International Journal of Grid and Distributed Computing
Vol. 14, No. 01, (2021), pp.1868-1876

Opinion Mining and Web mining using Sentiment Analysis

Shruti S. Gosavi
Junior Research Fellow
shruti.gosavi@tmv.edu.in
Tilak Maharashtra Vidyapeeth, Pune, India

Abstract

The modern research is focusing on the area of opinion mining also called as sentiment analysis due to steep degree of opinion thriving web assets such as argument forums, assessment sites and blogs are available in digital outline. Sentiment analysis is one of the machine learning technique in which machines scrutinize and categorize the sentiments, emotions, opinions of then human's about various topics which are articulated in the form of either wording or verbal communication.[24]Sentiment analysis or Opinion mining is nothing but which aims at influential what other public thinks, observes and states. Sentiments or Opinions includes unrestricted generated substance about goods, services, policies and political principles. In accumulation to satisfactory work being performed in text analytics, feature mining in sentiment analysis is at this moment flattering a dynamic area of research. Web content mining is the kind of mining, extraction and combination of functional data, information and understanding from web page filling. Mining of Frequent patterns from website data can help to improve the organization of web site and develop the recital of web server. There are several algorithms in data mining for mining frequent patterns such as Apriori Algorithm, FP-Growth Algorithm, etc.[26] Web usage mining is the method of discovering what users are looking for on Internet. Web has huge database and it is challenging job for data mining. By using such method we can discover and analyze the web data also we can accumulate more work time and get more practical information. Web mining is a permutation of the data mining technology and the web mining technology. Web mining is combination of web content mining, web structure mining and web usage mining. This survey will provide the areas in which issues and challenges arise in the field of opinion mining and sentiment analysis.

Keywords: *Challenges, Issues, Opinion Mining, Product Reviews, Sentiment Analysis, Web Mining*

Introduction:

The evaluation of teachers in a division has been focal point of attention for most educational examiners. The quality of teaching is not only calculated from ability and understanding of teacher but also their allegiance and assurance in classroom. [1] The prevailing view of an effective teacher is the teacher who possesses a broad knowledge of techniques and is able to skillfully use these techniques to meet the changing demands of the classroom. Furthermore in order to make improvements to teaching, it is essential to know what students deem of the way they are trained. Web Mining is the data mining technique which is used to discover and extract information from the World Wide Web or Internet. Web mining can be divided into of following task such as,

- Resource discovery: The assignment of retrieving proposed web documents.
- Information collection and Pre-processing: Automatically selecting and pre-processing precise information from the retrieved web assets.

ISSN: 2005-4262IJGDC

- Generalization: Automatically discovers common patterns at individual web sites as well as across numerous websites.
- Analysis: Validation and/or explanation of the mined patterns.
  For example, one of the analysis on a website might be mostly positive about a mobile handset, but can be particularly negative regarding how heavy in weight it is.

The rising importance of sentiment analysis coincides with the expansion of social media such as reviews, forum discussions, blogs, micro-blogs, Twitter, and social networks. One practice for accessing student necessities in the educational progression is to permit the chance to give feedback on the recital of their teacher which includes their viewpoint of instruction, institute, classroom atmosphere, and quality of the quantity learned. [2]

Opinion Mining:

The recent research is focusing on the area of Opinion Mining also called as sentiment analysis due to speedy degree of opinion thriving web assets such as disagreement forums, extent sites and blogs are on hand in digital sketch. [1] Sentiment analysis or Opinion mining is nothing but which aims at influential what other public thinks, observes and states. Opinions include unrestricted generated substance about goods, services, policies and political principles. [2]

Currently, there are many challenges in opinion mining and sentiment analysis some of them are listed below:

1. **Spam Detection and Fake Reviews**
   - Such type of review can be collected from content of review, abnormal behaviour of the reviewer
2. **Identification of comparitive words**
   - Such type of review can be collected from equalitive, non gradable, superlative relations of the review.
3. **Sentences with mixed views**
   - Such type of review can be collected from complex reviews through informal reviews such as blogs and forums
4. **Domain Independence**
   - Such type of review can be collected from negative reviews or fake reviews.
5. **Grouping synonyms**
   - Such type of review can be collected from various words and phrases which are then used to refer the same feature of that product

