Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 6, June 2015, pg.1075 – 1080

RESEARCH ARTICLE

Profile Ranking Using User Influence and Content Relevance with Classification Using Sentiment Analysis

Ashwini Sopan Shidore[1], M.R.Sindhu[2]
[1]PG Scholar, Department of Computer Engineering, GHRCOEM, Ahmednagar, University of Pune, India
[2]Asst. Professor, Department of Computer Engineering, GHRCOEM, Pune, University of Pune, India
[1] ashidore419@gmail.com; [2] sindhu.mrs@gmail.com

*Abstract— **It is a fundamental step towards the design of effective information diffusion mechanisms to understanding information diffusion processes that take place on the Web, specially in social media. Influence and relevance are two key concepts in information diffusion. Ability to popularize content in an online community is the influence. Influentials introduce relevant content, in the sense that such content satisfies the information needs. We propose Profile-Rank is a new information diffusion model supported random walks over a user content graph. Profile-Rank could be a Page-Rank inspired model that described the principle that relevant content is created and propagated by prestigious users and prestigious users create relevant content. We also propose a sentiment analysis. In this we get weight for profile ranking with its approach (Like positive approach or negative approach).***

Keywords— Information Diffusion, Content Relevance, User Influence ,Twitter ,Profile-Rank, Sentiment Analysis.

I. INTRODUCTION

Powered by the remarkable success of Twitter, Facebook, Youtube, and the blogosphere, social media is taking over traditional media as the major platform for content distribution. The combination of user-generated content and online social networks is the engine behind this revolution in the way people share news, videos, memes, opinions, and ideas in general. As a consequence, understanding how users consume and propagate content in information diffusion processes is a most important step towards the design of effective information diffusion mechanisms, viral marketing and recommendation systems on the Web. Influence and relevance are the two key concepts in information diffusion. In social networks, influence can be defined as the capacity to affect the behavior of others [2].However, in information diffusion scenarios, influence is usually a measure of the ability of popularizing information. Relevance is a relationship between a user and a piece of information, in the sense that relevant information satisfies a users information needs/interests, being a fundamental concept also in information retrieval and recommender systems. This work focuses on the link between user influence and informa- tion relevance in information diffusion data, which describe how users create and propagate information across time. As we are interested in the diffusion of content (e.g., news,videos) on the Web, we use the terms content and information interchangeably. Here, we present Profile-Rank, a random walk based information diffusion model that computes user influence and content relevance using information diffusion data.

Profile-Rank is based on the principle that influential users create relevant content and relevant content is created and propagated by influential users. If we consider Twitter as an information diffusion platform and tweets as content propagated through retweets. Profile-Rank can be intuitively described in terms of the behavior of a random tweeter (or twitterer) that navigates through Twitter Profile by clicking on random tweets (or retweets from these same tweets). Every click on a tweet leads the random tweeter to the profile of the original author of the tweet. We measure user impudence as the frequency with which the random tweeter visits a given profile[1].Likewise, content relevance is measured as the frequency with which the random tweeter clicks on a tweet and its retweets.

Social media has become one of the biggest forums to ex- press ones opinion. Aim of sentiment analysis is to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation, affective state (the emotional state of the author when writing).Classifying the polarity of a given text at the document, sentence, or fea- ture/aspect level, whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral is the basic task in sentiment analysis.

II. LITERATURE REVIEW

M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. [3]Described that using a large amount of data collected from Twitter, we present an in-depth comparison of three measures of influence: indegree, retweets, and mentions. Based on these measures, we investigate the dynamics of user influence across topics and time. We make different interesting observations. First, popular users who have high indegree are not necessarily influential in terms of spawning retweets or mentions. Second, most influential users can hold significant influence over avariety of topics. Third, influence is not gained spontaneously or accidentally, but through concerted effort such as limiting tweets to a single topic. We believe that these findings provide new insights for viral marketing and suggest that topological measures such as indegree alone reveals very little about the influence of a user.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon.[4] start with the network analysis and study the dis- tributions of followers and followings, the relation between followers and tweets, degrees of separation. Next by consid- ering the number of followers, the number of retweets and PageRank and present quantitative comparison among them we rank the users. The ranking by retweets pushes those with fewer than a million followers on top of those with more than a million followers. Through our topic analysis we show what different categories trending topics are classied into, how long they last, and how many users participate. Finally, we study the information diffusion by retweet. We construct retweet trees and examine their spatial and temporal characteristics.On the entire Twittersphere and information diffusion on it, this work is the first quantitative study.

