo (SW), which is a prominent Chinese social network webpage like Twitter. In SW, re-tweet process is unique in relation to Twitter. In Twitter, users can just re-tweet a post with no altering a first tweet. Be that as it may, in SW user can change or include data from other clients' in the re-tweeting way in their own re-tweet [5].

Various methodologies has been proposed to show the re-tweet practices dependent on a variety of social features, for example, textual-feature, social-feature, social-influence, visual-feature, emotion, or a blend of above features [6]. While these techniques have gained some growth to some level, the outcomes are inadequate, and can still be enhanced in a specific space. To enhance the performance of forecasting, ongoing works consolidate the observed explicit social data into network factorization structures to model novel methods. Indeed, it is normally that the re-tweet anticipation can be seen as the issue of matrix completion by integrating additional repositories of information about social influence among clients and message semantic between tweets [7].

As revealed in the example of Figure 1, when clients choose to re-tweet text, they are keen on the substance of message and more probable to re-tweets posted by his dear friends because of social relationships. We call this incident for social-context. This familiarity can be gained from social influence and text semantic data. Above two perspectives are significant for re-tweet anticipation. Nevertheless, the vast majority of the current strategies just overlooks such contextual information, or don't exploit these potential highlights [8,19].

Yadala Sucharitha*, Assistant Professor of CSE Dept., CMR Institute of Technology & Research Scholar of JNTUH, Hyderabad, Telangana, India.
Y Vijayalata, Professor of CSE Dept., GRIET, Hyderabad, TS, India.
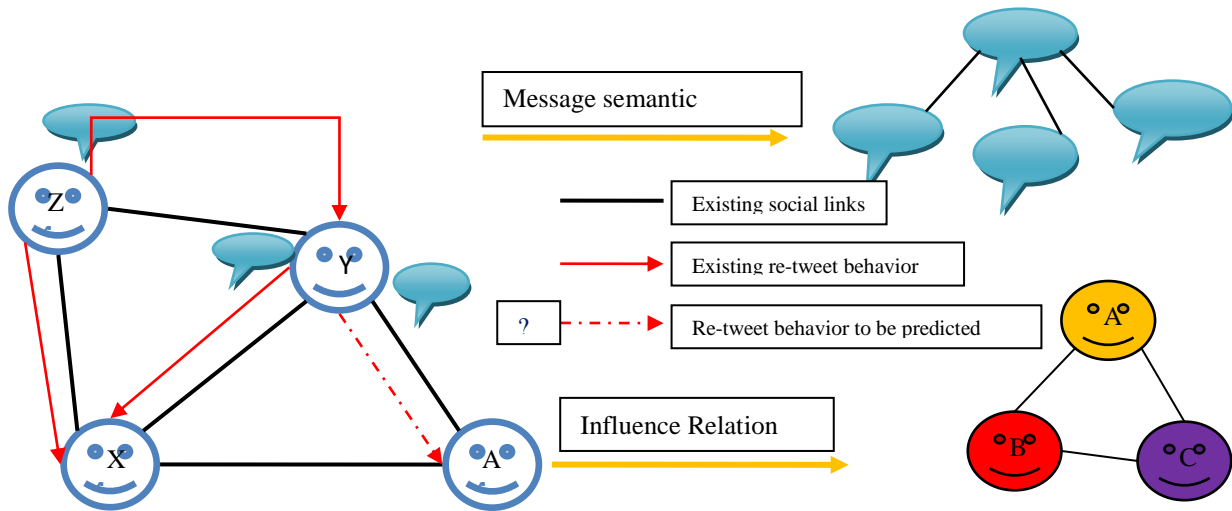**V Kamakshi Prasad**, Professor of CSE Dept., JNTUCE, Hyderabad, TS, India.

