ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 3, March 2019

Intelligent Data Mining of Social Media for Better Decision Making

Manoj Mansukh[1], Rohit Marathe[1], Sagar Mali[1], Chaugule Balaji A.[2]

BE Students, Department of Computer Engineering, Zeal College of Engineering and Research Pune, India[1]

Professor, Department of Computer Engineering, Zeal College of Engineering and Research Pune, India[2]

ABSTRACT: The creation of social media and the rapid improvement of mobile communication technology have dramatically changed the way to express the feeling, attitude, temper, passion and so forth. People often express their reaction, fancies and predilections through social media by means of short texts of epigrammatic nature rather than writing long text. Many social websites like Twitter, Google Review, Just Dial, Book my show, etc enables people to share and discuss their thoughts, opinion and view in the form of short text, which can be useful for other unknown peoples, customers and service users to decide whether that service or product is good or not. In this paper, whole process is divided into two steps. In first step through intelligent data mining data will be abstracted and in second step analysis frame work will be there, which will focuses on positive and negative opinion and through R-Programming the visualization will be created will be helpful for peoples to make decision on their subject (i.e. on institute, services, products, movies, tourist spot etc.). In visualization section it will contain graph, pictures, etc. Comparison parameter can also be implemented which will be again very much helpful for user and peoples.

KEYWORDS: co-extracting algorithm, co-extracting model, opinion targets, Opinion Relation Graph, opinion words, Topical Word Trigger Model

I. INTRODUCTION

Online networking is giving boundless chances to patients to talk about their encounters with drugs and gadgets, and for organizations to get input on their items and administrations. Pharmaceutical organizations are organizing informal organization checking inside their IT offices, making an open door for quick spread what's more, input of items and administrations to enhance and improve conveyance, increment turnover and benefit, and lessen costs. Online networking information collecting for bio-surveillance has additionally been accounted.

The idea of informal groups makes statistics accumulating tough. A few techniques have been utilized, for example, linkmining, classification through links, predictions based on objects, links, presence, estimation, protest, aggregate, and subgroup location, and mining the information. Connection forecast, viral showcasing, online talk gatherings (and rankings) take into consideration the advancement of arrangements in view of client criticism.

In the first stage of our current work, we employ exploratory analysis using the Self Organizing Maps to assess correlations between user posts and positive or negative opinion on the drug. In a second stage, we model the users and their posts using a network-based approach. Conduct a manual qualitative analysis of a large sample of software-relevant data to determine the information value of software users' posts. Use text classification techniques to effectively capture and categorize the various types of actionable software maintenance requests present in such posts.Investigate the performance of various text summarization techniques in generating compact summaries of the common technical concerns raised in software systems'Twitter feeds. Our main objective is to lay down an infrastructure for a more responsive and a more adaptive software engineering process that can achieve user satisfaction in aneffective and a timely manner.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 3, March 2019

II. RELATED WORK

Artificial Societies and Social Simulation Using Ant Colony, Particle Swarm Optimization and Cultural Algorithms, This system proposes Ant Colony System Algorithm [1]. Artificial Societies and Social Simulation using exceptional techniques to analyze and model the important data to assist appropriate decisions of the evolving models. Advantages: Improves good quality in a short time. It has better performance. Disadvantage: Community of agents is not in application.

Collective Extraction for Opinion Targets and Opinion Words from Online Reviews, Proposes a method to extract opinion targets and opinion words collectively based on the word alignment model [2]. Acollective extraction for opinion targets and opinion words based on the word alignment model, in which the extraction can be treated as a classification problem. Design a semi-supervised extraction method based on active learning, since labeling training samples is time-consuming and error-prone. Advantages: Higher accuracy. Effectively ignore the problem of error propagation. Greatly reduce the work of manually labeling samples. Disadvantage: It does not apply parallel extraction method.

Tracing Information Flow and Analyzing the Effects of Incomplete Data in Social Media, Proposes a k-tree model of cascades is generated from a balanced tree of height and branching factor. The goal of [3] paper is to address methods to collect massive amounts of social media data and what techniques can be used for correcting for the effects and biases arising from incomplete and missing data. Advantages: The information flow is unambiguous and precise. We can have the time, so it's easy to trace the information. In a very large network, it becomes easy to collect data. However, if data is incomplete cascades break into pieces. Many different diffusion mechanisms. Disadvantage: Not all the links transmits the information. Sometimes the links get missing due to Blogger forget to attach a link or mainstream media does not provide the source links. Not clear whether hashtags really diffuse. Due to "personalization" easier to argue URLs diffuse. Problem with all is that we do not realize the "influencer".

