FINDING THE ENDORSEMENT MEASUREMENT FOR SELECTING THE TWEET IN TWITTER TRENDING TOPIC

Mr. Rahul Y. Modi, Prof. Indr Jeet Rajput
Department of Computer Engineering, HGCE, VAHELAL, Gujarat, India

***Abstract*- In the last few year many people wants to create your own network in social site because the social media is very fast growing filed in recent year and every people want to connect to their friends and family with each and every time. So the each user can aware of the rules and regulation of the social media. The topic evolution is created using the time line by summarizing tweet in tweet stream and the sub topic is presented by the chronological way. There are different way to select the sub topic one of them is to select the salient tweet to generate the ranking record based on the salience. Tweet is suffering because the writing style and information is loss so it is differentiate from the traditional document.so in this research work we manly focus on performance using the text based sequential summarization and network based sequential summarization for sub topic detection and we also improve the quality of the tweet is also increases. Also get accurate tweet difference in different field like coffee, gold, silver, cocco and many more so using this proposed flow we can reduce time and don't get manipulate. We can get the batter improvement if the quality of the tweet content is good, if the quality of the content is batter then we can give the proper relationship to each tweet.**

I. INTRODUCTION

The information is acquiring using the process of data mining techniques. The data mining is the statistical techniques for knowledge discovery process. The knowledge discovery process is to extract the unknown, valid and important data form large information database. "The process requires a definition of the project goals, dataset acquisition, data cleaning and preprocessing, data mining, data interpretation and utilization. We will define data mining as the application of statistical techniques, e.g. predictive modeling, clustering, link analysis, deviation detection and disproportional measures, to databases. The modern technologies of computers, networks, and sensors have made data collection and organization an almost effortless task. However, the captured data need to be converted into information and knowledge from recorded data to become useful. Analysis specialist can performed the different task and extract the meaningful information form the large amount of database. Modern business and science have increases the volume of data for computer based approaches, but the data have been grown in terms of size and complexity. The automatic data analysis using more complex and sophisticated tools. Evaluating the part of the sub-topic it is major task for user to divide the part of the sub-topic from the tweet stream because there are hundreds of thousand tweets in same sub topic. For example in this work we can use the different sub topic like energy, grain, soft, metals and etc Supposed in metal part there are different categories like gold, silver, platinum and many more so in same time the price of the each categories is either increases or decrees and it is also possible to steady the price of the metal as per the demand of the metal. So that the cluster of tweet is specified in each sub topic. The clustering task is differentiating using text clus tering task. In this text clustering task we have to indicate the each and every segment of the text is adjacent using the average cosine. To get the better result we have to separate the lower average cosine because the date are generate on online social network is generated by different user on different network . The data generated by user is unstructured, noisy, vast, and dynamic so we can use the proper method to achieve batter results. So performing this operation we use the text base sequential summarization in our proposed work for text clustering task. The following charteristic is use to get the batter results using data mining techniques for online social network analysis (OSNA).

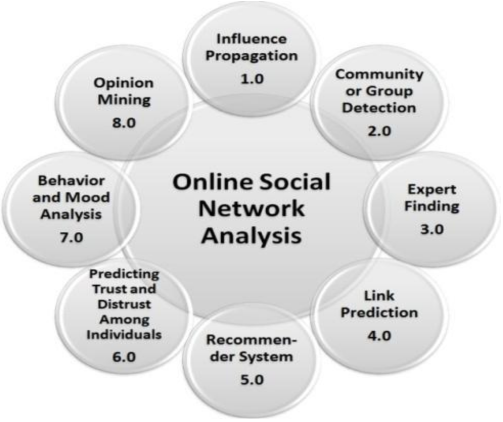

Influence Propagation 1.0
Community or Group Detection 2.0
Expert Finding 3.0
Link Prediction 4.0
Recommen-der System 5.0
Predicting Trust and Distrust Among Individuals 6.0
Behavior and Mood Analysis 7.0
Opinion Mining 8.0
Online Social Network Analysis

Fig 1. Key Research Issues in Online Social Network Analysis [13]

II. RELETED WORK

Link prediction is the problem of predicting the existence of a link between two entities, based on attributes of the objects and other observed links. The predicting task, such as predicting trust, distrust, friendship, co-authorship and other interactions, could be represented as predicting the link's value in social networks. Link prediction problem is relevant for different domains; several techniques have been proposed to solve it. Most of them are usually based on structural features and supervised methods. The base algorithm has several disadvantages. So we use the text based and network based sequential summarization techniques to solve the disadvantage of the base algorithm.

