International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 19 (2015) pp 39956-39962

Improving User Hypothesis Using Locational-Social-Tropical Prediction Framework

J. S. Kanchana

Associate Professor, Department of IT, K. L. N College of Engineering, Sivagangai District, Tamil Nadu, India

Dr. S. Sujatha

Assistant Professor, Department of MCA, Anna University Regional Centre, BIT Campus, Trichy, Tamil Nadu, 620 024, India

Abstract
The advent of social networks, web communities and blogs changed the ways of users to express their sentiments and opinions. The marketers identify the knowledge from the huge amount of unstructured information to create an image or identity in the customer minds for their product or brand. In order to predict the customer minds sentiment analysis is needed. Opinions are key influencer's of user behaviors. The user beliefs and perceptions of reality are based on how others see the world. Whenever the user needs to take a decision, they often seek out the opinions of others. Microblogging services like Twitter, Google Buzz users are sharing their information in a more convenient way. Twitter uses a follower and following social structure to subscribe their friends feeds. Twitter datasets is thus regarded as a data to correlate the topics with emotion analysis and learning. The observed tweets and topic context information to predict the user's opinion toward specific topics had not directly given yet, is a challenge. In this paper, the proposed Locational-Social-Topical (LST) prediction framework process the opinion of users and predicts them based on locational, social and topical factors. The locational context along with the topical or social factors influences on the user emotions for better prediction of user opinion. The location factor in turn influences the social as well as topical factors.

Keywords: Sentiments, Opinion Mining, Microblogging, LST framework.

Introduction

The world is moving in a warp speed and people are travelling in it-Social Network. The Social Network is gaining a critical mass of users where they feel very comfortable to share their emotions, information with all others. Most of the users use micro blogs (short text messages) to share their information. Micro blogging provides informational content for the marketers and also researchers for analyzing things. The problem here is due to limited size, people use short form of text. Extracting information from such text is a difficult task. Opinion mining plays a major role in extracting and analyzing the content from such micro blogs. For analyzing the reviews of a particular product, it is necessary to correlate the blog text to the topic.

Topic Correlation relates the micro blog to a particular topic. Hash tags play a major role in identifying the topics and also in identifying the people who are having similar thoughts on the same topic. This identification of similar people will help the marketers to recommend for their product. Topic correlation is not the only factor that helps in recommendation. As everyone knows the fact "A man is known by the company he keeps", the relationship among two people can influence the recommendation. If two people are friends, then their thoughts may also be similar. Social networks help to find the social relationship between people. Using the topical and social information, the similar minded people can be predicted which further helps in recommendation system. The most popular microblogging service is Twitter. Tweets in twitter helps to identify the thoughts of other people and has attracted many researches to do work in it.

The information is crucial and when they do share it is entering in a public zone where they comment about others. The information is categorized into two ie, Location Context, Social Context and Topical Context. The Location context is The Social context is information where it can contain (URL, text) ie, it is simply a tweet (we consider with Twitter). The Topical context is most important and it is identified here with the help of hashtag concept introduced by Twitter. The (#) Hashtags plays major role in identifying the topics in the particular tweet. The goal is to identify the likeminded users who are around the particular location, topic and social context with relating social relationship of a user.

The current popular micro blogging service twitter provides efficient and communicating platform used to analyze the emotions, sentiment and opinion of the global users. In earlier work, the traditional long text tweeter data were processed by state of art sentiment analysis and opinion mining methods. Go et al. used distant supervision approach that automatically classify the tweet messages as either positive or negative [2]. Davidov et al. proposed a framework based on sentiment classification in micro blogs using different feature types [1]. In Jiang et al. 's paper [7], incorporated target-dependent features and considered the relationships between tweets to improve tweet sentiment classification. Then they used small size dataset as a test data and their results are at tweet level. Hu et al. Proposeda supervised learning method to find whether social relations can be used for sentiment analysis [10]. Tan et al. In his paper, exploiting social network information in twitter based user-level sentiment analysis model [3]. They collected data from groups of extremely similar opinion users to evaluate their model.

Proposed work

The proposed system uses three different factors such as social, topical and location factors for predicting the opinion of a user. The social factors deal with the social friend relationship among the users. The topical factors deal with the user's view about a particular topic. The location factors deal with whether the user's location affects their opinion in any way.