Fig. 1: design of re-tweet behaviors on micro blogs

Re-tweeting is a significant client behavior in micro-blogs. Clients can forward the tweets which they are keen on, with the goal that the followers of the clients can witness the tweets too. The tweet circulating pattern and proliferation form, just as its compact introduction with mixed media included, for instance, music, video, and picture format, make the data scattering quicker in micro-blog compared to customary media, with the content and structure being more varied. In this manner, how to anticipate the times of re-tweeting in social networks by investigating the features of tweets spread turns into a rising research event [9, 20]. The consequence of the exploration can be applied in several areas:

1. A tweet that is re-tweeted to a great extent speaks to an emerging issue, so the forecasting on the times of re-tweeting can help out in discovering emerging events in social media.
2. A hot tweet can speak to the center that the vast majorities are concerned about so we can observe people sentiments in a superior manner by anticipating the times of re-tweeting.

Table. 1. Quantity of original messages re-tweeted in 30 days

Re-tweets	Original posts		Annotated with events	
	#posts	%	#posts	%
<10	42,551,890	94.748	882,190	2.072
10-100	2,171,212	4.8348	65,808	3.030
101-500	173,801	0.3869	5,463	3.143
501-1000	10,282	0.0228	401	3.891
1001-5000	2,838	0.006	158	5.567
5001-10000	26	0.00006	2	7.692
>10000	11	0.00002	1	9.091
Total	44,910,062	100.00	954,023	2.124

The above Table 1 presents the sub-sets of posts that have been explained with events. From table 1, around 94 % of the original-tweets were re-tweeted under ten times, of which roughly 2% were commented on with events. Also, most original tweets were re-tweeted in three levels in thirty days as appeared in Table2 & Figure.2. With aim of understanding the re-tweet action, we need to consider the re-tweet movement by every time [18].

3. In addition, social networks responds more quickly contrasted with customary media, particularly on social crisis, so conventional medium like TV, newspapers can outline news dependent on the most recent rising tweets in social network.

The dataset which is utilizing for our study contains two sets of documents. First, follow-ship network, it incorporates the following network of clients dependent on user IDs. Next, Tweets, it incorporates fundamental data about posts, mentions, re-tweet ways, whether containing links. A few posts are explained with events and for every event, the terms that are utilized to distinguish a event and a link to Wikipedia containing depictions to the event are given datasets. Specifically, for the follow-ship dataset, we found that most of clients have less those 10 followers [10]. Furthermore, for the social media dataset, we positioned the circulation of the original posts dependent on quantity of re-tweets they got during 30 days as appeared in Table 1. We chose original-tweets related with events which have the quantity of more than 100 re-tweets for our investigation (6,932). Figure 3 presents quantity of re-tweets every day of week. In Figure 4 illustrates the quantity of re-tweets a day hour-wise. Throughout the day, the most re-tweet action occurs from 10a.m.-12p.m.
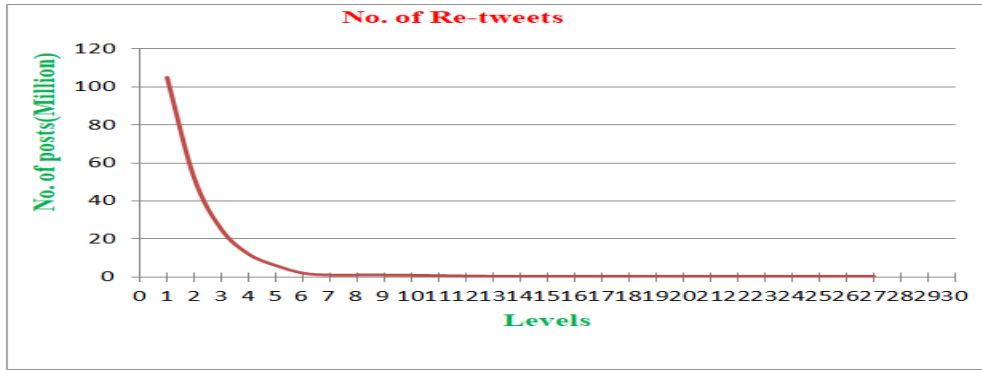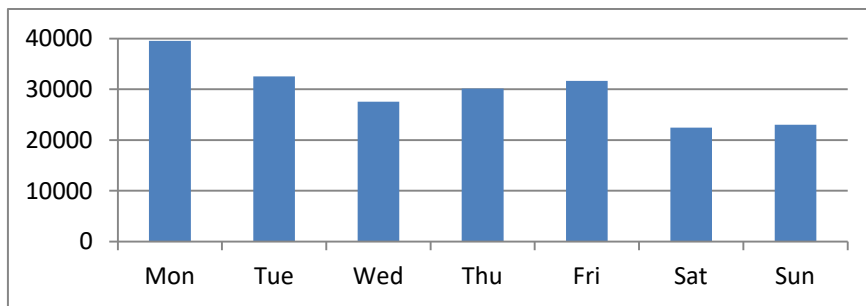
Figure2. Quantity of re-tweets in every level