III. PROPOSED WORK

We use different representations in the classification process: the Twitter features described in proposed algorithm and flowchart we use the text based and network based sequential summarization for textual content of tweets. We create two different groups of networks and we also create the vector pre trending topic of representation. For performing this proposed work a training set is using with different trending topic and test data set. And performing random selection from different trending topic and test data set and give the results. For performing this task we use the following representation.

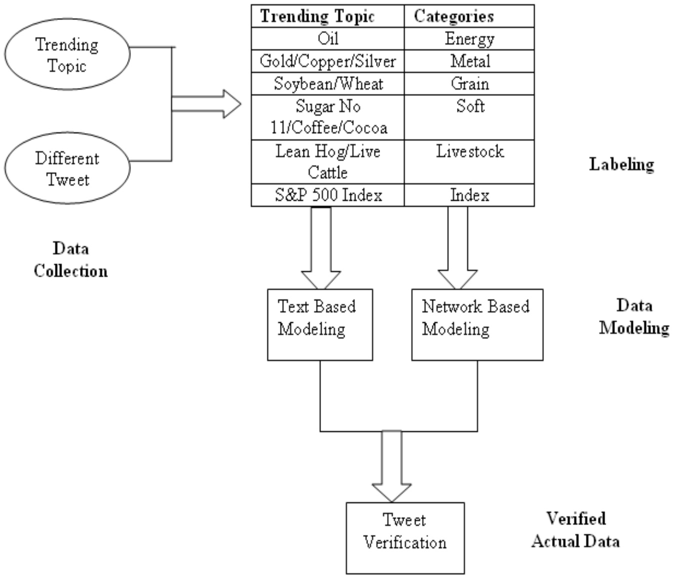
Data Collection: In Data collection representation the different data set is collected form the ONSs and they corresponding to the features introduced above mention in the proposed work. In this proposed work the data can be collected form the online social network site like twitter.

Labeling: When the sample data are selected from the training data set then after you have to give some label to this data for example different oil as give the label energy , gold ,silver have given the label metal and etc.

Data modeling: Data modeling is the formalization and documentation of existing processes and events that occur during application software design and development.

Verification: Verification is the final step of the proposed work and in this phase all the tweet is verified and it is classified according to the fields.

3.1 PROPOSED ARCHITECTURE:



Trending Topic	Categories
Oil	Energy
Gold/Copper/Silver	Metal
Soybean/Wheat	Grain
Sugar No 11/Coffee/Cocoa	Soft
Lean Hog/Live Cattle	Livestock
S&P 500 Index	Index

Fig 3.1 Architecture of Proposed Work

3.2 PROPOSED ALGORITHM.

Step 1: Select the appropriate dataset.

Step 2: Categories the data set and divided the tweet stream S into a serial of time Slices with the time window Δt.

Step 3: Initializing the mean and variance with the first time interval.

Step 4: if the tweet stream is greater then τ then tweet number (mean) is increasing other-wise Remain as it is. When the tweet number is increases then mean is Update. This process is Continuous up to the tweet number is reach to the starting number.

Step 5: Select quasi identifier, key attribute and Sensitive attribute from selected tweet Stream S.

Step 6: Each tweet S*i* is viewed as a mixture of topics in Z, and each topic Z*j*.

Step 7: Then after Rearrange element of tweet S*i*

Step 8: Categorize sensitive attribute values of table T into high (H) and low (L) sensitive class. (High value means Text based detection. And Low value means Network based detection)

a)Differentiate H and L value

Step 9: Combine H and L value

Step 10: After combine H and L value of the tweet stream it will be verified. And get the finally output.