Location factor: The concept of location factor is that if two users belong to a same location, then there may be chances for them to have a similar view on the particular topic. During preprocessing step, the user belonging to the same location is given weightage. The location information of a user is gained using google maps (longitude and latitude information). Based on this, if two users location is within a certain threshold value (approximately 200m), then the matrix is formed by given a weightage of 1. For others, the matrix value is calculated using the formula LI as given in equation (1).

$$L(I) = 1 - \frac{d_u}{\Omega * 10} \quad (1)$$

Where L(I) is the location similarity between two users u_i and u_j, d_u is the distance between two users and Ω is the threshold value.

Social factor: The concept of social factor is that if two users have a relationship then, there may be chances for them to have similar opinions. The social factor matrix is formed by mapping the user and their friends. The similarity between users based on this social contextual is formulated using cosine similarity.

$$R_s(u_i, u_j) = \frac{\sum_{k=1}^{n} SO_{ik}.SO_{jk}}{\sqrt{\sum_{k=1}^{n} SO_{ik}^2}\sqrt{\sum_{k=1}^{n} SO_{jk}^2}} \quad (2)$$

where $R_s(u_i, u_j)$ is the social relationship between two users u_i and u_j, SO_{ik} and SO_{jk} represents the opinion similarity of user u_i on u_k and user u_j on u_k respectively. Based on this relationship value, the similarity matrix is formed by

$$Sim_s = \begin{cases} R_s(i,j), & if i \wedge jare friends \\ o, & otherwise \end{cases} \quad (3)$$

Topical Factor: The concept of topical factor is that if two users have similar opinion on some topics, then they may have a possibility to have similar opinions on some other topics too. The topical factor matrix is formed by mapping the user and topics.

$$R_t(t_i, t_j) = \frac{\sum_{k=1}^{n} Tf_{ik}.Tf_{jk}}{\sqrt{\sum_{k=1}^{n} Tf_{ik}^2}\sqrt{\sum_{k=1}^{n} Tf_{jk}^2}} \quad (4)$$

where $R_t(t_i, t_j)$ is the topical relationship among two topics, Tf represents the term frequency among topics.

$$Sim_t = \begin{cases} R_t(i,j), & if i \neq j \\ o, & otherwise \end{cases} \quad (5)$$

In this proposed system the topical, social and location information are obtained through the users. After the information is gathered, preprocessing is done and the topical and social information based matrix is formed. These two matrices are given weightage with the location factor. After preprocessing, the matrix factorization is formed by

$$MF = Sim_{LS}.Sim_{LT} \quad (6)$$

where Sim_{LS} and Sim_{LT} are the location based social similarity matrix and location based topical similarity matrices respectively.

Algorithm:

Input: User-topic matrix before prediction

Output: User-topic matrix after prediction

1. Calculate similarity for social matrix using equation 2.
2. Form the social matrix S using equation 3
3. Calculate similarity for topical matrix using equation 4.
4. Form the topical matrix T using equation 5.
5. Calculate the location matrix using equation 1.
6. Compute the value of P using equation 6.

Experimental Results

We have created our datasets using Google web forms of 26 users and tweets. Although we have extracted the data with most appropriate hashtags like #google, #apple and etc. We also extracted smileys ie, emotions to analyse user feeling about that particular topic. We have created a simple database with following data to analyse the tweets in future with location update.

The topical context uses #hashtags and clearly exaggerates the relationship between users tweet with topics. With user topic ie, Hashtags and user tweet we generate the matrix (relationship of tweets and user). The matrix generated is factorized using Matrix Factorization method. The Factorization method decomposes the matrix value more efficiently.

The test environment we used is MATLAB. In this section, the sample of 5 users and their information is analyzed.

The users and their corresponding rating is as shown below.

$$BP = \begin{bmatrix} 0.1 & 0.8 & 0.9 & 0.6 & 0.8 \\ 0.3 & 0.3 & 0.7 & 0.9 & 0.9 \\ 0.7 & 0.7 & 0.8 & 0.2 & 0.2 \\ 0.5 & 0.6 & 0.5 & 0.1 & 0.6 \\ 0.9 & 0.8 & 0.1 & 0.7 & 0.4 \end{bmatrix}$$

The user who has initiated the topic is given with a value of 1. The other user's views are normalized based on that.