Text and Structural Data Mining of Influenza Mentions in Web and Social Media, Proposes a graph-based data mining technique [4] to detect anomalies and informative substructures among flu blogs connected by publisher type, links, and user-tags. Text mining of influenza mentions in WSM is shown to become aware of traits in flu posts that correlate to real-international ILI affected person reporting statistics. Advantages: To identify trends in flu posts that correlate to real-world ILI patient reporting data. Disadvantage: Content analysis does not provide.

Text Mining: Promises And Challenges, Proposes a text mining framework consists of two frameworks: Text refining and Knowledge distillation. The textual content refining that transforms unstructured text files into an intermediate form; and know-how distillation that deduces patterns or know-how from the intermediate form [5]. Advantages: Customer profile analysis, Patent analysis, Information dissemination and Company resource planning. Disadvantage: There are issues in this paper, semantic analysis, multilingual text refining, domain knowledge integration and personalized autonomous mining.

Social media competitive analysis and text mining: A case study in the pizza industry, Proposes competitive analysis for the user-generated data on Twitter and Facebook in three major pizza chains [6]. Results from the text mining and social media competitive evaluation show that these pizza chains actively engaged their clients in social media along with Twitter and Facebook. Advantages: Establishing effective and realistic benchmarks. Mining the content of social media conversations.Disadvantage: Does not track real-time data.

Mood Based Classification of Music by Analyzing Lyrical Data Using Text Mining, Proposes classification using Support Vector Machine algorithm [7]. As mood classes in music mental models may do not have the social connection of today's music listening environment, this research inferred an arrangement of mood classifications from social labels utilizing etymological assets and human skill. The resultant mood classes were contrasted with two delegate models in music brain science. Advantages: The framework may be utilized to hunt down female craftsmen, content melodies, or hallucinogenic music. Content features may prompt higher correctness's for most mood classifications.

Text Mining for the Hotel Industry, This paper proposes text mining as a means of information management. Text mining can examine the voluminous textual facts that can be found in a inn's internal databases and external sources. In [8] paper, illustrates how the text-mining technique may help to translate online textual information into meaningful competitive and customer intelligence for managerial decision making. Advantages: Reduce the use of

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 3, March 2019

manual labor in identification, storage, and analysis of business intelligence. Fully automated system. Disadvantage: Does not integrate text-mining tools with related technologies such as image recognition and Web-table mining.

Feature extraction and classification of proteomics data using stationary wavelet transform and naïve Bayes classifier, Proposes Naïve Bayes Algorithm and stationary wavelet transformation [9]. The data processes of MS signal in this paper mainly include two parts: preprocessing and biomarker selection, and the results are determined mainly by these two steps. To the denoising using SWT, compared to DWT, SWT it is very appropriate for this application for the characteristics of the MS data. Advantages: It requires a small amount of training data to estimate the parameters necessary for classification. High sensitivity, specificity and accuracy.

Semantic Data Analysis Algorithms Supporting Decision-making Processes, Proposes semantic data analysis processes, and theirs role in supporting decision-making tasks as well as intelligent management.The most important is that such systems may support financial or economy processes taken in different enterprises or institution. Wide information records obtained thanks to the application of cognitive information systems [10] allow finding many different applications both in local and global environment. Advantages: Cognitive systems are very efficient. Disadvantages: The structure of information record is too complex to perform full interpretation. The system has not enough knowledge to fully describe the semantic meaning of analyzed information record or complex structure.

III. SYSTEM OVERVIEW

The objective of feeling extraction is to recognize where in reports opinion, review or tweets are installed. Opinions are covered up in words, sentences and records. A sentiment sentence is the littlest complete semantic unit from which opinions can be removed. The opinion words, the opinion holders, and the logical data ought to be considered as hints while separating sentiment sentences and deciding their inclinations. Subsequently, the extraction algorithm is developed base by recognizing opinion words at in the first place, then distinguishing the feeling polarities of sentences lastly reports a short time later. Feeling scores of words, which speak to their opinion degrees and polarities. We used the Twitter Search API, to acquire our dataset of software program applicable tweets. In our analysis, we restriction our data collection manner to tweets addressed immediately to the Twitter account of a given software program machine.To automatically classify our data, we investigate the performance of sentiment analysis algorithm. The contribution work is to generate report in .pdf format of all APIs or single API of given query.