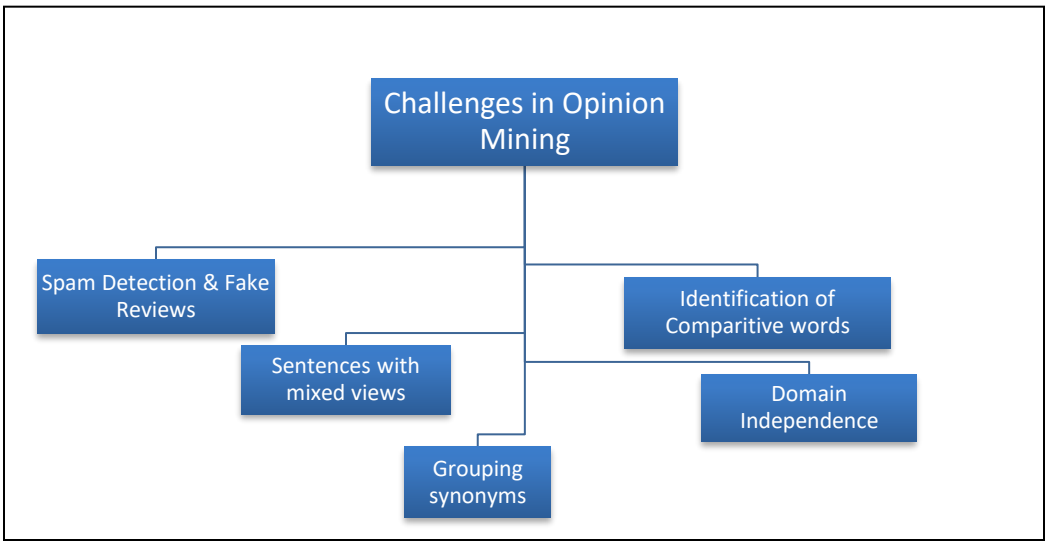

Fig (a): Challenges and issues in Opinion mining

Data Mining Algorithms:

1. Apriori Algorithm:

Apriori Algorithm is a typical algorithm for frequent item set mining and association rule learning over transactional databases. It profits by identifying the frequent individual items in the database and extending them to bigger and bigger item sets as extensive as those item sets become visible sufficiently often in the database. The frequent item sets resolute by Apriori Algorithm can be used to conclude association which underlines general trends in the database. This has applications in domains such as market basket analysis. Apriori algorithm combines with "bottom up" approach, in which frequent subsets are extends one item at a time also known as candidate generation, and collection of candidates are verified alongside the data. The algorithm terminates when no supplementary thriving extensions are found. Apriori algorithm combines breadth-first search and a Hash tree structure to count candidate item sets resourcefully. It generates candidate item sets of duration from item sets of length. Then it prunes the candidates which have an irregular sub pattern. According to the downward closure lemma, the candidate set contains all frequent -length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates.

Drawbacks of using Apriori algorithm is that in every count step we have to do a very costly scan over the complete database.

2. FP-Growth Algorithm:

In Data Mining the task of finding frequent pattern in huge databases is very significant and has been considered in large scale in the earlier period for few years. Regrettably, this task is computationally exclusive, especially when a large number of patterns is derived. The FP-Growth Algorithm, proposed by Han in, is an efficient and scalable process for mining the complete set of frequent patterns by pattern portion growth, using an complete prefix-tree structure for storing compressed and crucial information about frequent patterns named frequent-pattern tree (FP-tree). The FP-Growth Algorithm is an alternative way to find frequent item sets without by means of candidate generations, thus civilizing performance. For so much it uses a divide-and- conquer strategy. The core of this method is the practice of a particular data structure known as frequent-pattern tree (FP-tree), which recalls the item set association information.

In other words, such algorithm works as follows: [20]

Step 1: It compresses the input database by creating an FP-tree instance to correspond to frequent items.

Step 2: After this first step it separates the compressed database into a set of restrictive databases, each one associated with one frequent pattern. Later, each such database is mined independently. Using this policy, the FP-Growth reduces the search costs looking for tiny patterns recursively and then concatenating them in the extended frequent patterns, present good selectivity. In huge databases, its not probable to hold the FP-tree in the main memory. A strategy to cope with this problem is to first division the database into a set of smaller databases which are also known as projected databases, and then build an FP-tree from each of these smaller databases.

