Volume 5, Issue 10, October-2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Opinions Analysis of Unstructured Textual Reviews for Efficient Decision Making

Shatakshi Agrawal, Dr. Sadhna Mishra, Prof. Gaurav Nayak
Computer Science& Engineering, LNCTS, RGPV,
Bhopal, India

Abstract— The Technology have provided the people a global platform in forms of various social sites, blogs, online web discussion for sharing their opinions about the particular products or their services. The large no of public opinions are present on the web that provide valuable information for the further feedback and their improvement. The opinions are in the form of numerical and textual. The numerical calculation is easy to process but a lot of difficulties are faced in analysis and summarization of these textual opinions. In this paper a novel opinion mining and summarization application system is proposed and implemented that is able to analyze and summarize the textual opinions. Experimental results have indicated a satisfactory performance of the opinion mining and summarization application system tested over an organization

Keywords—Opinion mining, Summarization, Semantic Analysis, words polarity.

I. INTRODUCTION

Opinions are the best way in which the human express their feelings about the particular products or services they are using of any organization or also they express opinions about the place where they work. The various studies on opinions have come up to the point that opinions are generally the feedback the person provides after they experiences. The opinions are expressed in general form as there is no pattern or structure for expressing it.

In the early era, the opinions or feedback where collected numerically in the form of 0, 1, 2. The numerical data is easy to process but today due to the advancement of the modern technology opinions are expressed freely in the form of comments, post, blogs etc. The freely expressed opinions are processed by the use of natural language processing. The Study of natural language facilitates to understand the freely expressed opinions their ways of presentation as well as different rules that needs to follows in order to understand the structure of the language.

The Semantic analysis is an important point of analysis in these the clusters are created for the things that represent entity or semantic analysis can be also called as polymorphism one things having different name. In the opinions it is not guarantee that the people will use the same name they can use similar name for opinions expression.

The next and most important point of study is opinion analysis also called as Sentiment analyses which are use for predicating the opinion (good. Bad, excellent, worse) form the sentences. The opinion analysis plays an important for analysis. The opinion analysis helps to understand the polarity that the person is expressing about that particular entity. The needs for opinion mining and summarization emerge from these new way of expressing the opinions and feedback.

This paper presents a novel opinion mining and summarization application system for analysing the unstructured textual data. Section I deals with the introduction about the opinion mining. Section II presents an extensive literature survey about the methods and techniques for opinion mining. Section III presents the proposed work for the system. Section IV represents the experimental and result work the conclusion is stated in Section IV of this paper.

II. LITERATURE SURVEY

Katerina Kabassi & etal researches on multi criteria decision-making theories with a cognitive theory called human plausible reasoning (HPR) to provide personalized assistance via graphical user interfaces (GUIs). A GUI called intelligent file manipulator (IFM) helps with organizing computer file storage. The system reasons about user actions, goals, plans, and possible errors and offers automatic assistance in case of a problematic situation. Three multi criteria decision-making theories [simple additive weighting, multi attribute utility theory, and data envelopment analysis] were adapted, implemented, and combined with HPR, in turn. This process resulted in three different versions of IFM that were evaluated. [1]

Yingcai Wu & etal develop an opinion diffusion model to approximate opinion propagation among Twitter users. Accordingly, they design an opinion flow visualization that combines a Sankey graph with a tailored density map in one view to visually convey diffusion of opinions among many users. A stacked tree is used to allow analysts to select topics of interest at different levels. The stacked tree is synchronized with the opinion flow visualization to help users examine and compare diffusion patterns across topics. [2]

Desheng Dash Wu etal develops methodology that integrates popular sentiment analysis into machine learning approaches based on support vector machine and generalized autoregressive conditional heteroskedasticity modelling. A

corpus of financial review data was collected. Computational results show that the statistical machine learning approach has higher classification accuracy than that of the semantic approach. [3]

Fuji Ren, etal focus on a challenging problem of predicting users' opinions toward topics they had not directly given yet, which they define as user-topic opinion prediction. The main contributions of this paper are as follows: 1) Different from previous work recognizing emotional states/sentiments from online micro blogging data but ignoring whose they are, seek to find out who has what opinion of a specific topic in advance. Predicting individual's feeling about a given target is important for affective computing studies and able to be used to various applications. 2) To provide a solution, the author considers the opinion homophily among Twitter social friends and users' opinion consistency on content-related topics, and formulate them as social context and topical context mathematically. 3) Utilizing the learned emotional knowledge from the observed tweets and the social and topical context information, we propose a framework ScTcMF to predict the unknown user-topic opinions.[4]