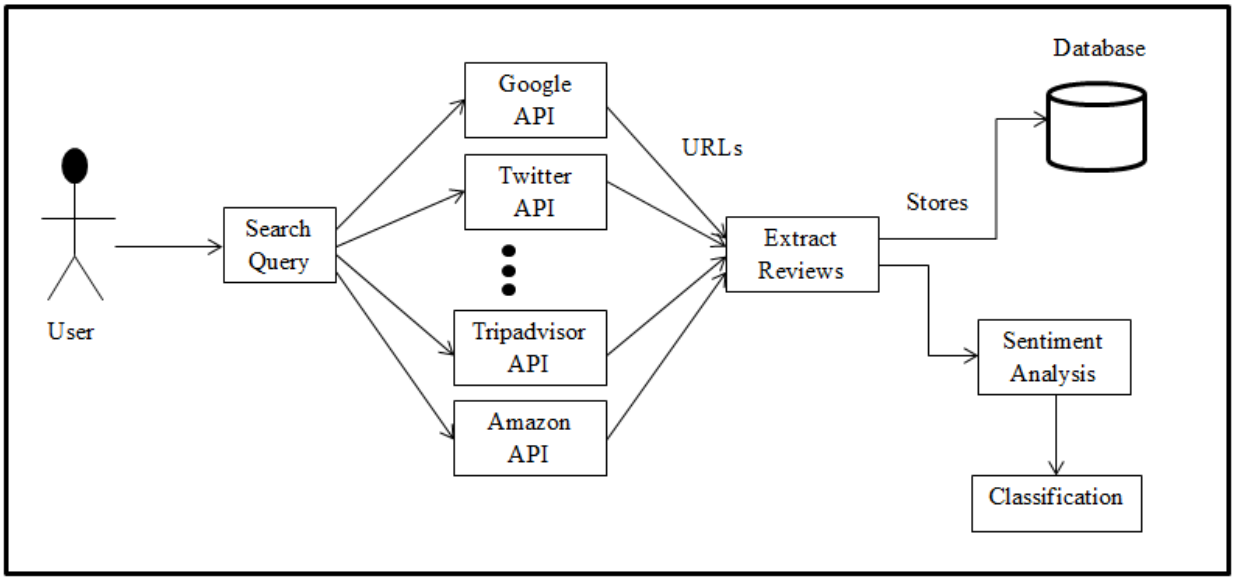

Fig.1 Proposed system architecture

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 3, March 2019

Advantages:

- Less time consumption.
- Easy access.
- Pervasiveness.
- Data from different websites can be found at one place.
- Data scraping only single click button.
- Automatic classification.

IV. MATHEMATICAL MODULE

1) **Web scraping**

Web pages are built the usage of text-based totally mark-up languages (HTML and XHTML), and regularly comprise a wealth of beneficial data in textual content form. However, maximum internet pages are designed for human quit-customers and not for ease of automated use. Because of this, tool kits that scrape internet content were created. A web scraper is an API or tool to extract information from a web web site. Companies like Amazon AWS and Google offer internet scraping tools, services and public data available freed from fee to stop users. Newer forms of web scraping involve being attentive to records feeds from internet servers. For instance, JSON is normally used as a transport storage mechanism among the customer and the internet server.

Recently, businesses have superior web scraping systems that depend on the use of strategies in DOM parsing, laptop vision and natural language processing to simulate the human processing that takes location whilst viewing a webpage to automatically extract useful facts.

Large web sites normally use protecting algorithms to guard their data from net scrapers and to restrict the number of requests an IP or IP community may ship. This has prompted an ongoing conflict between internet site builders and scraping developers.

2) Sentiment Analysis Algorithm:

Input: Text File(comment or review) T, The sentiment lexicon L.

Output: $S_{mt} = \{P, Ng\ and\ N\}$ and strength S where P: Positive, Ng: Negative, N: Neutral

Initialization: SumPos = SumNeg =0, where,

SumPos: accumulates the polarity of positive tokens ti-smt in T,

SumNeg: accumulates the polarity of negative tokens ti-smt in T,

Begin

1. **For each** $t_i \in T$ **do**
2. Search for t_i in L
3. **If** $t_i \in Pos - list$ **then**
4. SumPos ← SumPos + ti – smt
5. **Elseif** $t_i \in Pos - list$ ***then***
6. SumNeg ← SumNeg + ti – smt
7. **End If**
8. **End For**
9. **If** $SumPos > |SumNeg|$ ***then***
10. Smt = P
11. S=SumPos/(SumPos+SumNeg)
12. **Else If** $SumPos < |SumNeg|$ ***then***
13. Smt = Ng
14. S=SumNeg/(SumPos+SumNeg)
15. **Else**
16. Smt = N
17. S=SumPos/(SumPos+SumNeg)
18. **End If**

End

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 3, March 2019

V. EXPERIMENTAL RESULT

In this application when user enters search query with the help of using category like, product, hotel, restaurant, movie, book etc. We used the Google Search Engine for getting URLs of the user searches query. With the help of URLs connecting to the Twitter API, Google News API, Amazon and Tripadvisor sites for crawling reviews using category wise. The reviews are collected from given sites and apply sentiment analysis algorithm for classification of reviews in positive or negative. Then, verify the performance of the given sites reviews.