3.3 PROPOSED FLOWCHART.

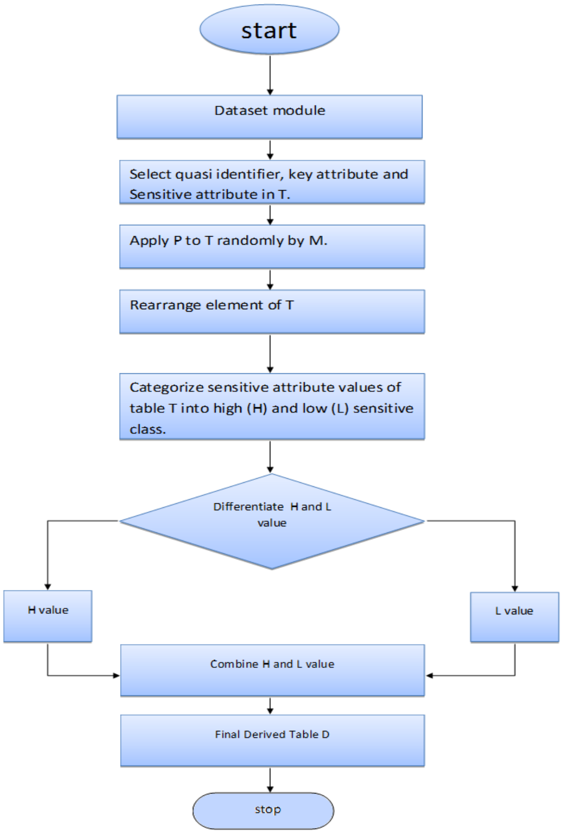

start
Dataset module
Select quasi identifier, key attribute and Sensitive attribute in T.
Apply P to T randomly by M.
Rearrange element of T
Categorize sensitive attribute values of table T into high (H) and low (L) sensitive class.
Differentiate H and L value
H value
L value
Combine H and L value
Final Derived Table D
stop

Fig 3.2 Flow chart for proposed method

3.4 Dataset Description

We take different gradient energy, metal, soft, grains as an example in our experiment. We can create the different summary for the sub topic detection using this gradient. In this research work use the number of tweets containing the key words metal, soft, grains between last few years using public twitter API 5. Form this we select the different categories for further processing as shown in the below figure .Matching keywords can be done in two way.

1) The corresponding has tag is selected from the given tweet
2) Hase tag has much concatenation word then it will divide in to number of parts.

So for the specific topic we can select the number of tweet and they are segment in to sub topic. The sub topic can contain the maximum number of tweet on average .Then after it will be filter some tweet in following condition.

1) Maximum number of length is up to 3 characters.
2) They can't contain extra words then main topic.
3) After filtering the other words are not allow excepted URL.

3.5 IMPLEMATATION DETAILS.

We define the Crowding Endorsement which suggests that the tweet gaining more endorsements from the crowds will be regarded as more important. In Twitter, a tweet can be re-tweeted by many others if they think it is important and we thus use re-tweeting to calculate the importance of crowding endorsement

Global Relevance: The global relevance of the tweet is defined as the cosine similarity between the tweet and the entire stream S.

$$\mathrm{GRel}(s_i) = \mathrm{cosine}(s_i, S) = \frac{V_{s_i} \cdot V_S}{\|V_{s_i}\| \, \|V_S\|}$$

Local Relevance: Assume that the tweets in a peak area represent a sub topic in the topic. The local relevance of the tweet is defined as the cosine similarity between the tweet and the tweets in the peak area that belongs to, i.e.

$$\mathrm{LRel}(s_i) = \mathrm{cosine}(s_i, p_j) = \frac{V_{s_i} \cdot V_{p_j}}{\|V_{s_i}\| \, \|V_{p_j}\|}$$

Crowding Endorsement: The endorsement of the tweet S*i* from the crowds is measured by the normalized re-tweeting count.

$$\mathrm{Eds}(s_i) = \frac{\mathrm{RetweetCount}(s_i)}{\|\mathrm{TotalRetweetCount}\|}$$

The weighed linear combination of the above three measures gives the final significance score of a tweet

$$Score(s_i) = \alpha_1 \cdot \mathrm{GRel}(s_i) + \alpha_2 \cdot \mathrm{LRel}(s_i) + \alpha_3 \cdot \mathrm{Eds}(s_i)$$

Where $\sum_i \alpha_i = 1$. The tweets in different peak areas are scored and ranked independently.