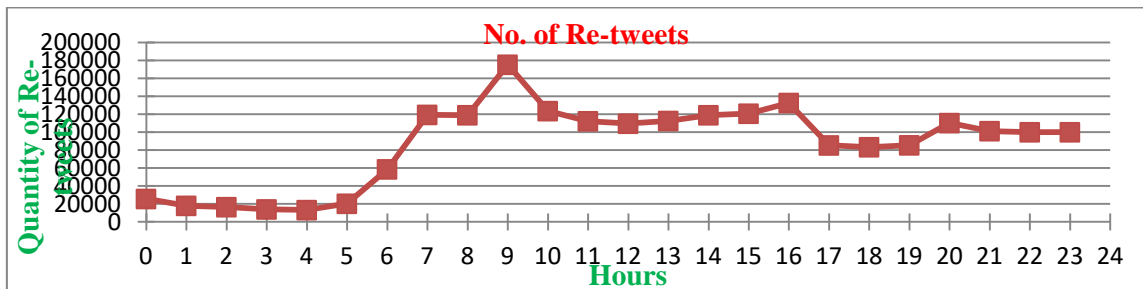

Figure3: Re-tweet activity of a week in day wise

Figure4: Re-tweet movement of a day in hour-wise

Table. 2. Quantity of retweets in 10 levels in a month

Levels	Quantity of Re-tweets	%
First	107,025,957	56.054
Second	49,401,732	25.873
Third	16,934,824	8.852
Fourth	8,045,273	4.221
Fifth	4,196,983	2.176
Six	2,315,741	1.209
Seven	1,294,645	0.686
Eighth	746,487	0.387
Ninth	428,149	0.231
Tenth	240,611	0.131

In this article, we design an innovative re-tweet forecast strategy dependent on client connectivity & event popularity by incorporating the observed re-tweet information, social-influence & message-semantic to enhance the exactness of forecasting. Specifically, we initially present social-influence matrix dependent on

network-structure, communication history, and message likeness-matrix dependent on report semantic [11]. At this point we use client and post idle feature spaces to learn social-influence and message-semantic respectively. At last, we lead a few examinations to approve the effectiveness of our model with the base line models. Empirical outcomes demonstrate our techniwue performs superior to than existing models [12].

The contributions of this study are as pursues:

(1) A broad statistical investigations on the re-tweeting behavior of clients' practices in the generally utilized micro blogs are given.

(2) Quantity of re-tweets is estimated to comprehend the clients' sharing for scattering data in micro blogs.

(3) A strategy to robotically foresee the quantity of re-tweets over social media networks is presented.

(4) Investigations are conducted on vast scale on real-world micro blogs dataset and outcomes show that our method can accomplish superior anticipation performance than base line approaches.

The rest of the article is structured as follows: In Section 2, we analysis the related work. Our proposed methodology is designed in Section 3. The model evaluation and experiential analysis are illustrated in Section 4, Results and discussions in Section 5, followed by the conclusion in Section 6.