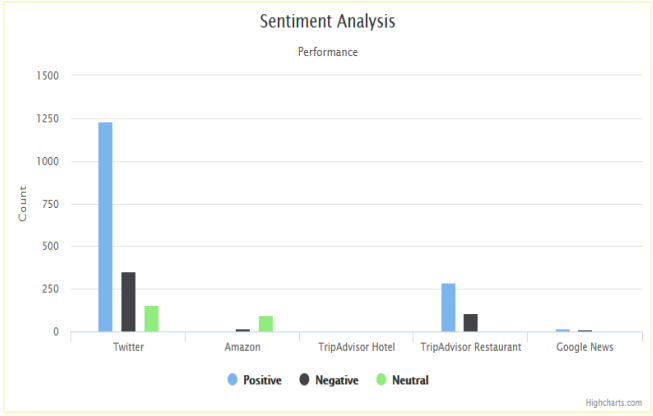

Fig. 2 Sentiment analysis performance

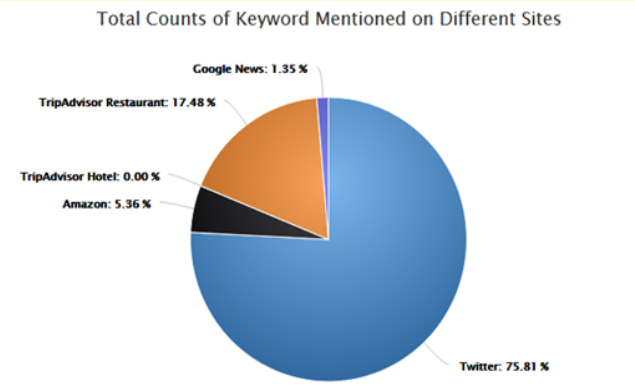

Fig. 3 Pie chart of total percentage of reviews extracted from different websites

Table I Performance of reviews on different sites

APIs	Positive Count	Negative Count	Neutral Count	Total Count
Twitter	1230	355	154	1739
Amazon	7	18	98	123
Tripadvisor Hotel	0	0	0	0
Tripadvisor Restaurant	289	105	7	401
Google News	17	12	2	31

VI. CONCLUSION

This project is implement using data scraping technique. After data scraping, the number of posts are getting to user. Apply the sentiment analysis on every post which leads to user decide which one is good. This project reduces the

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 7, Issue 3, March 2019

time consumption and the work of searching in many website. This provides an easy access platform for peoples, which help them to take better decisions. Visualization and comparison add more efficiency for decision making. In future scope, it can be made globalize, available all over the World. More advancement can be made in project, extracting data from various social media. Various classification algorithms can be implemented for better classifying the target words.

REFERENCES

[1] Ochoa, A. Hernandez, L. Cruz, J. Ponce, F. Montes, L. Li, and L. Janacek. "Artificial Societies and Social Simulation Using Ant Colony, Particle Swarm Optimization and Cultural Algorithms," *New Achievements in Evolutionary Computation*, Edition of book, Vol., P. Korosec,, Ed:, p. 267-297, 2010.
[2] Jiang, Xiangxiang, Yuming Lin, You Li and Jingwei Zhang. "Collective Extraction for Opinion Targets and Opinion Words from Online Reviews." 2016 7th International Conference on Cloud Computing and Big Data (CCBD) (2016): 367-373.
[3] Bindra, Gundeep Singh et al. "Tracing Information Flow and Analyzing the Effects of Incomplete Data in Social Media." 2012 Fourth International Conference on Computational Intelligence, Communication Systems and Networks (2012): 235-240.
[4] Corley, Courtney, Diane Joyce Cook, Armin R. Mikler and Karan P. Singh. "Text and Structural Data Mining of Influenza Mentions in Web and Social Media." International journal of environmental research and public health (2010).
[5] Kent, Ah-Hwee Tan. "Text Mining: Promises and Challenges." (1999).
[6] He, Wu, ShenghuaZha and Ling Li. "Social media competitive analysis and text mining: A case study in the pizza industry." Int J. Information Management 33 (2013): 464-472.
[7] Kashyap, Nirbhay, TanupriyaChoudhury, Dev Kumar Chaudhary and R. Lal. "Mood Based Classification of Music by Analyzing Lyrical Data Using Text Mining." 2016 International Conference on Micro-Electronics and Telecommunication Engineering (ICMETE) (2016): 287-292.
[8] Kin-Nam Lau, Kam-Hon Lee, and Ying Ho. "Text Mining for the Hotel Industry" Vol 46, Issue 3, 2005, pp. 344 - 362
[9] Liu, Dan, Yuan-Yuan Huang and Chen-xiang Ma. "Feature Extraction and Classification of Proteomics Data Using Stationary Wavelet Transform and Naive Bayes Classifier." 2010 4th International Conference on Bioinformatics and Biomedical Engineering (2010): 1-4.
[10] Ogiela, Lidia and Marek R. Ogiela. "Semantic Data Analysis Algorithms Supporting Decision-Making Processes." 2015 10th International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA) (2015): 494-496.