Alekh Agarwal et al., [7] proposed a machine learning method incorporating linguistic knowledge gathered through synonymy graphs, for effective opinion classification. This approach shows the degree of influence among relationships of documents have on their sentiment analysis. This is brought about by the use of graph-cut technique and opinion words got through synonymy graphs of Wordnet.

Arlei Silva, Hrico Valiati, Sara Guimares, Wagner Meira Jr.[5] described how individual behavior data may provide knowledge regarding influence relationships in a social net- work and also dene what we call the influence network discovery problem, which consists of identifying influence relationships based on user behavior across time. Several strategies for influence network discovery are proposed and discussed.

III.PROPOSED SYSTEM

3.1 Problem Statement:

Here, we study the problem of identifying influential users and relevant content in information diffusion data. We propose ProfileRank, a new information diffusion model based on random walks over a user-content graph. ProfileRank is based on the principle that relevant content is created and propagated by influential users and influential users create relevant content. If we consider Twitter as an in- formation diffusion platform and tweets as content propagated through retweets, ProfileRank can be intuitively described in terms of the behaviour of a random tweeter (or twitterer) that navigates through Twitter profiles by clicking on random tweets (or retweets from these same tweets). Every click on a tweet leads the random tweeter to the profile of the original author of the tweet. We measure user influence as the frequency with which the random tweeter visits a given profile. Likewise, content relevance is measured as the frequency with which the random tweeter clicks on a tweet and its retweets.

3.2 System Architecture:

In this portion, architecture of proposed system is given. Profile-Rank is a model for information diffusion that computes user influence and content relevance based on a bipartite directed graph that describes the flow of information among users. Data gathering for this task involved more effort than expected and required hand-tagging posts for sentiment in relation to a query.

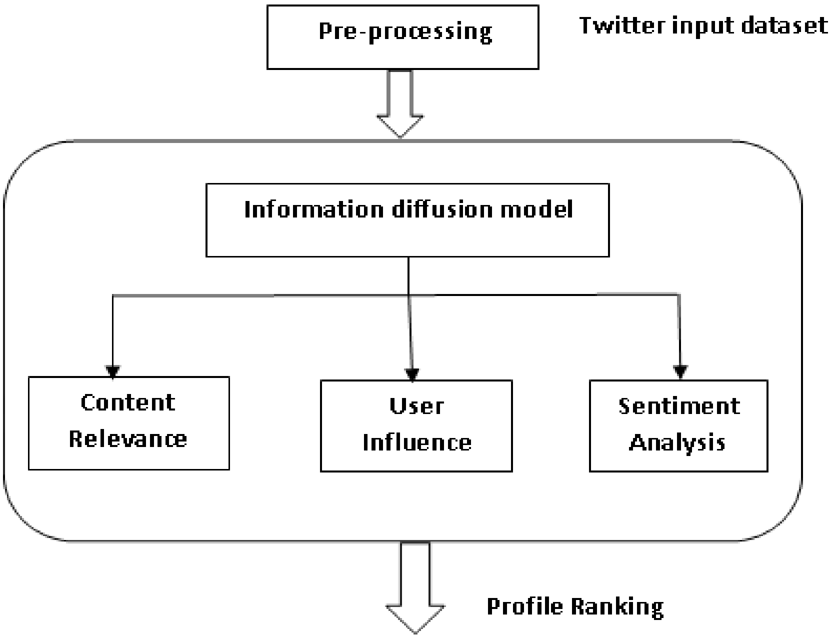


Fig 1.System Architecture

1. **Preprocessing :**

Due to the nature of language used in micro-blogging posts, preprocessing on the messages is very necessary and has been shown to improve performance, especially for smaller training sets. The pre-processing is necessary because there are some words or expressions in the review don't return any meaning and by the presence of those words we cannot get the correct sentiment analysis. So by doing pre-processing we get higher accurate results. In pre-processing we do Remove URLs, Remove Repeated Letters, Remove Special Symbols and Remove Questions.

2. **Information Diffusion Model:**

Information diffusion data is a sequence of occurrences of content. Each occurrence of a piece of content is defined as a tuple in the form ¡ u; c; t ¿, where u is a user from the set of users U,c is a piece of content from the content set C, and t is a timestamp. For a given tuple ¡ u; c; t ¿, we say that the user u propagated c at time t. Therefore, information diffusion data describes associations between users and content across time. Using this notation, we define an information diffusion dataset as a triple D = (U;C; T).