II. RELATED WORK

The bloom of social networks stirred wide consideration of numerous researchers. Now a days, they start to carry out research on the issues associated to social networks, as well as examining the contents of social networks, pulling out the association relation among micro-blogs and real society, and foreseeing whether a tweet will be re-tweeted just as the behavior of re-tweeting. In the literature on the examination of micro-blog contents, analysts establish that micro-blog assumes a significant role in numerous areas such as political decisions, marketing investigation, natural calamities analysis and emerging events spreading. Sayan Unankrd [13], designed a model to forecast the quantity of re-tweets for a particular original tweet using probabilistic collaborative filtering anticipation approach called match box depends on client connectivity and event reputation. Authors utilize the dataset crept by the WISE-2012challenge from SW which a reputed Chinese micro-blog just like twitter. They concluded as the proposed model achieved better accuracy compared to baselines approaches. Peng HK et al. [14] developed an approach for predicting a possible number of re-tweets for an original post within a month based on time series circulation of its top n re-tweets called life cycle concept. The proposed method majorly depends on response time, content and time interval. The investigational outcomes show that the proposed model produces satisfactory accuracy in predicting quantity of re-tweets in micro-blogs. Yongjin Bae et al. [3] presented a new approach for forecasting the re-tweet times of posts, creditability of tweet and life span of tweets. Authors extracted the information from re-tweet graphs such as number of times re-tweeted, content features and tweet behavior. Re-tweet forecasting is done based on similar event, re-tweet trend and tweet properties. Authors collected posts from tweeter during June 2012 to December 2012 and they compared the performance with conventional models, the proposed model producing 50 % more precise outcomes than baselines. Paolo Nesi et al. [1] presented a

novel technique called usage of the classification trees with recursive partitioning for forecasting number re-tweets for an original tweet. Dataset is collected from Twitter Vigilance for last 18 months used for investigation. The empirical results compared with baselines in terms of accuracy and processing time and proposed model results are better than existing models. Letierce J et al. [15] developed a link prediction approach for forecasting re-tweets on social media networks based on user interest and event popularity. It gives best prediction results over traditional methods. Gao S et al. [16] present a novel practice for forecasting the spread of messages in a micro blogs. Utilizing information of what and who was re-tweeted, authors train a probabilistic collaborative filter model to forecast future re-tweets and they discover that the most significant features for forecasting are the identity of the source of the tweet and re-tweeter. Tang X et al. [17] presented a new re-tweet prediction approach depends on probabilistic-matrix factorization model by combining the observed re-tweet data, social-influence and tweet semantic to enhance the accuracy of forecast. They concluded as the proposed model outcomes are superior over baseline methods in terms of accuracy and precision. They argue that the re-tweet forecasting model might give enhanced accuracy outcomes when the dissimilarity among the behavior of the user and re-tweeters is considered; determining the subject of interest of a user depends on his previous tweet and re-tweet.

III. METHODOLOGY

3.1 Event Category

In WISE-2012Challenge, the prearranged original tweets are clarified with some social events collectively with their matching keywords. It is hard to automatically aggregate events into various classifications and it is neither in our concentration in this report because certain events are simply marked by people names or by place names [19]. In addition, their pertinent keyword records are subjective and don't show clear relevant data among the keyword list & the event name. To solve this difficulty, we manually segregate the WISE-2012 provided 44 events that have links to Wikipedia pages into 11 categories for instance Natural Disaster, Celebrities, Product Release, Sports, and etc.

To facilitate anticipate the quantity of re-tweets, we propose a probabilistic collaborative filtering anticipation model called Matchbox which is a probabilistic model for creating customized recommendations of things to clients of a web service. It is utilized for the forecast of rating that clients are probably going to allocate to things. It utilizes content data in the structure of client and thing meta-data to learn relationships among them. This approach can be functional to adapt to our issue by the forecasting of re-tweeting probability rather than the forecast of rating. In proposed model, every client and things are signified to by a vector of features. Every feature is related with a latent trait-vector and the linear combination of these for a specific client or thing [20]. We implement this model to foresee whether followers of client will re -tweet the tweet posted by client who has posted an original-tweet. For our methodology, every tweet is viewed as a thing while re-tweeter is considered as a client.

