ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

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

Visit: www.ijirset.com

Vol. 7, Issue 5, May 2018

Intelligent Data Mining On Reviews for Better Decision Making

Shaikh Ashfak[1], Sonali Ransure[1], Anup Bhosale[1], Subhash Kumar[1], Dr.Arati Dandavate[2]

Students, Department of Computer Engineering, JSPM's Jayawantrao Sawant College of Engineering, Savitribai Phule Pune University, Pune, India[1]

Professor, Department of Computer Engineering, JSPM's Jayawantrao Sawant College of Engineering, Savitribai Phule Pune University, Pune, India[2]

ABSTRACT: The advent of social media and the rapid development of mobile communication technologies have dramatically changed the way to express the feeling, attitude, mood, passion etc. 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, FaceBook, Amazon, 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 Jquery the visualization will be created will be helpful for peoples to make decision on their subject (i.e. on Hotels, resturants, services, products, movies, etc). In visualization section it will contain graph. Comparison parameter can also be implemented which will be again very much helpful for user and peoples.

KEYWORDS: co-extracting algorithm, predilections, epigrammatic.

I. INTRODUCTION

If any customer or service consumer wants to use or try some new or unknown product or services then before buying that product they are checking in various no of websites for review and opinion on the same. And in today's era of digitalization, internet and social websites it become easy to get opinion and review, but in unorganized form.

In the first stage of our current work, user has to create an account on our web application. Second he/she has to select category of query that he/she want to search. In a background process, we first collect information from Twitter API, Facebook APIs, Amazon and Tripadvisor.Then after it by using text classification techniques to effectively capture and categorize the various types of words. And by using sentiments analysis algorithm we analyse the sentence and opinion of the peoples. In last we visualize the summery of whole sentiments by drawing the graph by jquery.

II. LITERATURE SURVEY

1. **A new approach towards co-extracting opinion- targets and opinion words from online reviews,** Efficient extraction of data can be done.
2. **Artificial Societies and Social Simulation Using Ant Colony, Particle Swarm Optimization and Cultural Algorithms,** This system proposes Ant Colony System Algorithm. Artificial Societies and Social Simulation using different strategies to analyze and model the necessary information to support the correct 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.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

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

Visit: www.ijirset.com

Vol. 7, Issue 5, May 2018

3. **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. a collective 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.
4. **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 this 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 know the "influencer".
5. **Text and Structural Data Mining of Influenza Mentions in Web and Social Media,** Proposes a graph-based data mining technique 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 identify trends in flu posts that correlate to real-world ILI patient reporting data. Advantages: To identify trends in flu posts that correlate to real-world ILI patient reporting data. Disadvantage: Content analysis does not provide.
6. **Text Mining: Promises And Challenges,** Proposes a text mining framework consists of two frameworks: Text refining and Knowledge distillation. The text refining that transforms unstructured text documents into an intermediate form; and knowledge distillation that deduces patterns or knowledge from the intermediate form. 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.
7. **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. Results from the text mining and social media competitive analysis show that these pizza chains actively engaged their customers in social media such as Twitter and Facebook. Advantages: Establishing effective and realistic benchmarks. Mining the content of social media conversations. Disadvantage: Does not track real-time data.
8. **Mood Based Classification of Music by Analyzing Lyrical Data Using Text Mining,** Proposes classification using Support Vector Machine algorithm. 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.
9. **Mining Twitter Data for a More Responsive Software Engineering Process,**Better analyzing can be done.
10. **Feature extraction and classification of proteomics data using stationary wavelet transform and naïve Bayes classifier,** Proposes Naïve Bayes Algorithm and stationary wavelet transformation. 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.
11. **Network-Based Modeling and Intelligent Data Mining of Social Media for Improving Care** Effective treatments can derive from the opinion of large patient and doctors. Lack of advancement in technique of mining, sorting data and algorithms.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

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

Visit: www.ijirset.com

Vol. 7, Issue 5, May 2018

12. **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 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.
13. **Text Mining for the Hotel Industry**, Proposes online text mining with analysis. Reduce the use of manual labor in identification, storage, and analysis of business intelligence. Fully automated system. Does not integrate text-mining tools with related technologies such as image recognition and Web- table mining.

III. PROPOSED SYSTEM

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 collect our dataset of software relevant tweets. In our analysis, we limit our data collection process to tweets addressed directly to the Twitter account of a given software system. To automatically classify our data, we investigate the performance of sentiment analysis algorithm.