3. **Content Relevance and User Influence:**

Random tweeter starts from a random profile and keeps clicking on tweets and retweets at random. The random tweeter is redirected to the profile of the original author of a tweet by clicking on it. The relevance of a tweet is the relative frequency that the random tweeter clicks on a tweet, or one of its retweets. Moreover, the frequency that the random tweeter visits a users profile is a measure of this users influence. We calculate content relevance and user influence by using diffusion data and it is calculated by using user-content matrix M and content user matrix L. Information diffusion graph is bipartite graph. Profile Rank computes user influence and content relevance based on a user-content bipartite directed graph.

4. **Do Sentiment Analysis:**

As a proposed work we are doing profile ranking and also sentiment analysis of profile. To improve profile ranking we classify profile in three classes :

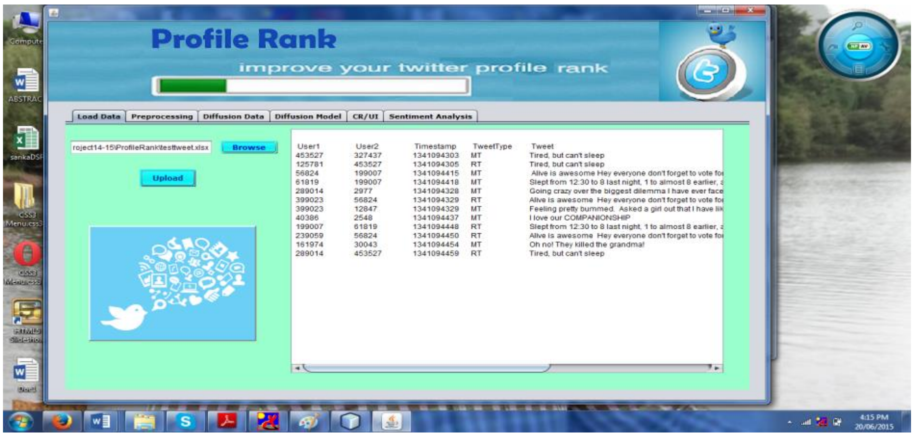
1) Positive 2) Negative 3) Neutral .

This classification will be done by doing sentiment analysis of tweets and retweets on that profile. In this navie bayes we calculate posterior for each class for every profile. Depending on posterior value i.e. greater posterior value for class is assigned to that profile. This will used to find status of profile.Weight consideration for each distinct item in a transaction in independent manner adds effectiveness for finding infrequent itemset mining.

IV. RESULT

Proposed system evaluated and studied through simulating the opportunistic environment and generating results.

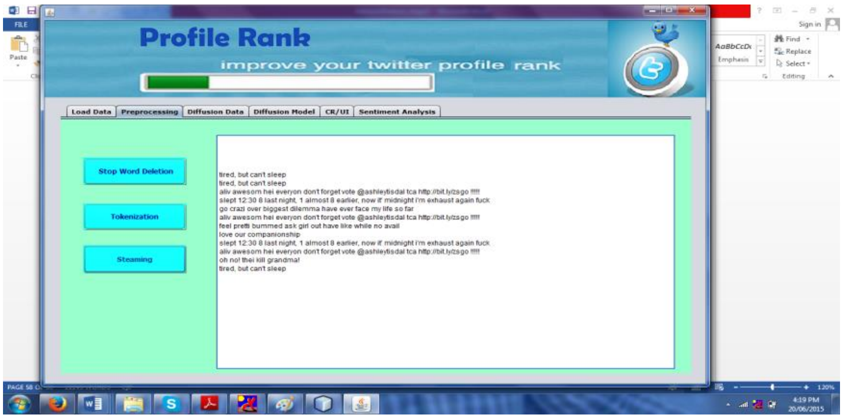
1. Firstly we have to upload a twitter dataset so we can do operations on that:



User1	User2	Timestamp	TweetType	Tweet
453527	327437	1341094303	MT	Tired, but can't sleep
125781	453527	1341094305	RT	Tired, but can't sleep
56824	199007	1341094415	MT	Alive is awesome Hey everyone don't forget to vote fo
61819	199007	1341094418	MT	Slept from 12:30 to 8 last night, 1 to almost 8 earlier,
289014	2977	1341094328	MT	Going crazy over the biggest dilemma I have ever face
399023	56824	1341094329	RT	Alive is awesome Hey everyone don't forget to vote fo
399023	12847	1341094329	MT	Feeling pretty bummed. Asked a girl out that I have lik
40386	2548	1341094437	MT	I love our COMPANIONSHIP
199007	61819	1341094448	RT	Slept from 12:30 to 8 last night, 1 to almost 8 earlier,
239059	56824	1341094450	RT	Alive is awesome Hey everyone don't forget to vote fo
161974	30043	1341094454	MT	Oh no! They killed the grandma!
289014	453527	1341094459	RT	Tired, but can't sleep

Fig : Twitter Dataset

2. After that have to preprocess input dataset by doing stopword deletion from it, tokenization and stemming process:



```
tired, but can't sleep
tired, but can't sleep
aliv awesom hei everyon don't forget vote @ashleytisdal tca http://bit.ly/zsgo !!!!!
slept 12:30 8 last night, 1 almost 8 earlier, now it' midnight i'm exhaust again fuck
go crazi over biggest dilemma have ever face my life so far
aliv awesom hei everyon don't forget vote @ashleytisdal tca http://bit.ly/zsgo !!!!!
feel pretti bummed ask girl out have like while no avail
love our companionship
slept 12:30 8 last night, 1 almost 8 earlier, now it' midnight i'm exhaust again fuck
aliv awesom hei everyon don't forget vote @ashleytisdal tca http://bit.ly/zsgo !!!!!
oh no! thei kill grandma!
tired, but can't sleep
```

Fig: Preprocessing dataset

3. From dataset content we have to create diffusion data & diffusion graph.

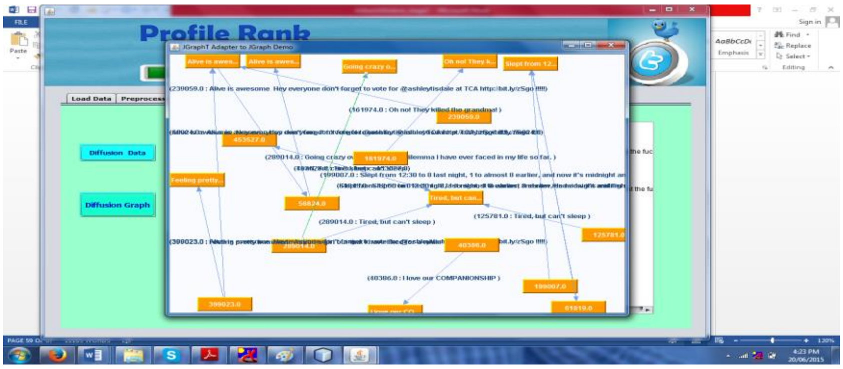
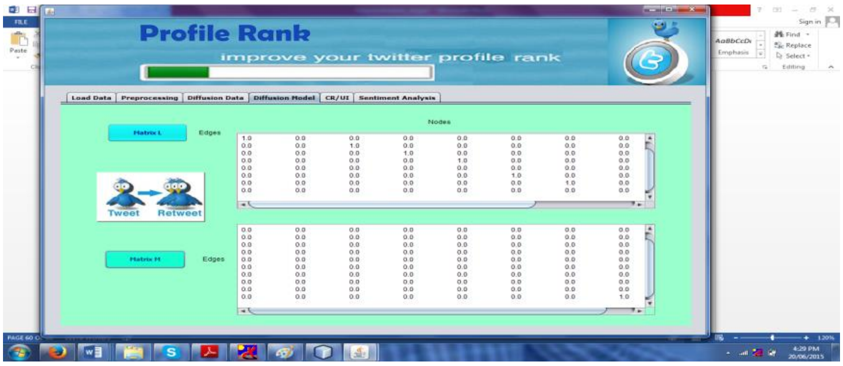