3.2 Tweet and Re-tweeter Features

As per the data-sets which

have been pre-processed by WISE-2012Challenge, we have Followship system &Tweets information with no content. While keyword records are given, they are subjective and don't show clear relevant data between the keywords and the event. For our methodology, all tweets are viewed as a thing while re-tweeter is taken as a client to train the model.

Tweet features comprise of tweet-id, client id who posted the original tweet, number of followers, followees, day of the week, time and category of occasion. Re-tweeter features incorporate client id who re-tweeted the tweet, number of followers & followees. Re-tweeters are mined from all clients who have re-tweeted in the past of each tweet. The binary input is one indicates that the re-twitter re-tweeted the tweet inside 30 days and zero if-not. The result of the model will be the probability of a re-tweet of the tweet by the re-tweeter.

3.3 Training Data

With the aim of train the model, it is needed the positive binary input and also negative input. The positive inputs are from all re-tweet activity in the past of each tweet in a similar event group. For a given tweet, the negative inputs are from all folowers in the re-tweet system who didn't re-tweet a given tweet. For each test event, we train the model by random select 1,000 original tweets in a same event group as things and mine re-tweeters from re-tweet history of each tweet.

3.4 Prediction

To foresee the quantity of re-tweets, for given original tweet and set of clients if client has the high probability of a re-tweet more than threshold, the client is probably going to re-tweet the original tweet. In an attempt to locate the most appropriate value for threshold, we did the forecasting on various threshold values. When threshold value is 0.4 it indicates the best performance of the model. The algorithm is shown in Algorithm 1.

Algorithm 1. Predict Re-tweet via Matchbox

```
Input   : mid     : message id
Output: num_r   : predicted quantity of re-tweets
tweets      = GetPrevious100Messages(mid);
users       = GetRetweeters(tweets);
retweets    = GetRetweetHistory(tweets);
tw_vectors  = CreateTweetF eatures(tweets);
usr_vectors = CreateUserFeatures(users);
model       = TrainModel(tw vectors, usr vectors, retweets);
foreach u ∈ usr vectors do
predict = model.predict(u, mid);
 if predict.getP robT rue() ≥ threshold then
num rt     = num rt + 1;
end
end
return num rt;
```

IV. EXPERIMENTS AND EVALUATION

4.1 Baselines

The two baseline models were compared with our proposed method outcomes.

Baseline 1: Popularity and Connectivity based Regression.
It is a technique to forecast re-tweet activities depends on event popularity and client connectivity by utilizing a naïve technique. The insight is that a tweet is more no. of times to

be re-tweeted if it is regarding a popular event and its writer is extremely connected with others. The forecast will be the evaluation of the probabilities of popularity of category and client connectivity. Re-tweet prediction is calculated by following equation.

$$\text{No. of Re-tweets} = 19.950(0.024C(\text{uid}) + 0.976P(\text{uid, category})) \quad \text{Eq. (1)}$$

Where functionC(uid) is to compute the quantity of re-tweets a uid may have dependent on the quantity of followers, function P(uid, category) is to anticipate how the event category recognition impacts a tweet being re-tweeted.

Baseline 2: User preferences based on classification.
User preferences are utilized to train a classifier to anticipate the conceivable number of re-tweets in thirty days for a given original tweet. Given a unique tweet, the writers require to calculate how conceivable a client will re-tweet the original tweet in the group. The aspirant users are mined from re-tweet repository in a structure of "who-retweet-who". The writers use P(r, u, c) to signify the interestingness of aspirant re-tweet client r to original client u on group c. The function is represented as follows.

$$P(r, u, c) = \frac{\sum RT(r,u,c)}{\sum T(u,c)} \quad \text{Eq. (2)}$$

Where RT(r, u, c) returns the quantity of re-tweets by client r from client u on category c, T(u, c) returns the total number of u's tweet on category c.