Xiaohui Yu etal presents explored the predictive power of reviews using the movie domain as a case study, and studied the problem of predicting sales performance using sentiment information mined from reviews. The author has approached this problem as a domain-driven task, and managed to synthesize human intelligence (e.g., identifying important characteristics of movie reviews), domain intelligence (e.g., the knowledge of the "seasonality" of box office revenues), and network intelligence (e.g., online reviews posted by moviegoers). The outcome of the proposed models leads to actionable knowledge that can readily be employed by decision makers. [6]

Jingbo Zhu, etal presents aspect-based opinion polling from unlabeled free-form textual customer reviews without requiring customers to answer any questions. First, a multi-aspect bootstrapping method is proposed to learn aspect-related terms of each aspect that are used for aspect identification. Second, an aspect-based segmentation model is proposed to segment a multi-aspect sentence into multiple single-aspect units as basic units for opinion polling. Finally, an aspect based opinion polling algorithm is presented in detail. Experiments on real Chinese restaurant reviews demonstrated that our approach can achieve 75.5 percent accuracy in aspect-based opinion polling tasks.[7]

III. PROPOSED WORK

Natural language based Opinions Mining and Summarization system for Unstructured Textual Data for extracting the valuable information for Decision as well as it will be capable of deciding the opinions polarity of those particular Features of the institute from the textual data.

The proposed system uses supervised machine learning approach for semantic datasets and polarity datasets generation. The student's opinions about their educational institutes will be considered in order to test the working and effectiveness of the system.

The Proposed system is able to solve all of the following objectives.

- To provide user with space to write their own opinions as well as taking the opinions from the internet that are available in the form of web blogs, and comments.
- To identify the feature of the organization from the opinions in the pos tagging module.
- To Create the Semantic Datasets for similarity matching
- To find out the features that are relevant to the organization in the semantic mapping module
- To calculate polarity of the identified features for result formulation in the opinion module.
- Structuring the result for efficient decision making.
- Efficiency of organization can be greatly improved and Solutions can be searched fast for Updation or making improvement.

The proposed system is divided into different modules for efficient processing. The modules are as follows:

- Textual Opinions Collection.
- Pre processing & Pos tagging
- Features Identification
- Semantic Database Creation
- Semantic Analysis & Similarity matching
- Opinion Analysis

- **Opinions Data**

The Opinions are collected from the students or taking the opinions from the internet that is available in the form of web blogs, and comments.

Example:

Financial support should be given to all.
Air conditioner should be installed.
Company Package is very less.
Trainer should be available for sports activities.
Space in Canteen is very limited.
Playground is not available.

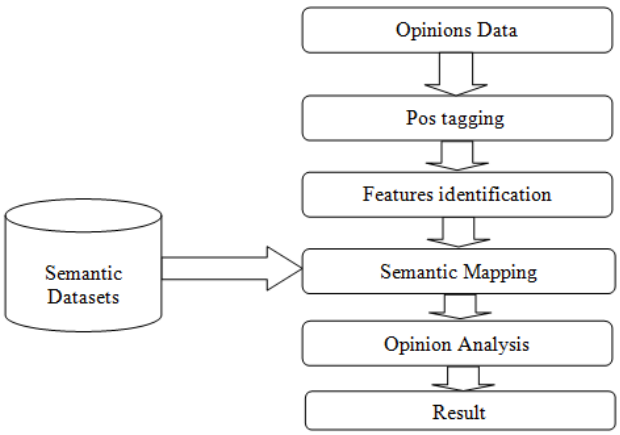

Fig1: Opinions Mining and Summarization System

- **Preprocessing or Postaging**

The pos tagging is important for finding the pos tagged words from the sentences. The Stanford tagger is considered for the system. It is trained with 10, 00,000 words from the Oxford dictionary with the help of standard Maxnet tagger methods.