For predicting the opinion of user 1 on topic t_2, we perform the following strategies.

3. 1. Social and Topical Information

The social information matrix for 5 users are formulated using the cosine similarity S is as shown below.

S=

$$\begin{bmatrix} 1 & 0.45 & 0.753 & 0.444 & 0.455 \\ 0.45 & 1 & 0.594 & 0.962 & 0.559 \\ 0.753 & 0.594 & 1 & 0.727 & 0.494 \\ 0.444 & 0.962 & 0.727 & 1 & 0.533 \\ 0.455 & 0.539 & 0.494 & 0.533 & 1 \end{bmatrix}$$

The topical information matrix for 5 users are formulated using the cosine similarity T is as shown below.

T=

$$\begin{bmatrix} 1 & 0 & 0.5527 & 0.6263 & 0.5377 \\ 0 & 1 & 0.7247 & 0.6922 & 0.4490 \\ 0.5527 & 0.7247 & 1 & 0.8760 & 0.7072 \\ 0.6263 & 0.6922 & 0.8760 & 1 & 0.5555 \\ 0.5377 & 0.4490 & 0.7072 & 0.5555 & 1 \end{bmatrix}$$

The prediction matrix (P) is calculated for social and topical information using matrix factorization as

$P_{st} = S * T$

$$\mathbf{P}_{st} = \begin{bmatrix} 0.93 & 0.97 & 0.90 & 0.94 & 0.91 \\ 0.9 & 0.93 & 0.84 & 0.9 & 0.87 \\ 0.77 & 0.82 & 0.86 & 0.89 & 0.88 \\ 0.91 & 0.95 & 0.95 & 1 & 0.97 \\ 0.80 & 0.85 & 0.87 & 0.92 & 0.90 \end{bmatrix}$$

3. 2. Locational, Social & Topical Information

With the social and topical information, the location information is also introduced. The location information is embedded into the social and topical information and the matrix is formulated.

$$L = \begin{bmatrix} 1 & 0.55 & 0.78 & 0.58 & 0.85 \\ 0.55 & 1 & 0.8 & 0.55 & 0.83 \\ 0.78 & 0.8 & 1 & 0.75 & 0.9 \\ 0.58 & 0.55 & 0.75 & 1 & 0.65 \\ 0.85 & 0.83 & 0.9 & 0.65 & 1 \end{bmatrix}$$

The prediction matrix is formulated as **$P_{lst} = L* S * T$**

$$\mathbf{P}_{lst} = \begin{bmatrix} 0.88 & 0.87 & 1 & 0.92 & 0.89 \\ 0.84 & 0.83 & 0.95 & 0.88 & 0.86 \\ 0.79 & 0.79 & 0.91 & 0.86 & 0.82 \\ 0.89 & 0.89 & 0.97 & 0.95 & 0.93 \\ 0.81 & 0.81 & 0.93 & 0.87 & 0.84 \end{bmatrix}.$$

The RMSE value is calculated as

$$RMSE = \sqrt{\frac{1}{N}\sum_{i,j}\left(O_{ij} - \hat{O_{ij}}\right)^2}$$

Table 1: Comparison of RMSE values for ST and LST information

Topic	RMSE for ST	RMSE for LST
T1	0. 493	0. 461
T2	0. 439	0. 375
T3	0. 506	0. 447
T4	0. 527	0. 506
T5	0. 414	0. 371

Table 1 contains three columns which represents the topic, RMSE values for Social-topical information and RMSE values for Locational-Social-topical information