FP-Growth Algorithm operates in the following four modules[23]

- ✓ Pre-Processing
- ✓ FP Tree an FP Growth Module
- ✓ Association Rule Generation
- ✓ Results

1. The Preprocessing module extracts the log file, which usually is in ASCII format, into a database format, which can be processed by FP Growth Algorithm.
2. The Second Module is performed in two steps.

- FP-Tree generation

Using FP Growth to construct association rules FP tree is a compact data structure that provisions significant, critical and quantitative Information regarding frequent patterns.

Mainly components used in FP tree are: It combines of one root mentioned as root, a bunch of item set prefix sub-trees as the children of the root node, and a frequent-item set header table. Every node in the item prefix sub-tree contributes of three fields: item-name, count, and node-link, in which item-name registers which item this node represents, count registers the number of transactions represented by the portion of the path attainment this node, and node-link links to the next node in the FP tree carrying the same item-name, or null if there is not a bit. Every entry in the frequent-item header table combines of two fields, first is item-name and second is head of node link, which denotes to the first node in the FP-tree receiving the item-name.

- Secondly, an FP-tree-based pattern-fragment enlargement mining method is developed, which starts from a frequent length pattern (as an initial suffix pattern), examines only its conditional-pattern base (a sub-database which consists of the set of frequent items compatible with the suffix pattern), builds its (conditional) FP-tree, and conducts mining recursively with such a tree. The pattern growth is achieved via concatenation of the suffix pattern with the new ones generated from a conditional FP-tree. From the time when the frequent item set in any transaction is constantly encoded in the equivalent course of the frequent-pattern trees, pattern growth ensures the completeness of the result. FP-Growth algorithm is further divided into two categories:

Step1: Built a compact data structure called FP-Tree Extract Frequent Item sets Directly from FP-Tree

Step 1: FP-Tree Construction FP-Tree is constructed using two passes over the data sets

Pass 1: Scan the data and find support for each item Discard Infrequent Items Sort frequent Items in decreasing order based on their support

Pass2: FP-Growth reads one transaction at a time and maps it to a path. Fixed order is used, so paths can overlap when transactions share items Pointers are maintained between nodes containing the same item, creating singly linked lists Frequent item sets extracted from the FP-Tree

ISSN: 2005-4262IJGDC

Step 2: Frequent Item set Generation FP-Growth extracts frequent item sets from the FP-tree. Bottom-up algorithm - from the leaves towards the root Divide and conquer: First look for frequent item sets conclusion. First, extract prefix path sub-trees ending in an item (set). (indication: use the linked lists) Every prefix path sub-tree is processed recursively to extract the frequent item sets. Solutions are then combined later on.

Benefits in using FP-Growth Algorithm:

- FP-Growth algorithm is much quicker than apriori algorithm.
- Only two passes is obligatory over data sets.
- No Candidate Generation
- It eliminates repetitive database inspection.

Drawbacks in using FP-Growth algorithm:

- Although the transaction information well done in form of tree, but multiple recursive operation perform on tree at dynamic times so at every recursive operation accessing database and interact with hardware become time consuming so performance degraded.
- It is very expensive to build

Sentiment Analysis:

The textual data or information accessible in the web is growing day by day. Sentiment analysis is nothing but machines learning line of attack in which machines scrutinize and categorize the human's sentiments, emotions, and opinions about some product or process which are then expressed in the structure of either text or speech. In order to improve the sales of a product and to increase the customer satisfaction, almost all of the on-line shopping sites provide the opportunity to customers to express reviews about products. Such types of reviews are large in number and to mine the entire sentiment or opinion divergence from all of them, sentiment analysis can be used.[3]

Some challenges faced in sentiment analysis

1. **Sarcasm and Irony**
   - Such types of reviews are collected from the person who has studied psychology, cognitive science.
2. **Types of Negations**
   - Such type of review can be collected from traditional method based on static review and improper word sense.
3. **Word Ambiguity**
   - Such type of review can be collected from syntactic ambiguity, semantic ambiguity, pragmatic ambiguity and lexical ambiguity.
4. **Multipolarity**
   - Such type of review can be collected from polarity inconsistencies in the word in sentiment dictionary

