ANALYSIS OF SOCIAL RELATIONSHIPS ON TWITTER DATA FOR ONLINE ANALYTICAL PROCESSING

[1]Sana, [2] CH. B. N. Lakshmi
[1] M. Tech Scholar, [2] Professor,
[1, 2] Department of Computer Science and Engineering,
[1, 2] TKR College of Engineering and Technology, Hyderabad, India.

Abstract: Social media platforms, for instance as twitter has become increasingly popular as an emerging platform for messaging and communication. On the other side, online meticulous dealing with multidimensional sorted out data. The purpose of using Hierarchical topic modeling i.e., THLDA to mine the dimension hierarchy of tweets. It can be applied for text OLAP on the tweets. And it uses word2vec to analyze the semantic relationships of words in tweets to obtain more effective dimensions. To improve the model effectiveness the bicliques to calculate the semantic impact of the topic of two tweets. Focusing on how the social impact factors and word semantic similarity influence the experimental results separately. In this we are considering how social relationships impact on the hierarchical topic model. In social relationships we are focusing direct and indirect relationships to follow the unrelated tweeters. We conduct extensive experiments on real twitter data to evaluate to effectiveness of THLDA. We are using hash tags to improve the models.

Index Terms – Social Media Relationship, Bicliques, Semantic Relationship, Online Analytical Processing, Hierarchical Latent Dirichlet Allocation(THLDA).

I. INTRODUCTION

During past few years, Twitter has become increasingly popular as an social networking platform on which users post and communicate with messages .the huge amount of twitter data gather so far make it possible To find the distribution and drift of mass choice and opinions, which greatly assists in product recommendation, target marketing. the OLAP(online analytical processing) interactively view data from all aspects in layered graininess, which has already been proven especially usefull for business intelligence. so successfully apply OLAP techniques to twitter, it is difficult to find the hidden dimensions from its extensive data. so we using latent dirichlet allocation(THLDA) model we analyzing textual data from unstructured content. we focus on how to discover the underlying topics of tweets from tweeters and social relationships mainly direct and indirect social relationship and from their published tweets.

II. LITERATURE SURVEY

[1] S.ChaudhuriandU.Dayal, author developed OLAP is an approach to answering multidimensional analytical queries over the cube data. It provides the operations such as rolling up, drilling down and slicing the dilemma of the physical architecture of data warehouses should draw attention to the well-known issues of index placement, data partitioning and materialised view placement. Data storage control poses new problems as well. Detecting runaway issues, administration and resource preparation are critical but not well-resolved challenges.

[2] M. Michelson and S. A. Macskassy, ‘‘Discovering users’ topics of interest on Twitter: How to accurately and effectively mine tweets’ topics from social data has long been the focus of research in the field of natural language processing. For example Michelson and Macskassy et al. present a topic profile to characterize tweets’ topics. We propose an approach to detect temporally active and dense communities, making use of the biased density metric and the influence of active users with the frequency of their interactions with the neighbourhood.

[3] A. Cuzzocrea, C. De Maio, G. Fenza, V. Loia, and M. Parente, ‘‘OLAP analysis of multidimensional tweet streams for supporting advanced analytics, introduce an aggregation operator for tweets’ content by

using formal concept analysis theory. We applied the OLAP technology to enable comprehensive analysis of unstructured data generated by social networks. We have proposed a new multidimensional model (unified multidimensional social media data model) dedicated to the storage of social media unstructred data streams to enable OLAP analysis.

[4] X. Liu et al., ‘‘A text cube approach to human, social and cultural behavior in the Twitter stream,the main purpose of Using linguistic features and the text cube approach for the HSCB dimension of sentiment appears to be quite promising. In particular, the use of the text cube provides a useful way to explore and analyze complex data, and in particular to connect language patterns to potential HSCB dimensions. In the present paper we focused on the HSCB dimension of sentiment. While sentiment analysis in the literature has gained significant attention, we think it is new to apply the method to the text cube in this context.