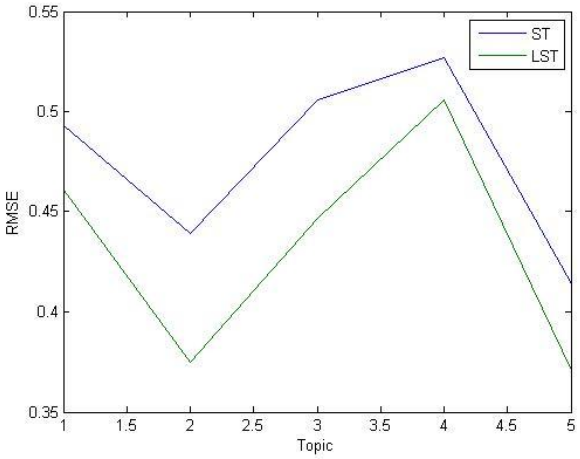

Fig. 1. RMSE Graph for ST and LST Information

Fig 1 represents the RMSE graph for Social Topical and Locational Social Topical Information where the X-axis represents the RMSE values and the Y-axis about the topic values.

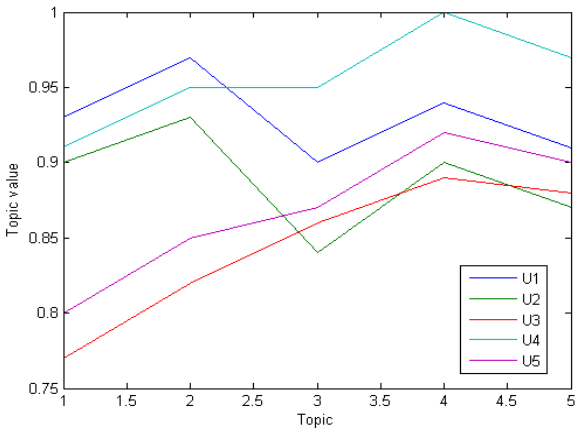

Fig. 2. User Topic Opinion using Social and Topical Information.

Fig 2 represents the User Topic Opinion using Social and Topical Information where the X-axis represents the Topic values and the Y-axis about the topic.

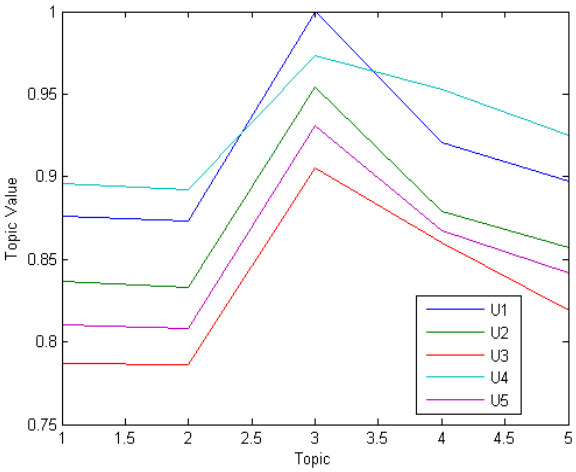

Fig. 3. User Topic Opinion using Locational, Social and Topical Information

Fig 3 represents the User Topic Opinion using Locational, Social and Topical Information where the X-axis represents the Topic values and the Y-axis about the topic.

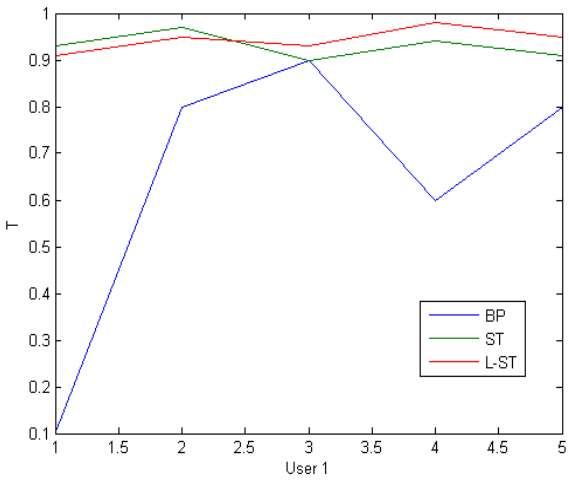

Fig. 4. User 1 opinion with all topic information

Fig 4 represents the User1 Opinion with all Topical Information where the X-axis represents the user1 values and the Y-axis about the topic