Profile Rank
JGraphT Adapter to JGraph Demo
Load Data
Preprocess
Diffusion Data
Diffusion Graph
Alive is awes...
Alive is awes...
Going crazy o...
Oh no! They k...
Slept from 12...
(239059.0 : Alive is awesome Hey everyone don't forget to vote for @ashleytisdale at TCA http://bit.ly/zSgo !!!!!)
(161974.0 : Oh no! They killed the grandma!)
239059.0
453527.0
(289014.0 : Going crazy o
161974.0
ilemma I have ever faced in my life so far.)
(199007.0 : Slept from 12:30 to 8 last night, 1 to almost 8 earlier, and now it's midnight an
Feeling pretty...
56824.0
Tired, but can...
(125781.0 : Tired, but can't sleep)
(289014.0 : Tired, but can't sleep)
125781.0
(399023.0 :
289014.0
40386.0
bit.ly/zSgo !!!!!)
(40386.0 : I love our COMPANIONSHIP)
199007.0
399023.0
I love our CO...
61819.0

Fig : Diffusion Graph

4. From this diffusion model we achieve matrix L & M

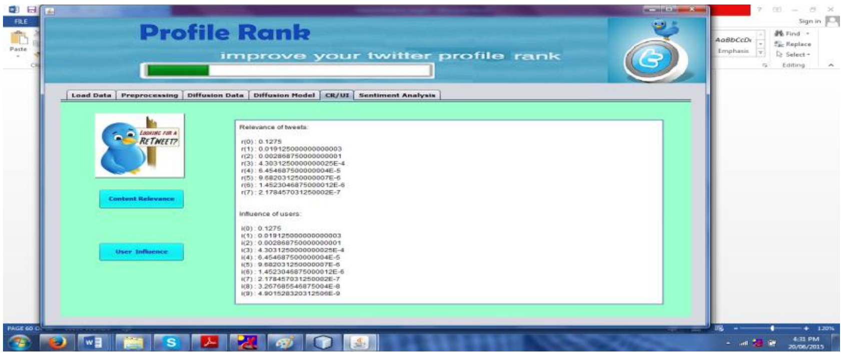


```
1.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
0.0   0.0   1.0   0.0   0.0   0.0   0.0   0.0
0.0   0.0   0.0   1.0   0.0   0.0   0.0   0.0
0.0   0.0   0.0   0.0   1.0   0.0   0.0   0.0
0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
0.0   0.0   0.0   0.0   0.0   1.0   0.0   0.0
0.0   0.0   0.0   0.0   0.0   0.0   1.0   0.0
0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
```

```
0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
0.0   0.0   0.0   0.0   0.0   0.0   0.0   0.0
0.0   0.0   0.0   0.0   0.0   0.0   0.0   1.0
```

Fig : Matrix M & L

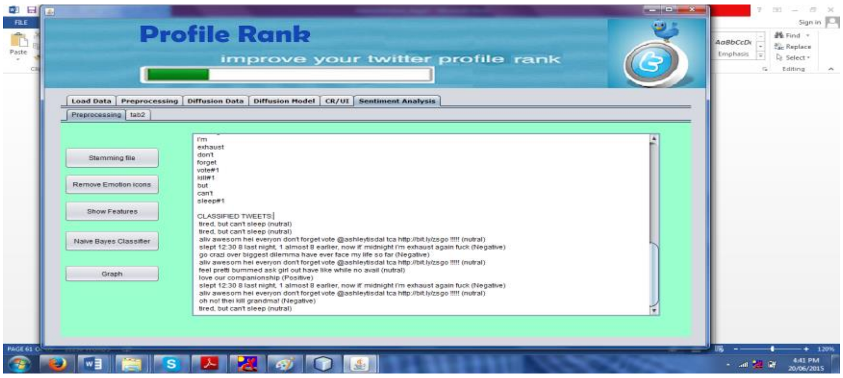
5. From this matrices and using formulae we can find a data relevance and use influence value for each tweet and user respectively. From this we can get a profile ranking.



```
Relevance of tweets:

r(0) : 0.1275
r(1) : 0.019125000000000003
r(2) : 0.0028687500000000001
r(3) : 4.3031250000000025E-4
r(4) : 6.454687500000004E-5
r(5) : 9.682031250000007E-6
r(6) : 1.4523046875000012E-6
r(7) : 2.178457031250002E-7

Influence of users:

i(0) : 0.1275
i(1) : 0.019125000000000003
i(2) : 0.0028687500000000001
i(3) : 4.3031250000000025E-4
i(4) : 6.454687500000004E-5
i(5) : 9.682031250000007E-6
i(6) : 1.4523046875000012E-6
i(7) : 2.178457031250002E-7
i(8) : 3.267685546875004E-8
i(9) : 4.901528320312506E-9
```