The Stanford tagger uses the following ideas (i) explicit use of both preceding and following tag contexts via a dependency network representation, (ii) it uses a broad view of lexical features, on multiple consecutive words including jointly conditioning, (iii) in conditional log linear model priors are effectively used, and (iv) fine-grained modelling of unknown word features. Using these ideas together, a bi-directional dependency network tagger in bidirectional/wsj3t0-18 holder gives 97.24% accuracy on the Penn Treebank WSJ, an error reduction of 4.4% of the best previous single automatically learned tagging result. The tagger uses a bi-directional dependency network tagger for tagging the words the tagger is composed of both the features of tagging; it uses a CMM method of left to right and right to left for extracting the tagged tokens.

Examples:

Financial/NN support/NN should/MD be/VB given/VBN to/TO all/DT./.
Air/NN conditioner/NN should/MD be/VB installed/VBN./.
Company/NN Package/NN is/VBZ very/RB less/JJR./.
Trainer/NN should/MD be/VB available/JJ for/IN sports/NNS activities/NNS./.
Space/NNP in/IN Canteen/NNP is/VBZ very/RB limited/JJ./.
Playground/NN is/VBZ not/RB available/JJ./.

Feature Identification

The words Tag with NN/NP/NNS in the preprocessing step are considered as noun from the statement. The noun words in English are used to represents the entity on which the adjective is expressed. So all the NN tag words are taken as feature from the sentences. The basic rules of English Grammar are used for feature identification and these features are considered for polarity finding and calculation.

Examples:

Financial/NNS support/NN should/MD be/VB given/VBN to/TO all/DT./.
--The important entity word for this sentence is <Financial support >
Company/NN Package/NN is/VBZ very/RB less/JJR./.
--The important entity word for this sentence is <Company Package >
Trainer/NN should/MD be/VB available/JJ for/IN sports/NNS activities/NNS./.
--The important entity word for this sentence is <Trainer sports activities >

- **Semantic Datasets**

The semantic datasets are created in order to map the entity words extracted from the pre processing step.

Words or phrases	Entity	Properties
Contents	Teaching	Contents
Books issue	Library	Books
Trainer	Extracurricular activities	Sports activities
Financial supports	Extracurricular activities	Sports activities
Drinking	Canteen	Items

- **Semantic matching**

Jaccard method calculates similarity between words and provides similarity coefficient value. A threshold value is used to decide whether that two word are similar or not. Word pair which qualifies threshold value is considered as similar. In this way a misspelled word can be match with the entity in the ontology. Jaccard similarity determines the Jaccard coefficient. The Jaccard similarity coefficient is a statistical measure of similarity between sample sets. It is defined as the cardinality of their intersection divided by the cardinality of their union.

Mathematically,
J (A, B) = |A ∩ B| / | A U B|
Eg: X= {A, B, C, D, E}, Y = {B, C, D, E, F}
X and Y are words.
Jaccard similarity=4/6=0.67.

- **Opinion analysis**

For Calculating the Polarity from the sentences the Rule based approach are used. The rule based approach works on the principle of language grammar and the condition that needs to be satisfied in order to express the polarity for the entity in the sentences.

Examples:
Considering the sentences for the polarity algorithms
Air conditioners are installed.
Financial supports are very good for sports facility.
Company Package is not very bad.
Trainer should be available for sports activities.
Space in Canteen is very limited.
Playground is not available.

In the first sentence the word "installed" defines polarity for the air conditioner the word "installed" is checked for the polarity it is found that "installed" word belongs to the positive set of words now the pointer PC tracks the position of that word in the sentence and store the word index in index 1. After that set of more word is searched, if it is not found then inversion set word is searched in the sentence. If it is also not found then good count is incremented by 1 Result of this search is good count = 1

Sentence: Air conditioners are installed; Related to infrastructure; Good: 1; Bad: 0; Feedback: good

In the second sentence the word "very good" defines the polarity for the financial support entity the pointer pc searches for the word and finds "good" related to positive set then its index is again stored in index 1 and then the more set word is search the more set word "very" is found its index is store in index 3 and good count is increment by 1. After that inversion word is searched the inversion word is not found. So the final output will be

Sentence: Financial supports are very good for sports facility; Related to: extracurricular activities; Good: 2; Bad: 0; Feedback: very good.