IV. RESULTS AND ANALYSIS

Results are generated to taking the number of experiment on given data set and compare with the skewed class distribution. The proposed algorithm is use on the different tweet data set for evolution. In the proposed algorithm we firstly use the dynamic and semantic model with supervised and unsupervised algorithm to improve the accuracy of the link predication algorithm. The output of the dynamic and semantic model are directly access in the network base and text base sequentially summarization for getting the better results. After doing implementation of the existing approach, the comparison of the result is to be analysed. The result analysis is done by comparing existing result of the twitter trending table which is give in the base paper and result taken by our implementation. The comparison of the network based and text base sequential summarization approaches. Due to the shortage of gold-standard sequential summaries, we invite two annotators to read the chronologically ordered tweets, and write a series of sub-summaries for each topic independently. Each sub-summary is up to 140 characters in length in base paper implementation but in our proposed work we can give the maximum number of length of tweet to comply with the limit of tweet, but the annotators are free to choose the number of sub-summaries.

The link predication method is use the standard data set for performing the proposed algorithm and flowchart. The standard data set are classified on to two different data set. Two networks are created NET1 and NET2 with the number slot between 0 to 500.we must careful that each group NET1 and NET2 have different tweet. Also remove the nodes that has only 20 activities and 2 links to the given time period. NET1 and NET2 have extracted the maximum connected component in NET1 and NET2. First the NET1 pair is refer and then set the activity in NET1 within number slot between 0 to 500 stander dataset then NET2 pair is refer and then set the activity in NET2 within number slot between 0 to 500 stander dataset then As show in the fig 5.1 part A. as a results analysis the proposed method is select the random outperformed for batter accuracy. The result of the proposed method is shown in fig 4.5.

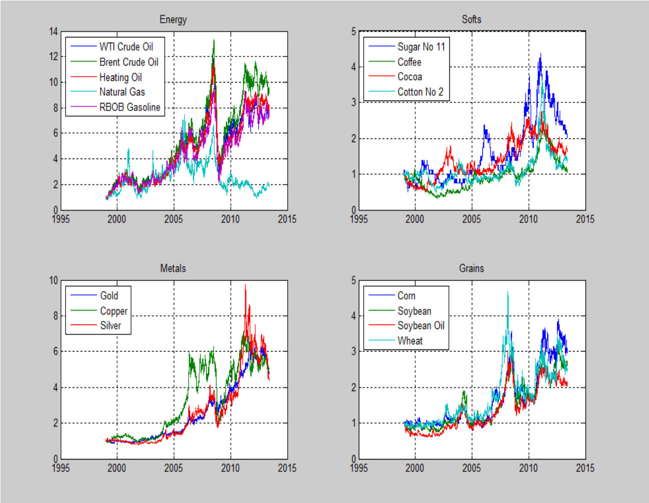


Fig 4.1 Positive Twit on Energy, Metal, Soft, Grains

We can see that precision decreases when *k* ranges from o to 5 while the recall increases as well in the year 1995 to 2015 the reason is that the more links we predict, the more correct links are revealed in the test data. However, the predictions ranking in the bottom do not capture as many correct links as the top ones.

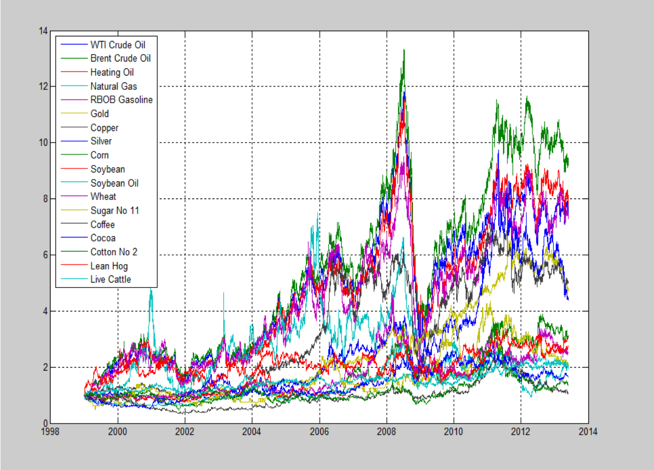


Fig 4.2 Total Tweet on Each And Every Year on Different Gradient

Show in Fig 4.2 Efficiency and performance under different thresholds. We can find that smaller threshold leads to less time complexity but poorer performance. For this experiment we can collect the different files from the year 1998 to 2014.