V. RESULTS AND DISCUSSIONS

For assessment of proposed method, we anticipated 33 test tweets and the ground truth of 33 tweets are given by WISE-2012 Challenge. For every tweet, we compute the prediction error score (*PES*) utilizing Eq. 3.

$$\text{PES}_i = \frac{|A_i - P_i|}{A_i} \quad \text{Eq. (3)}$$

Where A_i denotes the real value for tweet *i* and P_i is the forecasted value for tweet *i*.

For each method, the average of PES is calculated as follows utilizing Eq. 4.

$$\text{Avg}_j = \frac{\sum_{t=1}^{n} \text{PES}_t}{N} \quad \text{Eq. (4)}$$

Where *n* is the quantity of test tweets. The less number is the superior prediction outcome. Table.3 displays the performance of the proposed method against baselines and its shows; the proposed model performs better in prediction of re-tweets compared to baseline approaches.

Table 3: Average forecast error (AFE) scores

Method	AFE scores
Baseline 1: Popularity and Connectivity based Regression	0.700
Baseline 2: User preferences based on classification	0.666
Proposed Method	**0.568**

VI. CONCLUSIONS AND FUTURE WORK

In this article, we proposed a method to automatically anticipate the quantity of
re-tweets over social networks. The anticipation on the hours of re-tweeting in social media is to quantize the speed of data spread in micro-blogs and to discover the focal point of public attention consistently, which is the key purpose of our examination. In this research, we investigate the behavior properties of re-tweeting in social network and foresee the times of re-tweeting of an original tweet in one month using proposed method. The investigation displays that our methodology can work out the forecast on re-tweeting times, and the average prediction error is controlled inside 20%. Our empirical outcomes the proposed method dependent on re-tweet posts accomplishes better forecast accuracy over existing methods. Our outcomes additionally indicated that the precision of proposed method can be improved by combining re-tweet records and system topology. As future work, we intend to extend this work integrating time delay factor among the posted message and the received client and investigate how the deep learning model can be utilized so that feature vectors of clients and posts can be additionally adapted efficiently.