In the third sentence the word "not very bad" defines polarity for the company package the word "bad" is checked for the polarity it is found that bad belongs to the negative word so the pointer nc searches for that word and store its index in index1 then the more set words are searched. The word "very" is available so its position is stored in index3 and the bad count is increment by 1 then the set of inversion word is searched the inversion word "not" is also present, it totally changes the polarity of entity from negative to positive so the output result for this will be good.

Sentence: Company Package is not very bad; Related to: placement; Good: 2; Bad: 0; Feedback: very good

In this way the polarity is find out as well as calculated for the rest of the sentences

Sentence: Trainer should be available for sports activities; Related to: extracurricular activities; Good: 0; Bad: 1; Feedback: Bad

Sentence: space in canteen is very limited; Related to: canteen; Good: 0; Bad: 2; Feedback: Very Bad

Sentence: Playground is not available; Related to: extracurricular activities; Good: 1; Bad: 0; Feedback: good

IV. EXPERIMENTAL RESULT

Precision is one measure of the effectiveness of some computer applications for finding search words, candidate terms, and other items. Precision is a measure of the proportion of the results of a computer application that are considered to be pertinent or correct. For example, the system searches similarity match between the words and the semantic lexicon for the test data set and finds 304 sentences, 285 of which are really correct, then the system precession is 93.57%.

Recall is one measure of the effectiveness of some computer applications for finding search words, candidate terms, and other items. Recall is a measure of the proportion of all possible correct results of a computer application that the application actually produces. For example, suppose you are using a computer application to search for terms in a document that has 80 terms in it. (You know because you counted them.) If the application finds 55 of these terms, then the recall of the application is 55 out of 80, or 0.62. The system searches similarity match between words and semantic lexicon that has 326 sentences in it. If the system finds 303 of these terms, then the recall of the system is 303 out of 326, or 92.90%.

In opinions analysis system these two parameters are evaluated for checking the accuracy of the system. In semantic analysis evaluation the three training dataset were considered of 400, 900, 1100 sentences. The semantic lexicon was trained with these sentences and precision and recall were calculated by varying this dataset. The test dataset contains 326 sentences these all sentences were tested on different semantic lexicon and result was calculated.

Table 6.1: Semantic analysis data

SR NO	TRAINING DATA	PRECISION (%)	RECALL (%)
1	400	93.05	54.29
2	900	99.75	88.34
3	1100	99.87	92.94

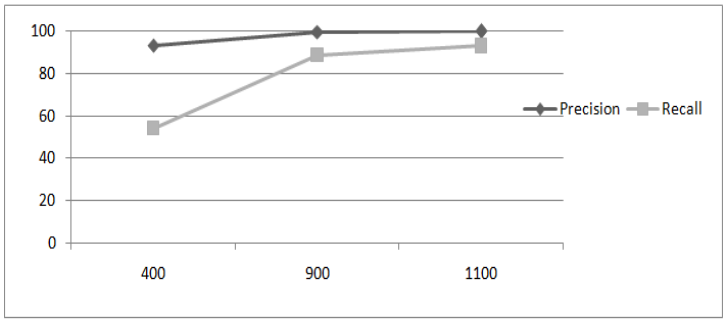

Fig6.1: precision and recall

From the result of the semantic analysis it is understood that the system will work more efficiently with the increase in size of the training data.

In opinion analysis evaluation, all the possible positive, negative, inversion and quantifier words are added to the respective sets. The test dataset of 326 sentences evaluated on the feedback system and the result is calculated in the terms of precision.

The result contains all the entity of the institute and their precision value. The table below shows all the sub domain entity value the number of right and wrong value that have been calculated by the system for each entity based on these precision values a graph is plotted.