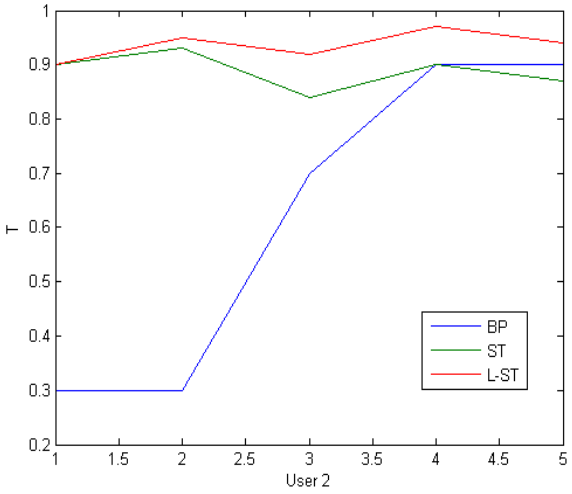

Fig. 5. User 2 opinion with all topic information

Fig 5 represents the User2 Opinion with all Topical Information where the X-axis represents the user2 values and the Y-axis about the topic

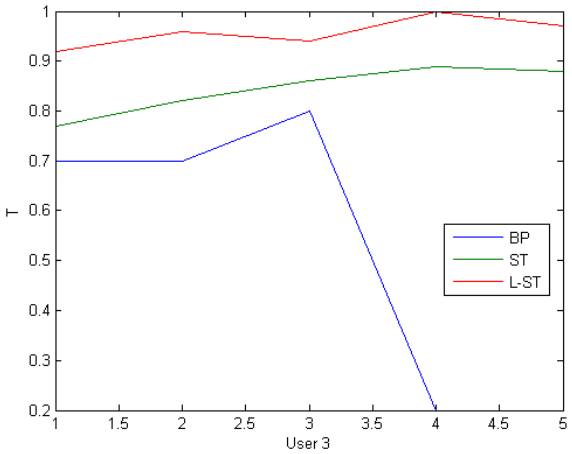

Fig. 6. User 3 opinion with all topic information

Fig 6 represents the User 3 Opinion with all Topical Information where the X-axis represents the user 3 values and the Y-axis about the topic.

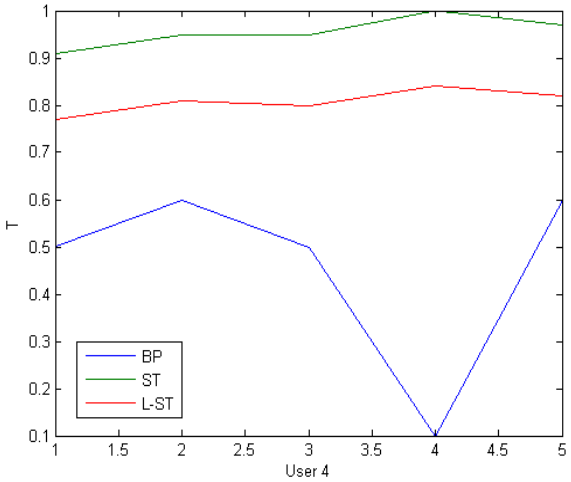

Fig. 7. User 4 opinion with all topic information

Fig 7 represents the User4 Opinion with all Topical Information where the X-axis represents the user4 values and the Y-axis about the topic.

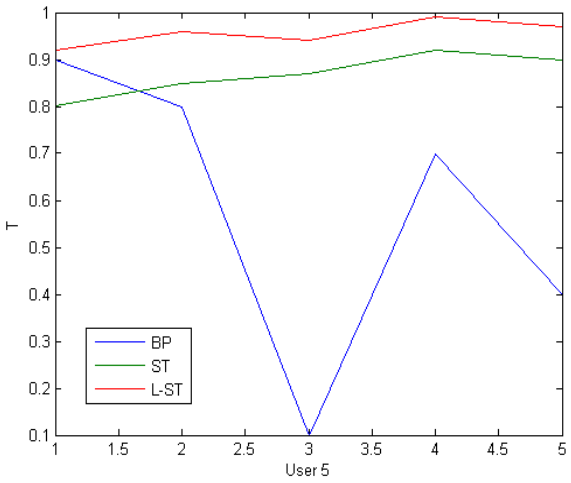

Fig. 8. User 5 opinion with all topic information

Fig 8 represents the User5 Opinion with all Topical Information where the X-axis represents the user5 values and the Y-axis about the topic.