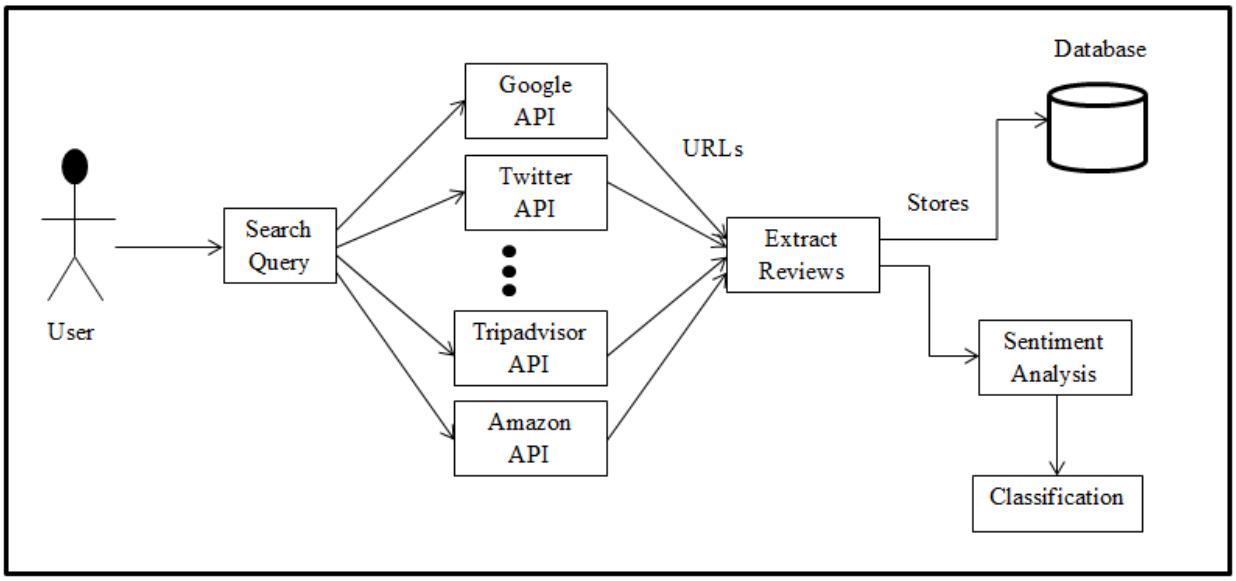

Fig.1 Proposed system architecture

Advantages:

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

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

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

Visit: www.ijirset.com

Vol. 7, Issue 5, May 2018

- Automatic classification.

IV. MATHEMATICAL MODULE

1) Web scraping

Web pages are built using text-based mark-up languages (HTML and XHTML), and frequently contain a wealth of useful data in text form. However, most web pages are designed for human end-users and not for ease of automated use. Because of this, tool kits that scrape web content were created. A web scraper is an API or tool to extract data from a web site. Companies like Amazon AWS and Google provide web scraping tools, services and public data available free of cost to end users. Newer forms of web scraping involve listening to data feeds from web servers. For example, JSON is commonly used as a transport storage mechanism between the client and the web server.

Recently, companies have developed web scraping systems that rely on using techniques in DOM parsing, computer vision and natural language processing to simulate the human processing that occurs when viewing a webpage to automatically extract useful information.

Large websites usually use defensive algorithms to protect their data from web scrapers and to limit the number of requests an IP or IP network may send. This has caused an ongoing battle between website developers 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

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, Facebook API, Amazon and Tripadvisor sites for crawling reviews using category wise.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

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

Visit: www.ijirset.com

Vol. 7, Issue 5, May 2018

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.

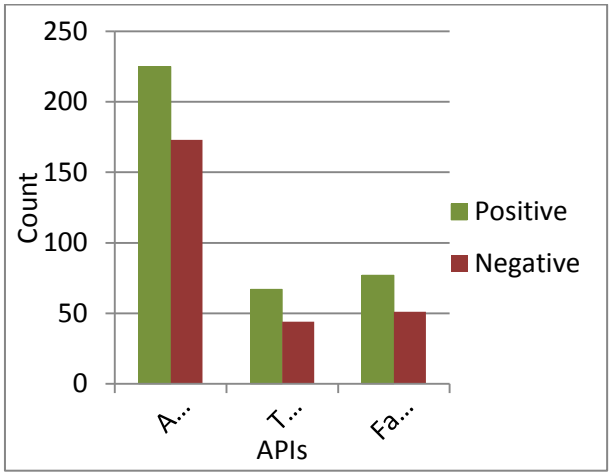

Fig.2 Sentiment Analysis on reviews graph of different APIs

Table I Classification of reviews using different APIs

APIs	Positive	Negative
Amazon	225	173
Twitter	67	44
Facebook	77	51

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 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. A new approach towards co-extracting opinion- targets and opinion words from online reviews
2. Network-Based Modeling and Intelligent Data Mining of Social Media for Improving Care
3. Mining Twitter Data for a More Responsive Software Engineering Process
4. Artificial Societies and Social Simulation Using Ant Colony, Particle Swarm Optimization and Cultural Algorithms
5. Collective Extraction for Opinion Targets and Opinion Words from Online Reviews
6. Tracing Information Flow and Analyzing the Effects of Incomplete Data in Social Media
7. Text and Structural Data Mining of Influenza Mentions in Web and Social Media
8. TEXT MINING: PROMISES AND CHALLENGES
9. Social media competitive analysis and text mining: A case study in the pizza industry
10. Mood Based Classification of Music by Analyzing Lyrical Data Using Text Mining
11. Text Mining for the Hotel Industry
12. Feature extraction and classification of proteomics data using stationary wavelet transform and naïve Bayes classifier
13. Semantic Data Analysis Algorithms Supporting Decision-making Processes