[5] N. U. Rehman, A. Weiler, and M. H. SchollSocial networks are channels in this paper where millions of people regularly connect and exchange digital content. Users share their thoughts and beliefs on all matters of concern. These opinions have import importance for private, academic and commercial purposes, but it is a challenge for the researchers and the underlying technologies to offer meaningful insights into these data due to the amount and speed at which they are generated. We attempt, by implementing text and opinion mining methods into the data storage system, to expand the current OLAP (Online Analytical Processing) technology to render multidimensional analysis of social media data through knowledge discovery techniques to deal with semi-structured.

[6] E. Siswanto, M. L. Khodra, and L. J. E. Dewi, Numerous studies have been conducted toExplore the social network of Twitter; some have been conducted to predict the interestor the topic of the user's tweet. In this study, we investigate the best classification model fordetermining the user's interest based on the bio and a collection of tweets. We use thesupervised learning-based classification with the lexical features.Two approaches wereproposed; they are the classification that was made based on the user's tweet using multilabelmethod and the classification that was made based on specific accounts. From the result ofexperimental result, it could be concluded that the employment of the classification using specific accounts approach led to better accuracy.

III. PROPOSED METHODOLOGY

In this proposed methodology we are analyzing social relationship that are direct and indirect social relationship between tweeters to enhance the model.and we are using bicliques to calculate the semantic impacts of the topic of two tweets and to improve model effectiveness and we focus on how the social relationships and word semantic similarity influence the experimental results and we are also using hashtags to improve our model

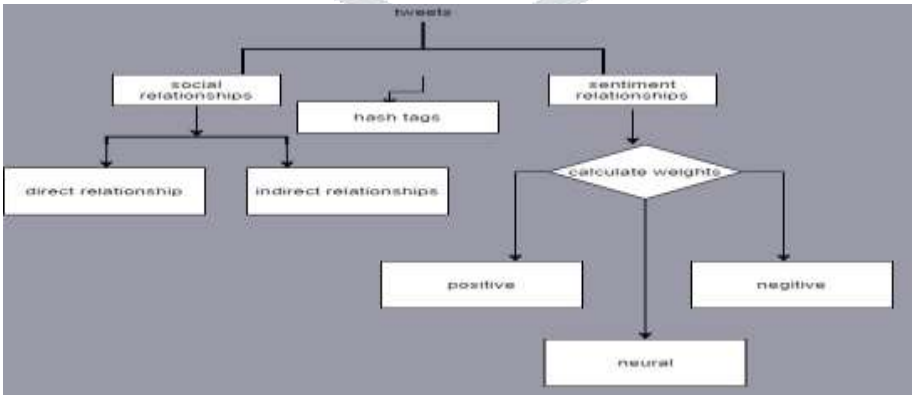


Figure 1.Whole process of twitter data and their techniques.

A. Twitter Data

Twitter have two entities,such as tweets and tweeters.the tweets have the content published by the tweeters and the tweeters have their own properties.on the other side twitter data can also be divided into two parts

such as structured and unstructured data.the structure data require additional preprocessing for olap.however ,the unstructured data require special treatment for olap.

B. Social Relationships

We have to find social relationship between twitter. So we focus indirect and indirect social relationships.

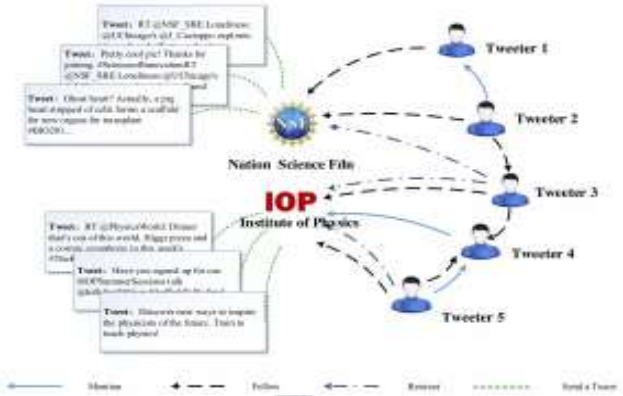


Figure 2.social relationships between twitters.

Figure 2 shows social relationships among tweeters the tweeters find social behaviors, following, and mentioning, retweeting. in the fig, intitute of physics sends a tweets, tweeters 3 and 5, who follow it, and it receive a notification, and many retweet the tweet if they are interested in it. meanwhile ,tweeter 5 can also mention it to his friend, tweeter4, when sending and retweeting tweets.and also retweet the suggested friend i.e indirect relationship.