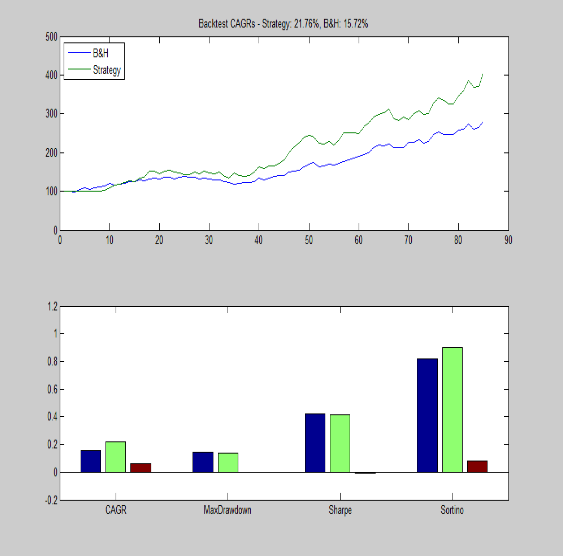


Fig 4.3 Shows Changes In Eigenvalues

Fig. 4.3 shows changes in eigenvalues between snapshots taken at different time instances for DBLP and different datasets (characteristics of the datasets are discussed in Section) It shows that *ith* eigenvalue increases over time and therefore curve fitting can be used to predict $\Lambda t+k$ i.e., curve fitting previous spectrum $\Lambda t, \Lambda t+1, \Lambda t+2, ..., \Lambda t+k-1$.

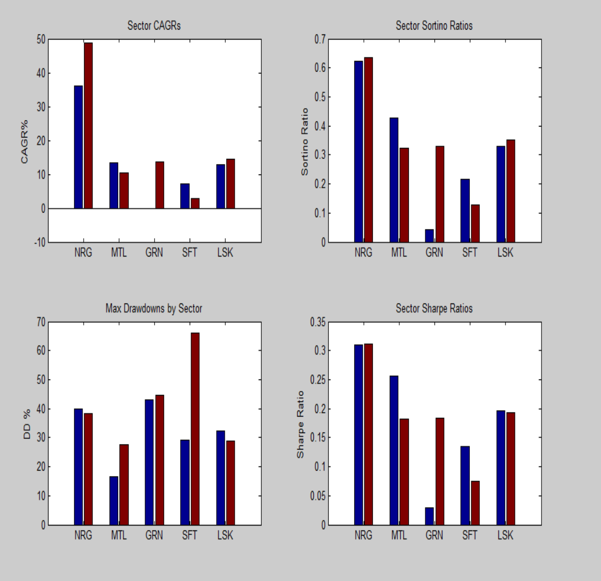


Fig 4.4 Tweet Streams Of Different Topic

Fig. 4.4 shows the tweet streams of different topic and the high-frequency words in Different areas of the tweet stream related to the topic. For different topics, an explosive growth is clearly observed at the beginning of the tweet stream. The volume of the tweets then fluctuates up or down for a while, either dramatically or slightly, and gradually declines with a long tail.

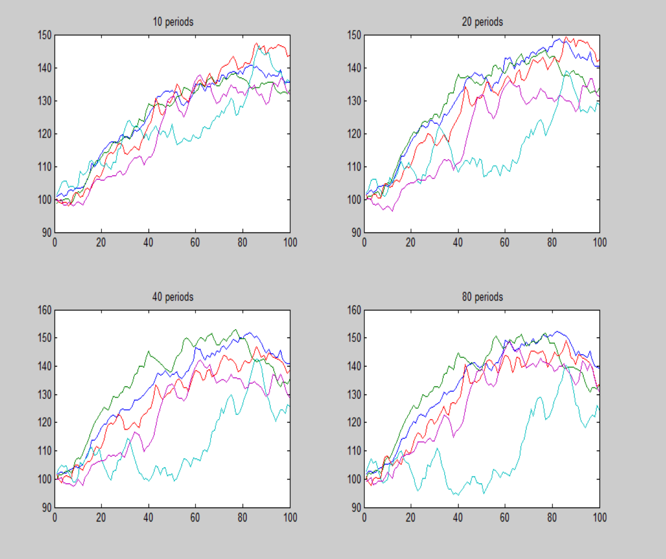


Fig 4.5 Describe Time Period on Tweet per hours

Fig 4.5 Part A, B, C, and D represent the tweet link between the given time period 10 ,20 , 40, and 80 hours respectively for the categories like energy, metal , soft, gain and Live stocks as mention in the data set table.