ISSN: 2005-4262IJGDC

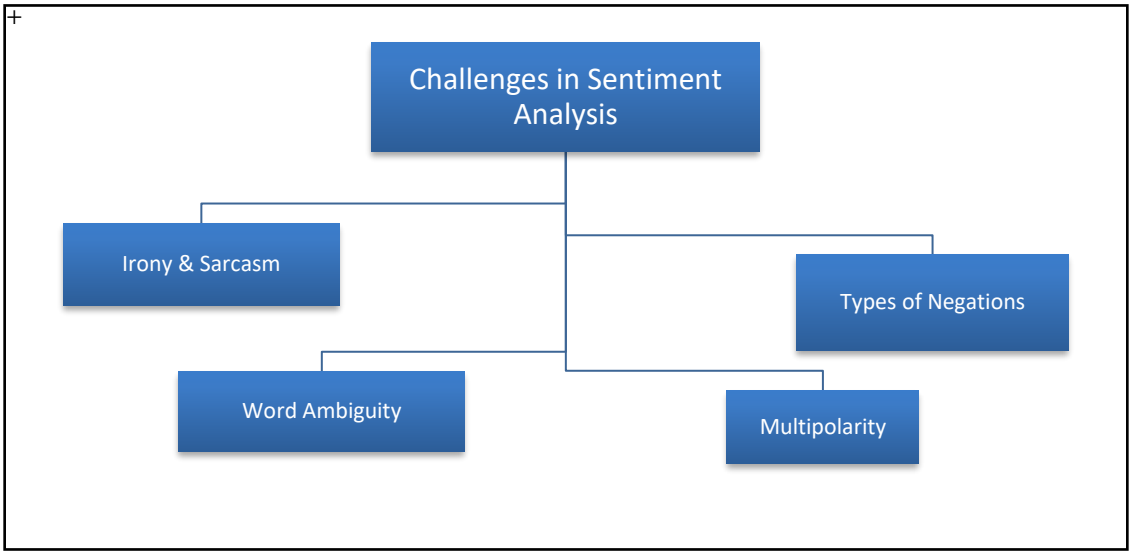

Fig (c): Challenges and issues in Sentiment Analysis

Conclusion:

The biggest challenge in opinion mining and sentiment analysis of product reviews is to manufacture a tumbledown of opinions based on product aspects. People can express attributes with many different terminology or phrases. The Frequent patterns mining algorithms like Apriori Algorithm, FP-Growth Algorithm, etc are used to mine frequent patterns. FP-growth algorithm is used for mining frequent item sets, candidate generation. There are some limitation of FP-Growth algorithm related to hardware so performance is declined. These terminologies are grouped underneath the same feature cluster to construct useful outline. Inadequate application has been done on combination of synonym features and clustering. In potential, Opinion Mining can be accepted out on a place of reviews and set of revealed feature expressions extracted from reviews. The state-of-art for existing methodologies, useful for producing better outline based on feature based opinions as positive, negative or neutral.

References:

1. Student Feedback Mining System Using Sentiment Analysis, R.Menaha, R.Dhanaranjani, T.Rajalakshmi, R.Yogarubini International Journal of Computer Applications Technology and Research Volume 6–Issue 1, 51-55, 2017, ISSN:-2319–8656
2. Sentimental Analysis of Student Feedback using Machine Learning Techniques, Daneena Deeksha Dsouza, Deepika, Divya P Nayak, Elveera Jenisha Machado, Adesh N. D. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1S4, June 2019
3. Sentiment Analysis and Opinion Mining: A Survey, G.Vinodhini, RM.Chandrasekaran, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012 I*SSN: 2277 128X
4. Bing Liu.2011.Sentiment Analysis Tutorial – Given at AAAI-2011, San Francisco, USA.
5. Andreas Auinger, Martin Fischer.2008.Mining consumers‟ opinions on the web