C. Latent dirchlet allocation

The LDA is a probability model for idea-based discrete data sets.in each document is a mixture of multi topics.the LDA is athree level hierarchical bayes model.the collection level,document level,word level in which each part of the model is treated as a finite mixing model based on the set of the topic probabilities.a document is probability distribution of different topics.

D. Word2vec

We are using algorithm i.e word2vec algorithm.that are used to produce word embeddings.the algorithm is a neural network model and it is trained to reconstruct the linguistic contexts of words.this algorithm takes as its input a large collection of text and produces a vectorspace,typically several hundred dimensions.we have calculate term-frequency or inverse document frequency(TF/IDF) to find similarity between two tweets.

For example:
tweet 1: An apple a day keep doctor away
tweet 2:Apple is good for health
Apple day kee doctor away good health(unique words from 2 tweets)
T1 1 1 1 1 1 0 0
T2 2 0 0 0 0 1 1
TF/IDF of word apple in tweet1=1/2=0.5

E. Sentiment Analysis

Sentiment analysis is process of find the opinions, attitudes and emotions expressed in a piece of text.the main purpose of sentiment analysis are to detect and extract information from the tweets.it detect the tweets has positive, neural or negative. Sentiment analysis is two types

1. Lexicon-based methods
2. Machine learning-based method

In this paper we are using lexicon-based method and its calculates sentiment of a given text from the words or phrases for this method a lexicon(dictionary) of words with assigned to them is required.

IV. EXPERIMENT

A. Data and Environment

We have to verify the effectiveness and efficiency of our model, we conducted extensive experiments on large amount of twitter data collected through the twitter app. first we choose twitter users with the large quantity of attention seeds and obtained all tweeters who followed the seeds, view their profiles, tweets, and social relationships .the number of tweets we removed the short tweets of less than 6 words, since we think such tweets generally have no clear semantics and also we removed duplicated tweets.

Algorithm 1 :modeling process of THLDA
Input: TDC - The set of Twitter document;
α, β, γ - hyperparameters;
L - the height of topic tree;
I - the iteration number of Gibbs sampling;
Output: TopicTree;
1: // Associate topic with node based on Dirichlet dist
2: **for** each t ∈ TopicTree **do**
3: draw a Dirichlet Process ϕ ~ Dir(β);
4: **end for**
5: // Generate a path for TweetDocm based on nCRP
6: **for** each TweetDocm ∈ TDC **do**
7: let c1 be the root node;
8: **for** each level l ∈ 1, 2, . . . , L **do**
9: draw the current level for each Tweetm,s ;
10: draw a occupied path cl using Eq. (5);
11: draw a unoccupied path cl using Eq. (6);
12: **end for**
13: obtain cm;
14: draw a L-dim. topic proportion vector θm from Dir(α);
15: **for** i = 1 to **I do**
16: **for** each word w ∈ W **do**
17: draw topic z ∈ 1, 2, . . . , L from Mult(θ);
18: draw w from the topic z;
19: **end for**
20: **end for**
21: **end for**
22: **return** TopicTree;