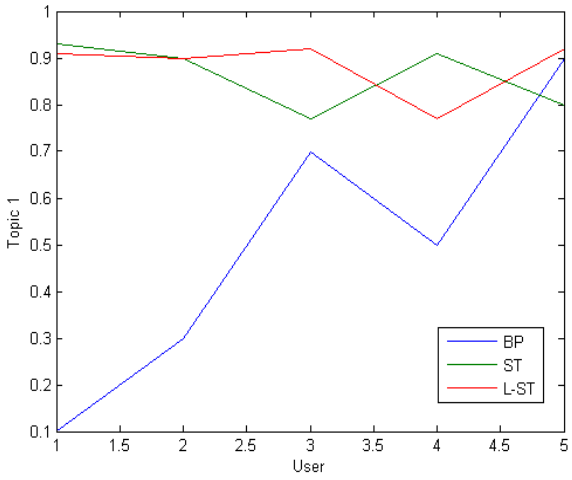

Fig. 9. All Users opinion with topic information 1

Fig 9 represents all User Opinion with all Topical Information where the X-axis represents the user values and the Y-axis about the topic1 value.

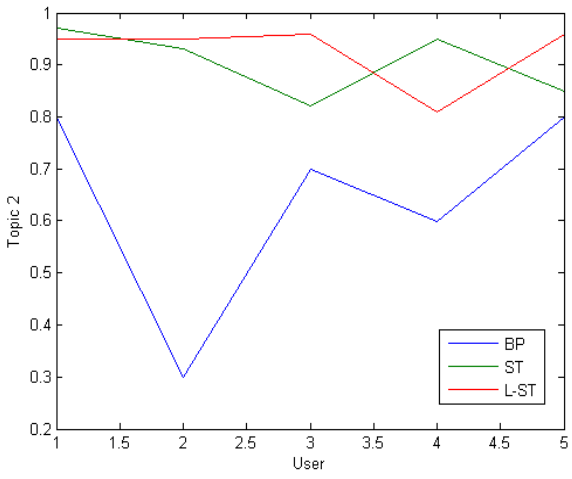

Fig. 10. All Users opinion with topic information 2

Fig 10 represents all User Opinion with all Topical Information where the X-axis represents the user values and the Y-axis about the topic2 values.

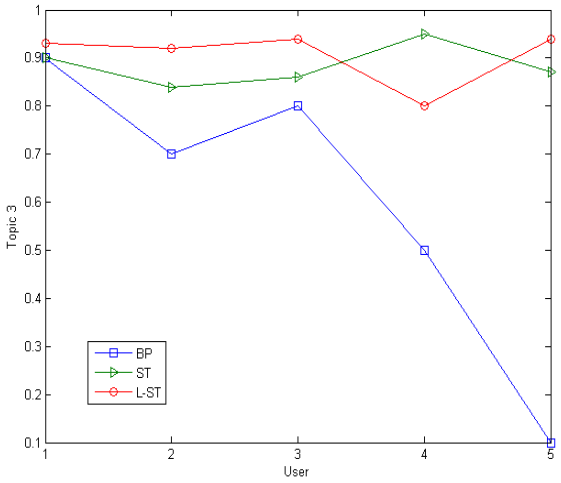

Fig. 11. All Users opinion with topic information 3

Fig 11 represents the User Opinion with Topical Information where the X-axis represents the user values and the Y-axis about the topic3 value

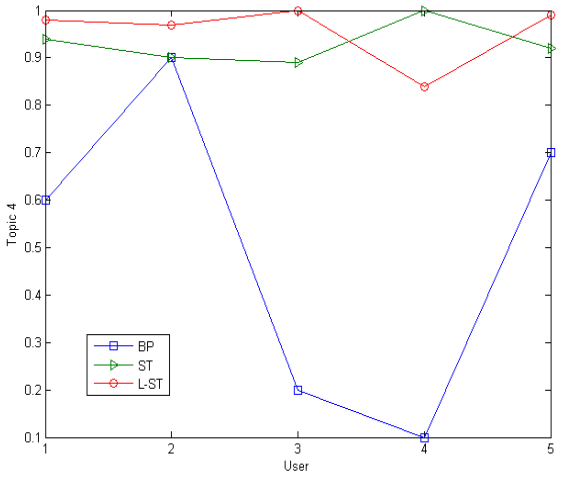

Fig. 12. All Users opinion with topic information 4

Fig 12 represents the User Opinion with Topical Information where the X-axis represents the user values and the Y-axis about the topic4 values