6. H D U K Kim, K. Ganesan, P Sondhi, C Zhai. 2011. Comprehensive Review of Opinion Summarization.
7. Classification of Product Reviews". In Proceedings of the 12th International Conference on World Wide Web, p. 519-528.
8. Pang, B., Lee, L. & Vaithyanathan, S, "Thumbs Up? Sentiment Classification Using Machine Learning
9. Gamon, M. (2004), "Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis". In Proceedings of the International Conference on Computational Linguistics (COLING 2004), p. 841-847.
10. Pang, B. & Lee, L. (2008), "Opinion Mining and Sentiment Analysis". In Foundations and Trends in Information Retrieval 2 (1-2), p. 1–135.
11. P. Sampath, C. Ramesh, T. Kalaiyarast, S. Sumaiya Banut, G. Arul Selvan An Efficient Weighted Rule Mining for Web Logs Using Systolic TreeIEEE-International Conference On Advances In Engineering, Science And Management (ICAESM - 2012) March 30, 31, 2012 432
12. J. Han, J. Pei, Y. Yin, and R. Mao, "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach," Data Mining and Knowledge Discovery, vol. 8, no. 1, pp. 53-87, Jan. 2004.
13. S. Sun and J. Zambreno, "Mining Association Rules with Systolic Trees," Proc. In!'1 Conf Field-Programmable Logic and Applications (FPL '08), Sept 2008.
14. R. Narayanan, D. Honbo, G. Memik, A. Choudhary, and J. Zambreno, "An FPGA Implementation of Decision Tree Classi_cation," Proc. Conf Design, Automation, and Test in Europe (DATE), pp. 189-194, Apr. 2007
15. Z. Baker and V. Prasanna, "Efficient Hardware Data Mining with the Apriori Algorithm on FPGAs," Proc. IEEE Symp. Field- Programmable Custom Computing Machines (FCCM), pp. 3-12, Apr. 2005
16. Y.-H.Wen, J.-W. Huang, and M.-S. Chen, "Hardware-Enhanced Association Rule Mining with Hashing and Pipelining," IEEE Trans. Knowledge and Data Eng.,vol. 20, no. 6, pp. 784-795, June 2008
17. A. Ghoting, G. Buehrer, Y.-K. Chen, and P. Dubey, "Cache-Conscious Frequent Pattern Mining on a Modern Processor," Proc. Int'I Conf Very Large Data Bases (VLDB), pp. 577-588, 2005 R. Bayardo, B. Goethals, and M. Zaki, eds., Proc. IEEE ICDM Workshop Frequent Itemset Mining Implementations (FIMI), Nov.2004.
18. T. Uno, M. Kiyomi, and H. Arimura, "LCM ver. 2: Efficient Mining Algorithms for Frequent/Closed/Maximal Itemsets," Proc. IEEE ICDM Workshop Frequent Itemset Mining Implementations (FIMI), 2004
19. C. Lucchese, S. Orlando, and R. Perego, "kDCI: On Using Direct Count Up to the Third Iteration," Proc. IEEE ICDM Workshop Frequent Itemset Mining Implementations (FIMl), 2004.
20. Jiawei Han Hong Cheng Dong Xin Xifeng Yan Frequent pattern mining: current status and future directionsReceived: 22 June 2006 / Accepted: 8 November 2006 Published online: 27 January 2007 Springer Science+Business Media, LLC 2007 M Suman , T Anuradha ,K Gowtham, , A Ramakrishna A frequent pattern mining algorithm based on fp-tree structure and apriori algorithm, InternationalJournal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 1, Jan-Feb 2012, pp.114-116