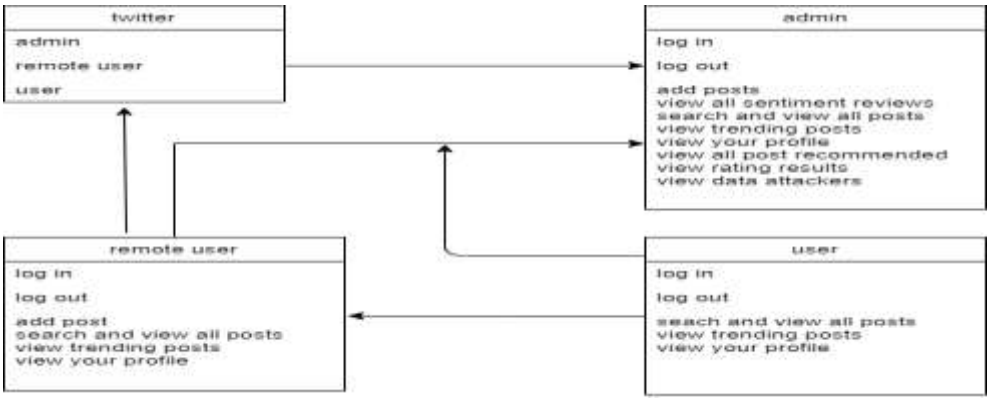


Figure 3.Galaxy Schema for Twitter Data

In figure 3.the tweeter have three platforms admin, user and remote user. The admin and remote user have a direct relationship they both can view their tweets. Same as user and remote user have direct relationship but the admin and user have indirect relationship. The user can't see the admin tweets without taking recommendation from remote user.

B. Hashtags

We are using hashtags to improve our model .mainly hashtags are used to discover tweets around those specific toipics,because hashtags aggregate all social media data with the same hashtag.in this model we are using hashtags to find the trending tweets that are usefull for us.

Figure 4.View Trending Posts by Hashtag

C. Ratings

We are rating tweets in tweeter app. The rating will be show in pie charts and bar charts.

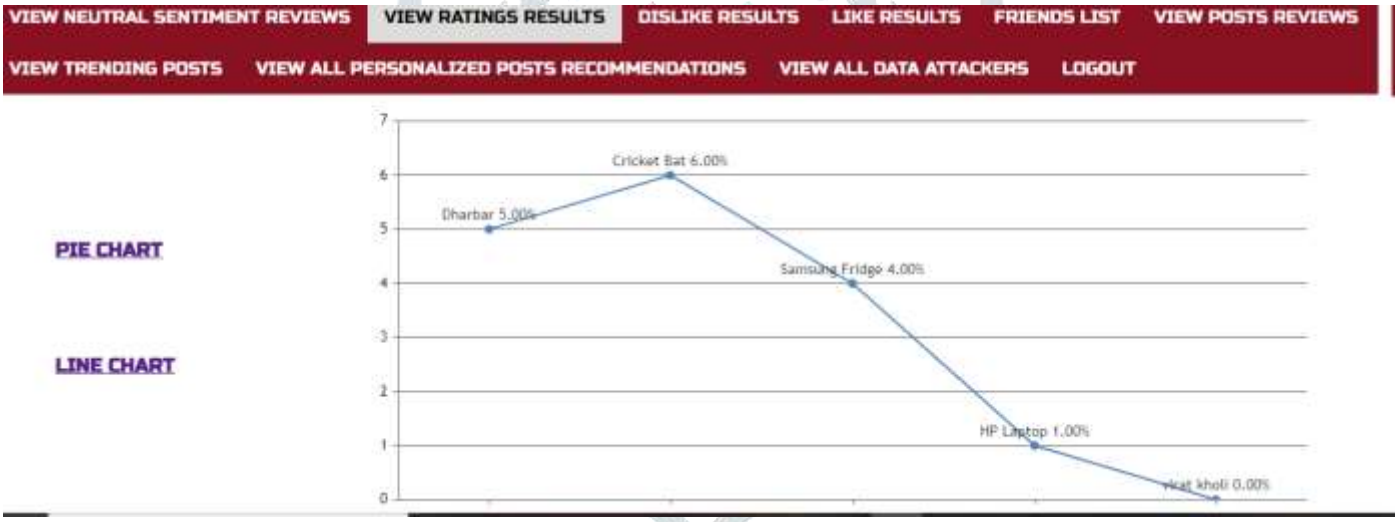


Figure 5.View Rating Results of Difference Tweets

D. Likes and dislikes

We have to see the likes and dislikes of the tweets in the twitter app. By using this we have to find the taste of the person.we showing dislikes in pie chart and bar charts

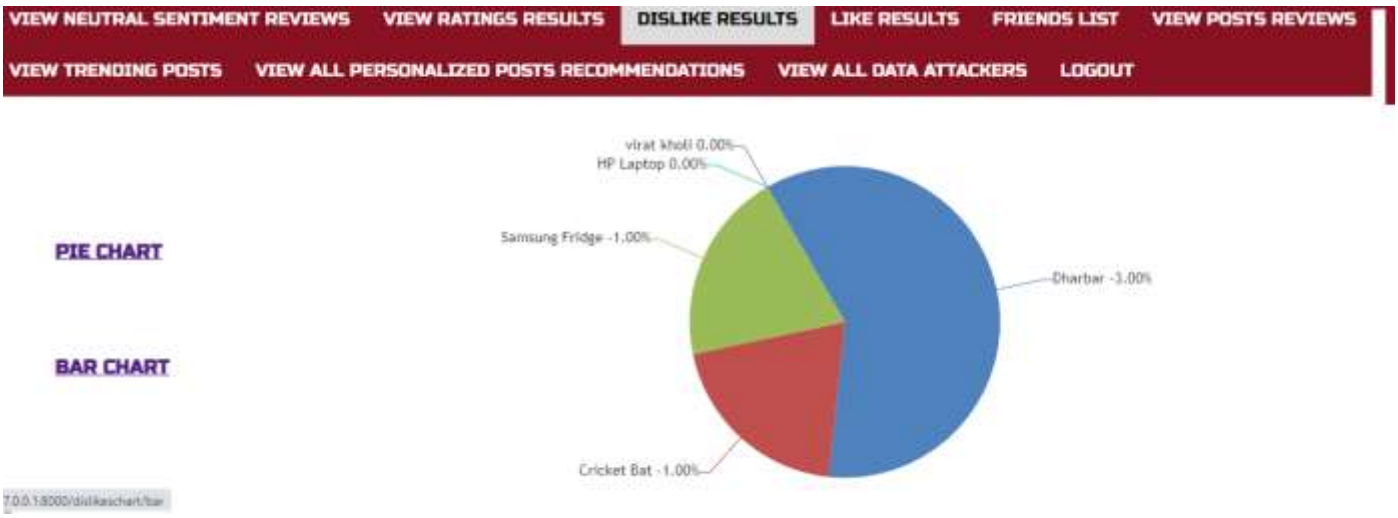
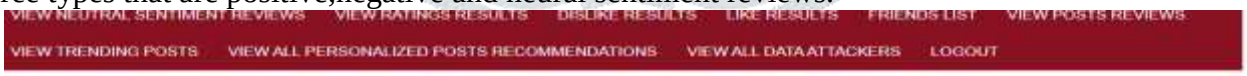


Figure 6.dislike results of tweets

Figure 7.Like Results of Tweets

E. Sentiment Reviews

We are using setiment reviews to find tweets reviews in the tweeter app. so the sentiment reviews are three types that are positive,negative and neural sentiment reviews.



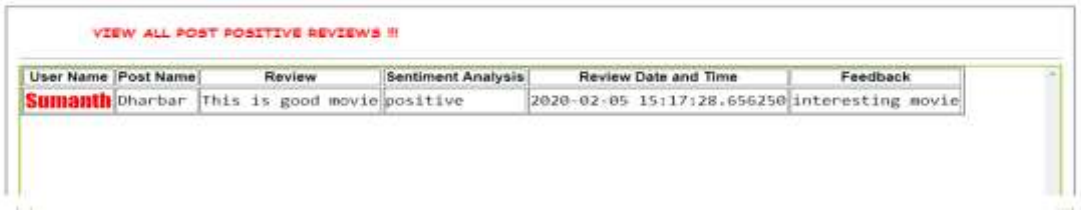


Figure 8.View All Positive Reviews of Tweets

V. RESULTS AND DISCUSSION

Here we have social networking service such as twitter on which remote user post and interact with messages i.e tweets. Registered users can upload, like and retweet messages, but users can only view them without registration. here we are post the tweet by using hashtag to view as a trending post and also we are focusing indirect relation for example user A and user B has a direct relationship they can see each other posts ,like and dislike post. same as user B and user C has direct relationship they see each other posts. but

user A and user C had indirect relationship they can't see each other post without taking recommendation from the user B.

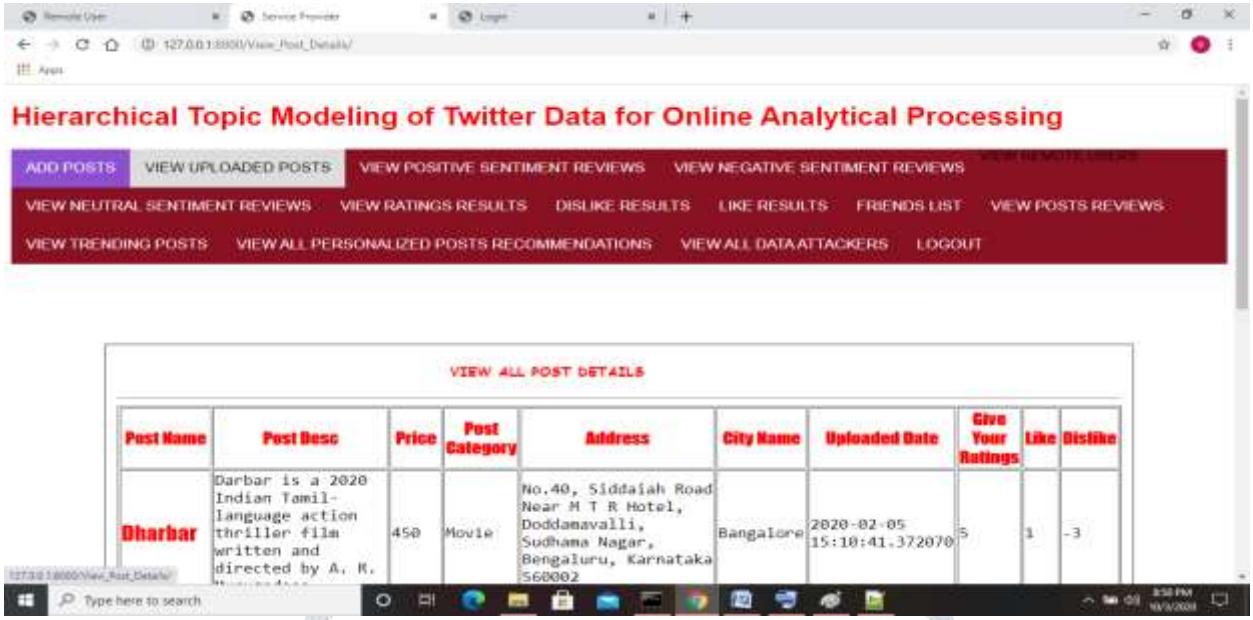


Figure 9.Results Of Viewing Uploaded Tweets

VI. CONCLUSION

In this paper we put a model i.e hierarchical topic model(thLDA),which is applied to find dimension hierarchy of tweets topics from quantity amount of unstructured twitter data. We focus on how social relationships impact on hierarchical topic model.we mainly focus on indirect social relationship.to improve our model effectiveness we consider bicliques to calculate the semantic relation impact on two tweets.and we focus on how word semantic similarity between tweets .and we are improving our model by using hashtags.

VII. REFERENCE

[1] D. Yu et al., ''Mining hidden interests from Twitter based on word similarity and social relationship for OLAP,'' Int. J. Softw. Eng. Knowl. Eng., vol. 27, nos. 9–10, pp. 1567–1578, 2017.

[2] D. Yu, J. Sun, Y. Wu, Z. Ni, and Y. Li, ''Discovering hidden interests from Twitter for multidimensional analysis,'' in Proc. 29th Int. Conf. Softw. Eng. Knowl. Eng., 2017, pp. 329–334.

[3] S. Chaudhuri and U. Dayal, ''An overview of data warehousing and OLAP technology,'' ACM SIGMOD Rec., vol. 26, no. 1, pp. 65–74, 1997.

[4] A. Inokuchi and K. Takeda, ''A method for online analytical processing of text data,'' in Proc. 16th ACM Conf. Conf. Inf. Knowl. Manage., 2007, pp. 455–464.

[5] F. Ravat, O. Teste, R. Tournier, and G. Zurfluh, ''Top_keyword: An aggregation function for textual document OLAP,'' in Proc. Int. Conf. Data Warehousing Knowl. Discovery. Berlin, Germany: Springer, 2008, pp. 55–64.

[6] C. X. Lin, B. Ding, J. Han, F. Zhu, and B. Zhao, ''Text cube: Computing IR measures for multidimensional text database analysis,'' in Proc. 8th IEEE Int. Conf. Data Mining (ICDM), Dec. 2008, pp. 905–910.