Fig : Relevance & Influence value

6. Now as our proposed approach for profile ranking we do sentiment analysis for the input data so it also helps to categories profiles.



```
i'm
exhaust
don't
forget
vote#1
kill#1
but
can't
sleep#1

CLASSIFIED TWEETS:
tired, but can't sleep (nutral)
tired, but can't sleep (nutral)
aliv awesom hei everyon don't forget vote @ashleytisdal tca http://bit.ly/zsgo !!!!! (nutral)
slept 12:30 8 last night, 1 almost 8 earlier, now if midnight i'm exhaust again fuck (Negative)
go crazi over biggest dilemma have ever face my life so far (Negative)
aliv awesom hei everyon don't forget vote @ashleytisdal tca http://bit.ly/zsgo !!!!! (nutral)
feel pretti bummed ask girl out have like while no avail (nutral)
love our companionship (Positive)
slept 12:30 8 last night, 1 almost 8 earlier, now if midnight i'm exhaust again fuck (Negative)
aliv awesom hei everyon don't forget vote @ashleytisdal tca http://bit.ly/zsgo !!!!! (nutral)
oh no! thei kill grandma! (Negative)
tired, but can't sleep (nutral)
```

Fig: Sentiment Analysis

V. CONCLUSION

We have analyzed the tweets of top trending topics and reported on the temporal behavior of trending topics and user participation. We have presented a family of naive-bayes classifiers for detecting the polarity of English tweets. The experiments have shown that the best performance is achieved by using a binary classifier trained to detect just three categories: positive, negative and neutral. We then classify the trending topics based on the senti words. A closer look at retweets reveals that any retweeted tweet is to reach an average of 1;. Once retweeted, a tweet gets retweeted almost instantly on the 2nd, 3rd, and 4th hops away from the source,signifying fast diffusion of information after the 1st retweet. And that tweet get a more data relevance. And the user that retweeted that tweet more times has greater user influence value and get highest rank accordingly. Based on an analysis of the results, we showed that relevant content and influential users discovered by Profile-Rank provide valuable knowledge in the analysis of information diffusion.

REFERENCES

[1] Arlei Silva, Sara Guimares, Wagner Meira, Jr., Mohammed ZakiProfile-Rank: finding relevant content and influential users based on information diffusion,2013.

[2] N. Friedkin. A structural theory of social influence, volume 13. Cam- bridge University Press, 2006.

[3] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In ICWSM, 2010

[4] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In WWW, 2010.

[5] A. Silva, H. Valiati, S. Guimares, and W. Meira Jr. From individual behavior to influence networks: A case study on twitter. In Webmedia, 2011.

[6] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, Sentiment Analysis of Twitter Data, In Proceedings of the ACL 2011 Workshop on Languages in Social Media,2011 , pp. 3038

[7] Alekh Agarwal and Pushpak Bhattacharyya, Sentiment analysis: A new approach for effective use of linguistic knowledge and exploiting similarities in a set of documents to be classified, In Proceedings of the International Conference on Natural Language Processing (ICON), 2005.

[8] F. Ricci, L. Rokach, and B. Shapira. Introduction to recommender systems handbook. Recommender Systems Handbook, 2011.

[9] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large- scale networks. In KDD, 2009.

[10] Antonio Fernandez Anta, Luis N unez Chiroque, Philippe Morere, and Agustn Santos. 2013. Sentiment Analysis and Topic Detection of Span- ish Tweets: A Comparative Study of NLP Techniques. Procesamiento del Lenguaje Natural, 50:4552.

[11] Stefano Baccianella, AndreaEsuli,and FabrizioSebastiani.2010. Senti-WordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In Human Language Technology Confer- ence - North American chapter of the Association for Computational Linguistics, pages 22002204.

[12] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: Finding topic- sensitive influential twitterers. In WSDM, 2010.

[13] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: experiments on recommending content from information streams. In CHI, 2010.

[14] F. Alkemade and C. Castaldi, "Strategies for the diffusion of innovations on social networks," Comput. Economics, vol. 25, no. 1-2, pp. 3–23, 2005.

[15] M. De Choudhury, S. Counts, and M. Czerwinski. Identifying relevant social media content: leveraging information diversity and user cognition. In HT, 2011.

[16] J.Wortman, "Viral marketing and the diffusion of trends on social networks," University of Pennsylvania, Tech. Rep. Technical Report MS- CIS-08-19, May 2008