21. K.Saravana Kumar, R.Manicka Chezian, A Survey on Association Rule Mining using Apriori Algorithm International Journal of Computer Applications (0975 -8887) Volume 45 No.5, May 2012
22. Chhavi Rana, A Study of Web Usage Mining Research Tools Int. J. Advanced Networking and Applications Volume:03 Issue:06 Pages:1422-1429 (2012) ISSN : 0975-0290
23. Rakesh Kumar Malviya, Mahesh Chandra Malviya, Vinay Kumar Soni, Ritesh Joshi, Preetesh Purohit Survey of Web usage Mining, IJCST Vol. 2, ISSue 3, September 2011
24. Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey. R. Agrawal, T. Imielienski, and A. Swami. Mining Association Rules between Sets of Items in Large Databases. Proc. Conf. on Management of Data, 207216. ACM Press, New York, NY, USA 1993
25. Hypothyroid disease using data mining techniques." IJERT, ISSN (2013): 2278-0181.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.
26. Patil, Tina R., and S. S. Sherekar. "Performance analysis of Naive Bayes and J48 classification algorithm for data classification." International Journal of Computer Science and Applications 6.2 (2013): 256-261.
27. Wisaeng, Kittipol. "A comparison of decision tree algorithms for UCI repository classification." Int. J. Eng. Trends Technol 4 (2013): 3393-3397.
    a. http://en.wikipedia.org
28. Banu, G. Rasitha. "A Role of decision Tree classification data Mining Technique in Diagnosing Thyroid disease." International Journal of Computer Sciences and Engineering 4.11(2016):64-70.
    a. http://www.cs.waikato.ac.nz/ml/weka/
29. Gaganjot Kaur Amit Chhabra,"Improved J48 Classification Algorithm for the Prediction of Diabetes", International Journal of Computer Applications (0975 – 8887) Volume 98 – No.22, July 2014
30. Guo, Yang, Guohua Bai, and Yan Hu. "Using Bayes Network for Prediction of Type-2 Diabetes." In Internet Technology and Secured Transactions, 2012 International Conference For, pp. 471-472. IEEE, 2012.
31. Hany A. Elsalamony, Helwan University, Cairo, "Bank Direct Marketing Analysis of Data Mining Techniques", Saudi Arabia International Journal of Computer Applications (0975 – 8887Volume 85 – No 7, January 2014
32. A. Floares., A. Birlutiu. "Decision Tree Models for Developing Molecular Classifiers for Cancer Diagnosis". WCCI 2012 IEEE World Congress on Computational Intelligence June, 10-15, 2012 - Brisbane, Australia.
33. Tina R. Patil, Mrs S. S. Sherekar"Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification", International Journal of Computer Science and Applications Vol. 6, No.2, Apr 2013 ISSN: 0974-1011.
34. Decision Tree Induction: An Approach for Data Classification Using AVL-Tree. Devi Prasad Bhukya1 and S. Ramachandram2.
35. Milos Ilic, Petar Spalevic and Mladen Veinovic, Wejdan Saed Alatresh, "Students' success prediction using Weka tool", in INFOTEH-JAHORINA Vol. 15, March 2016.
36. Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages. ACM Transactions on Information Systems, 26 (3), 1–34. doi: 10.1145/ 1361684.1361685

37. Agarwal, B. , Mittal, N. , Bansal, P. , & Garg, S. (2015). Sentiment analysis using common-sense and context information. Computational Intelligence and Neuro- science, 2015 , 30 .
38. Balahur and Montoyo, 2008 Balahur, A., Montoyo, A., 2008. A feature dependent method for opinion mining and classification.Paper presented at the International Conference on Natural Language Processing and Knowledge Engineering, NLP-KE 08.
39. Baroni and Vegnaduzzo, 2004 Baroni Marco, Vegnaduzzo Stefano, 2004. Identifying subjective adjectives through webbased mutual information paper presented at the German conference on natural language processing KONVENS-04
40. Abulaish et al., 2009 M. Abulaish, M.N. Doja, T. Ahmad Feature and opinion mining for customer review summarization Pattern Recognition and Machine Intelligence, Springer Berlin
    a. Heidelberg (2009), pp. 219–224
41. Subjectivity and sentiment analysis of modern standard Arabic. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers. 2011.
42. Barbosa, Luciano and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. in Proceedings of the International Conference on Computational Linguistics (COLING-2010). 2010.
43. Bespalov, Dmitriy, Bing Bai, Yanjun Qi, and Ali Shokoufandeh. Sentiment classification based on supervised latent n-gram analysis. in Proceeding of the ACM conference on Information and knowledge management (CIKM-2011). 2011.
44. Bollegala, Danushka, David Weir, and John Carroll. Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification. in Proceedings of the 49th Annual Meeting of the Association for computational Linguistics (ACL-2011). 2011.
45. Bollen, Johan, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. Journal of Computational Science, 2011.
46. Hassan, Ahmed, Amjad Abu-Jbara, Rahul Jha, and Dragomir Radev. Identifying the semantic orientation of foreign words. in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers (ACL-2011). 2011.

ISSN: 2005-4262IJGDC