V. CONCULTION

The overview of the topic evolution is created using the time line by summarizing tweet in tweet stream and the sub topic is presented by the chronological way. There are different way to select the sub topic one of them is to select the salient tweet to generate the ranking record based on the salience. Tweet is suffering because the writing style and information is loss so it is differentiate from the traditional document.so in this research work we manly focus on performance using the text based sequential summarization and network based sequential summarization for sub topic detection and we also improve the quality of the tweet is also increases. Also get accurate tweet difference in different field like coffee, gold, silver, cocco and many more so using this proposed flow we can reduce time and don't get manipulate. We can get the batter improvement if the quality of the tweet content is good, if the quality of the content is batter then we can give the proper relationship to each tweet. As a future work we can use real rime server for check and implement proposed flow. And match with different gradient and things like gold, silver, fruit, and coco. As a future work we can use real rime server for check and implement proposed flow. And match with different gradient and things like gold, silver, fruit, and coco.

REFERENCE

[1] Takayuki Kamei, Keiko Ono, Masahito Kumano, And Masahiro Kimura "Predicting Missing Links In Social Networks With Hierarchical Dirichlet Processes" , Ieee World Congress On Computational Intelligence, 2012.

[2] Feng Liu, Bingquan Liu, Xiaolong Wang, Ming Liu, And Baoxun Wang "Features For Link Prediction In Social Networks: A Comprehensive Study" Ieee International Conference On Systems, Man, And Cybernetics, 2012

[3] Jorge Valverde-Rebaza And Alneu De Andrade Lopes "Structural Link Prediction Using Community Information On Twitter" , Ieee,2102

[4] Ting Jin, Tong Xu, Enhong Chen, Qi Liu, Haiping Ma, Jingsong Lv, Guoping Hu "Random Walk With Pre-Filtering For Social Link Prediction" Ninth International Conference On Computational Intelligence And Security,2013

[5] Deepak Mangal, Niladri Sett, Sanasam Ranbir Singh, Sukumar Nandi "Link Prediction On Evolving Social Network Using Spectral Analysis" Ieee Ants 2013

[6] Ryan N. Lichtenwalter , Jake T. Lussier, And Nitesh V. Chawla "New Perspectives And Methods In Link Prediction" Acm,2010

[7] Dehong Gao, Wenjie Li, Xiaoyan Cai, Renxian Zhang, And You Ouyang "Sequential Summarization: A Full View Of Twitter Trending Topics" Ieee/Acm Transactions On Audio, Speech, And Language Processing, Vol.22, No. 2, February 2014

[8] Fenhua Li, Jing He, Guangyan Huang , Yanchun Zhang, And Yong Shi Iccs "Clustering Based link prediction Method In Social Networks ". 14th International Conference On Computational Science 2014.

[9] Wadhah Almansoori • Shang Gao • Tamer N. Jarada"Link Prediction And Classification In Socialnetworks And Its Application In Healthcare And Systems Biology" Springer-Verlag 2012

[10] Jianhan Zhu, Jun Hong, And John G. Hughes "Using Markov Chains For Link Prediction In Adaptive Web Sites" Springer ,2002

[11] Neil Zhenqiang Gong , Ameet Talwalkar, Lester Mackey And Ling Huang "Jointly Predicting Links And Inferring Attributes Using A Social-Attribute Network San " The 6th Sna-Kdd Workshop(Sna-Kdd'12), Aug.12, 2012, Beijing, China .2012

[12] Michael Fire, Lena Tenenboim-Chekina, Rami Puzis, Ofrit Lesser, Lior Rokach, And Yuval Elovici, "Computationally Efficient Link Prediction In A Variety Of Social Networks" Acm Transactions On Intelligent Systems And Technology,2013

[13] G Nandi, And A Das "A Survey On Using Data Mining Techniques For Online Social Network Analysis" Ijcsi International Journal Of Computer Science Issues, 2013

[14] Amir Hossein Rasekh, Zeinab Liaghat, and Ala Mahdavi "Predict Edges in Fliker Social Network Using Data Mining Method, Science direct 2012" Science Direct, 2012

[15] Bruce Hoppe and Claire Reinelt "Social network analysis and the evaluation of leadership networks "Elsevier, 2012

[16] D. M. Boyd and N. B. Ellison, "Social Network Sites: Definition, History and Scholarship," Journal of Computer-Mediated Communication, Vol. 13, No. 1, 2008, pp. 210-230.