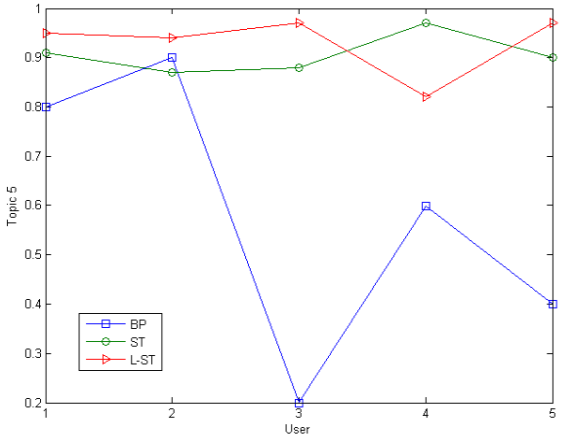

Fig. 13. All Users opinion with topic information5

Fig 13 represents the User Opinion with Topical Information where the X-axis represents the user values and the Y-axis about the topic information5 values

Conclusion

The LST prediction framework uses location context involving social and topical factors improves the prediction of users opinion. Location context has a greater influence on the user's emotions. The result of the LST framework provides better prediction on user opinion when compared to the prediction involving only social and topical factors.

References:

[1] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys, " in Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 2010, pp. 241-249.

[2] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision, " Tech. Rep., 2009

[3] J. Tang, H. Gao, X. Hu, and H. Liu, "Exploiting homophily effect for trust prediction, " in Proceedings of the sixth ACM international conference on Web search and data mining, 2013.

[4] B. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth, " Journal of the American society for information science and technology, vol. 60, no. 11, pp. 2169-2188, 2009.

[5] X. Hu, L. Tang, J. Tang, and H. Liu, "Exploiting social relations for sentiment analysis in microblogging, " in Proceedings of the sixth ACM international conference on Web search and data mining, 2013.

[6] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text, " Journal of the American Society for Information Science and Technology, vol. 61, no. 12, pp. 2544-2558, 2010.

[7] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification, " in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, 2011, pp. 151-160.

[8] B. Pang and L. Lee, "Sentiment analysis and opinion mining, " Foundations and Trends in Information Retrieval, no. 1, pp. 1-135, 2008.

[9] J. Bollen, B. Gonc¸alves, G. Ruan, and H. Mao, "Happiness is assortative in online social networks, " *Artificial Life*, vol. 17, no. 3, pp. 237-251, 2011.

[10] X. Hu, L. Tang, J. Tang, and H. Liu, "Exploiting social relations for sentiment analysis in microblogging, " in *Proceedings of the sixthACM international conference on Web search and data mining*, 2013.

[11] M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging topic detection on twitter based on temporal and social terms evaluation, " in *Proceedings of the Tenth International Workshop on Multimedia DataMining*. ACM, 2010, p. 4.

[12] R. Zafarani, W. D. Cole, and H. Liu, "Sentiment propagation in social networks: A case study in livejournal, " in *Advances in SocialComputing*. Springer, 2010, pp. 413-420.

[13] Q. Gu, J. Zhou, and C. Ding, "Collaborative filtering: Weighted nonnegative matrix factorization incorporating user and item graphs, " in *SIAM SDM*, 2010, pp. 199-210.

[14] Fuji Ren, *Senior Member, IEEE* and Ye Wu, " Predicting User-topic Opinions in Twitter with Social and Topical Context" in IEEE Transactions on Affective Computing, Vol. 4, Issue 4. 2014.

[15] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining, " in *Proceedings of LREC*, vol. 2010.

[16] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social web, " *Journal of the American Society forInformation Science and Technology*, vol. 63, pp. 163-173, 2012.