Table 6.2: Opinion analysis data

SR NO	ENTITY	CORRECT VALUE	INCORRECT VALUE	PRECISION (%)
1	Canteen	43	11	79.62
2	Infrastructure	34	12	73.91
3	Teaching	29	5	85.24
4	Library	32	9	78.04
5	Lab facility	34	7	82.92
6	Placement	45	7	86.53
7	Activities	38	9	80.85

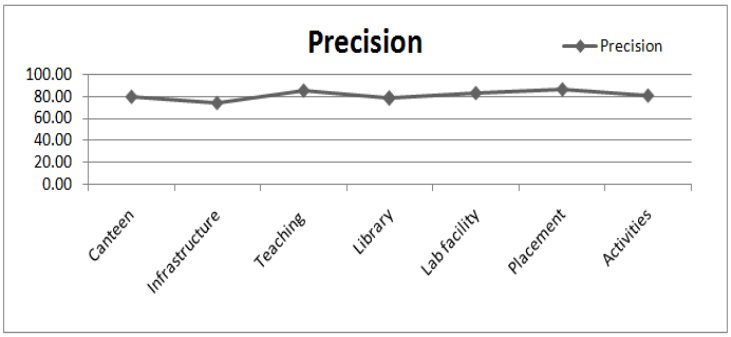

Fig 6.3: Opinion Precision graph

The working of both the semantic analysis and the opinion analysis are evaluated considering the evaluation parameters. Now the result for the system is calculated the result is shown in a more generalized way so that it can be easy for the feedback to understand. The output is shown by six sets very poor, poor, good, excellent, neutral and no comments all the sentences after opinion analysis are split in two these categories so that it will be easy for the institute to understand the problem.

V. CONCLUSIONS

In this paper, proposed Opinions Mining and Summarization System is developed and implemented. The proposed system is tested & evaluated and the result where satisfactory. In the future, the work will be carried out in order to improve the efficiency of the system.

REFERENCES

[1] Katerina Kabassi and Maria Virvou "Combining Decision-Making Theories with a Cognitive Theory for Intelligent Help: A Comparison" IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, VOL. 45, NO. 2, APRIL 2015

[2] Yingcai Wu, Member, IEEE, Shixia Liu, Senior Member, IEEE, Kai Yan, Mengchen Liu, Fangzhao Wu "OpinionFlow: Visual Analysis of Opinion Diffusion on Social Media" IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 20, NO. 12, DECEMBER 2014

[3] Desheng Dash Wu, Lijuan Zheng, and David L. Olson "A Decision Support Approach for Online Stock Forum Sentiment Analysis" IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, VOL. 44, and NO. 8, AUGUST 2014

[4] Fuji Ren, Senior Member, IEEE, and Ye Wu "Predicting User-Topic Opinions in Twitter with Social and Topical Context" IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 4, NO 4, OCTOBER-DECEMBER 2013

[5] Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu, and Emery Jou in their paper "Movie Rating and Review Summarization in Mobile Environment" IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 42, NO. 3, MAY 2012.

[6] Xiaohui Yu, Member, IEEE, Yang Liu, Member, IEEE, Jimmy Xiangji Huang, Member, IEEE, and Aijun An, Member, IEEE in their paper "Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 4, APRIL 2012.

[7] Jingbo Zhu, Member, IEEE, Huizhen Wang, Muhua Zhu, Benjamin K. Tsou, Member, IEEE, and Matthew Ma, Senior Member, IEEE in their paper "Aspect-Based Opinion Polling from Customer Reviews" IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 2, NO. 1, JANUARY-MARCH 2011.

[8] Nathalie Camelin, Associate Member, IEEE, Frederic Bechet, Member, IEEE, Géraldine Damnati, and Renato De Mori, Fellow, IEEE "Detection and Interpretation of Opinion Expressions in Spoken Surveys" IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 18, NO. 2, FEBRUARY 2010.

[9] Thanh-Son Nguyen, Hady W. Lauw, Member, IEEE, and Panayiotis Tsaparas, Member, IEEE "Review Selection Using Micro-Reviews" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 4, APRIL 2015

[10] Uros Krcadinac, Philippe Pasquier, Jelena Jovanovic, and Vladan Devedzic "Synesketch: An Open Source Library for Sentence-Based Emotion Recognition" IEEE Transactions on Affective Computing, Vol. 4, No. 3, July-September 2013

[11] Jelena Jovanovic, Ebrahim Bagheri John Cuzzol Dragan Gasevic, Zoran Jeremic, Reza Bashash "Automated Semantic Tagging of Textual Content" Published by the IEEE Computer Society IT Pro November/December2014 1520-9202/14.

[12] Chenghua Lin, Yulan He, Richard Everson, Member, IEEE, and Stefan Ruger "Weakly Supervised Joint Sentiment-Topic Detection" from Text IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 6, June 2012.