REFERENCES

1. Nesi P, Pantaleo G, Paoli I, Zaza I. Assessing the re-Tweet proneness of tweets: predictive models for retweeting. Multimedia Tools and Applications. 2018 Oct 1;77(20):26371-96.
2. Bunyamin H, Tunys T. A comparison of retweet prediction approaches: the superiority of Random Forest learning method. TELKOMNIKA, Vol.14 (3) 2016, pp. 1052-1058.
3. Bae Y, Ryu PM, Kim H. Predicting the lifespan and retweet times of tweets based on multiple feature analysis. ETRI Journal. 2014 Jun;36(3):418-28.
4. Sucharitha Y, Vijayalata Y, Prasad VK. Analysis of Early Detection of Emerging Patterns from Social Media Networks: A Data Mining Techniques Perspective. InSoft Computing and Signal Processing 2019 (pp. 15-25). Springer, Singapore.
5. P Chandra Shaker Reddy, A Sureshbabu, " An Adaptive Model for Forecasting Seasonal Rainfall Using Predictive Analytics", International Journal of Intelligent Engineering and Systems, Vol.1(5), 2019, pp. 22-32.
6. Can EF, Oktay H, Manmatha R. Predicting retweet count using visual cues. 22nd ACM international-conference on Information & Knowledge Management 2013 Oct 27 (pp. 1481-1484). ACM.
7. Chen J, Liu Y, Zou M. User emotion for modeling re-tweeting behaviors. Neural Networks. 2017 Dec 1;96:11-21.
8. Suh B, Hong L, Pirolli P, Chi EH. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. Second International-Conference on Social-Computing2010 Aug 20 (pp. 177-184). IEEE.
9. Yadala sucharitha, Y Viijayalata, and V Kamakshi Prasad, "Emergent Events Identification in Micro-Blogging Networks Using Location Sensitivity", JARDCS, Vol. 11(2), 2019, pp. 596-607.
10. Jiang B, Lu Z, Li N, Wu J, Jiang Z. Retweet prediction using social-aware probabilistic matrix factorization. Inter.-Conf on Computational Science 2018 Jun 11 (pp. 316-327). Springer, Cham.
11. Zaman TR, Herbrich R, Van Gael J, Stern D. Predicting information spreading in twitter. Workshop on computational social science and the wisdom of crowds, nips 2010 Dec 10 (Vol. 104, No. 45, pp. 17599-601). Citeseer.
12. Lee K, Mahmud J, Chen J, Zhou M, Nichols J. Who will retweet this?: Automatically identifying and engaging strangers on twitter to spread information. 19th inter.-Conf. on Intelligent User Interfaces 2014 Feb 24 (pp. 247-256). ACM.
13. Unankard S. Prediction of Re-tweeting Activities in Social Networks Based on Event Popularity and User Connectivity. Inter.-Conf. on Machine Learning and Data Mining in Pattern Recognition 2018 Jul 15 (pp. 357-368). Springer, Cham.
14. Peng HK, Zhu J, Piao D, Yan R, Zhang Y. Retweet modeling using conditional random fields. 11th Inter. Conf. on Data Mining Workshops Dec 11 (pp. 336-343). IEEE.
15. Letierce J, Passant A, Breslin J, Decker S. Understanding how Twitter is used to spread scientific messages.
16. Gao S, Ma J, Chen Z. Modeling and predicting retweeting dynamics on microblogging platforms. 8th ACM Inter. Conf. on Web Search and Data Mining 2015 Feb 2 (pp. 107-116).
17. Tang X, Miao Q, Quan Y, Tang J, Deng K. Predicting individual retweet behavior by user similarity: A multi-task learning approach. Knowledge-Based Systems. 2015 Nov 1;89:681-8.
18. P Chandra Shaker Reddy, A Sureshbabu, "A Personalized Artificial Neural Network Model for Crop Yield Forecasting", JARDCS, Vol. 11(05), 2019, pp. 2209- 2219.
19. Unankard S, Chen L, Li P, Wang S, Huang Z, Sharaf MA, Li X. On the prediction of re-tweeting activities in social networks–a report on WISE-2012challenge. Intern. Conf. on Web Information Systems Engineering 2012 Nov 28 (pp. 744-754). Springer, Berlin, Heidelberg.
20. P Chandrashaker Reddy and A Suresh Babu, "An Enhanced Multiple Linear Regression Model for Seasonal Rainfall Prediction", International Journal of Sensors, Wireless Communications and Control (2020) 10: 1. https://doi.org/10.2174/ 2210327910666191218124350.

AUTHORS PROFILE

Yadala Sucharitha, Assitanat professor, she holds bachelor's degree in CSE & Master degree in CSE from JNTUH, Hyderabad and pursing PhD in JNTUH, Hyderabad. Her research field is in the scope of dataminng, big data and advanced compiler design. She has become reviewer for reputed journal. Currently work as Assistant professor of CSE Dept. in CMR Institute of Technology, Hyderabad, India.

Y Vijayalata, an academician with more than 21 years of teaching and research experience. She holds Masters Degree in CS from BITS, Ranchi and PhD in CSE from JNTUH, Hyderabad. Her research field in scope of Image Processing, Cloud Computing and Big-Data Analytics. She have published more than 53 papers in various journals and conferences. She also mentored TCS student internships at GRIET. She has a reviewer for reputed journals and keynote speaker for various conferences. Currently work as professor of CSE Dept. in GRIET, Hyderabad, India.

V Kamakshi Prasad, an academician with more than 25 years of teaching and research experience. He holds Masters Degree in CST from Andhra University and PhD in Speech Recognition, from IIT-Madras. His research field in scope of Speech Recognition and Processing, Image Processing, Data mining and Adhoc Networks. He have published more than 100 papers in various journals and conferences. He has a reviewer for reputed journals and keynote speaker for various conferences. Currently work as professor of CSE Dept. in JNTUHCE, Hyderabad, India and Director of Evaluation in JNTUH.

Retrieval Number: B7365129219/2020©BEIESP
DOI: 10.35940/ijitee.B7365.019320

370

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication