Opinion Mining of Live Comments from Website using Fuzzy Logic and NLP

Pooja C. Sangvikar[1] Prof. V.S. Khandekar[2]
[1,2]Department of Information Technology
[1,2]Smt. Kashibai Navale College of Engineering Pune, India.

Abstract— For many Natural Language Processing tasks, Opinion Mining of text content is important. In recent years, Social Media plays important role for expressing and sharing of any valuable or important information in terms of text, SMSs, mails, reviews or comments, etc. Existing studies of Opinion Mining tend to extract less features and also used static datasets i.e. publicly available datasets. So the proposed system gives a heuristic approach for Dynamic Comment Classification. The approach focuses on crawling of web pages by entering seed URL and then parsing of web pages is done. After getting user and their comments, preprocessing of comments is done. Various features are extracted like term weight, Noun Identification, Thematic words, Bag of words, etc. Then by applying Fuzzy Logic and IF-THEN rules, comment classification is done: positive and negative. Evaluation is done by the evaluation parameters like Precision, recall and F-measure.

Key words: Web Crawler, Data Preprocessing, Feature Extraction, Fuzzy Logic

I. INTRODUCTION

Opinion Mining is language processing task that uses a computational approach to identify opinionated content and classify it as positive, negative or neutral [1]. The unstructured data on the Web often carries expression of opinions of users in the form of reviews, blogs, comments, etc. Opinion Mining attempts to identify the expressions of opinion and mood of writers. Most of the current Opinion Mining research is focused on business and e-commerce applications, such as review of products and movie reviews. Few researches have tried to understand opinions in the social and geopolitical context.

But it becomes more difficult for web users to find valuable or important information in such a huge repository when the quantity of evaluative texts expands, so sentiment classification becomes important. Sentiment classification has been applied to many areas. It is used to annotate the sentiment content in text, categorize opinions in product reviews, etc. Some of other terms used in previous papers are sentiment analysis, opinion extraction and affect analysis. Sentiment classification has become an overlapping research issue in multiple research areas, such as Data Mining (DM), Machine Learning (ML), and so on [3].

Sentiment analysis is an area of research that is closely related to text analytics, natural language processing (NLP), computational linguistics (CL), and information retrieval (IR). The general aim of sentiment analysis is to determine/extract the opinion contained within a piece of text. There has been an increase in popularity for sentiment analysis in recent years, mainly due to the many practical applications it supports. For example: tracking opinions in online forums, blogs and social networks; helping companies and organizations find customer opinions of commercial products or services; and helping individuals decide on which product to buy or which movie to watch.
Sentiment classification has increasingly gained attraction in recent years. It aims to divide text into different emotional polarities, such as positive, negative and neutral. Major approaches for sentiment classification fall into two categories: lexicon based methods and machine learning based methods. The performance of lexicon based methods strongly relies on sentiment lexicon [2]. As it is costly to build sentiment lexicons manually, most previous work has focused on the automatic or semi- automatic construction of sentiment lexicons. For machine learning based methods, sentiment classification is often treated as a traditional text categorization problem, and it's important to extract useful textual features for machine learning algorithms [3].

The remainder of this paper is organized as follows: Section II describes literature survey or related work on Sentiment Analysis. Section III includes overview of proposed system. Section IV describes the Results and Discussion. Section V gives Conclusion and Future Work.

II. RELATED WORK

Sentiment Analysis has been done using a different techniques or methods. Some works extract the meaning of the text, document, sentence or phrase level while others obtain connections between users to assign sentiment polarity for sentiment analysis. Many different approaches to solve sentiment analysis have been developed by researchers from information retrieval, most of which in this field use bag of words representations. In particular, Opinion Mining of tweets has been done using approaches based on text, which is lexicon based classifiers, also by combining Natural Language Processing and Machine Learning techniques. A lot of research has been done in this field by researchers and scholars all around the world.

Text Mining and Sentiment Analysis have received a great attention due to abundance of opinion data that exists in social networks such as Facebook, twitter, etc. Here author Akaichi (2013) in [4] focused on mining of Facebook status updates. For this, they constructed sentiment lexicon based on interjections, acronyms and emoticons. There are five main steps are followed: raw data collection, lexicon developments, feature extraction, training model for text polarity creation and machine learning method application. To evaluate the performance of sentiment classification, accuracy is calculated on different feature sets.

Stopword removing is one of the frequently used step in preprocessing. Stopwords are periodically occurring words that rarely carry any information and orientation. Ghag and Shah (2015) in [5] the effect of stopword removal on various sentiment classification models was analyzed. Sentiment Classification model were analyzed using movie review dataset. Classifiers focuses on proportional presence count distribution and proportional frequency count

distribution where as traditional approaches such as delta TFIDF and alternative term weighting techniques.

Cai and Spangler (2008) in [6] focused on techniques that detect the topics that are highly correlated with positive and negative opinions. By coupling this technique with Sentiment classification, sentiment score is calculated. For topic detection, Point-wise mutual information and term frequency distribution method is used.

A dependency Tree based Sentence-level Sentiment classification approach is presented by Li (2011) in [7]. Here flat features (Bag-of-words) is captured as well as structured features from dependency tree of a sentiments. Author introduced a convolution tree kernel based approach to the sentence level Sentiment classification. This approach achieved improvement for implicit Sentiment Classification. To identify the polarity SVM classifier is used. An approach which adopts empirical learning to implement the Sentiment Classification technology and used a distance based predictive model to bind computational efficiency and modularity proposed by Bisio (2013) in [8].

Li (2013) [9] performed Sentiment Classification with full consideration of polarity shifting phenomenon. Firstly, extraction of some detection rules for detection of polarity shifting of sentimental words from polarity shifted words in testing data, detection rules are applied. Lastly, term counting based classifier is designed by using polarity shifted words.

Lin (2015) [10] proposed a personality based Sentiment Classification method to capture more useful but not widely used sentiment words. To utilize both personality related and commonly used textual features, they adopt an ensemble learning strategy. Allocation of tweets is done to different groups according to personality traits of users for each group, Random Forests is trained separately.

Liu (2011) in [11] presented a Novel Approach for News Video Story Sentiment Analysis. In this, two challenges are addressed: new video story Sentiment Classification and ranking. Graph based approach is used to classify the news stories into sentiment classes. To add news videos into sentiment space, a multimodal approach is used and to rank the videos in each class visual representation scores is adopted. For sentiment representation, sentiment class analysis is done based on PageRank algorithm and affinity propagation clustering.

Mouthami (2013) in [12] proposed a New algorithm called Sentiment Classification algorithm with POS tags is used to improve classification accuracy on Movie Reviews Dataset. It approximately classifies the Sentiment using Bag of Words. Su (2012) in [13] explained a Semi-supervised learning method based on multi-view learning. Idea of approval is to generate multiple views by accomplishing both feature partition and language translation strategies and after that to perform multi-view learning for Semi-Supervised Sentiment Classification standard co-training algorithm is applied. To generate different views two strategies are used: Feature Partition which splits whole and other strategy is language translation which translates original text into another language.

III. PROPOSED METHODOLOGY

A. *Overview*

The proposed method of Opinion Mining of live comments from websites using fuzzy logic and NLP is described efficiently according to the steps which are depicted in the Fig.1.The following steps are used for comment classification.

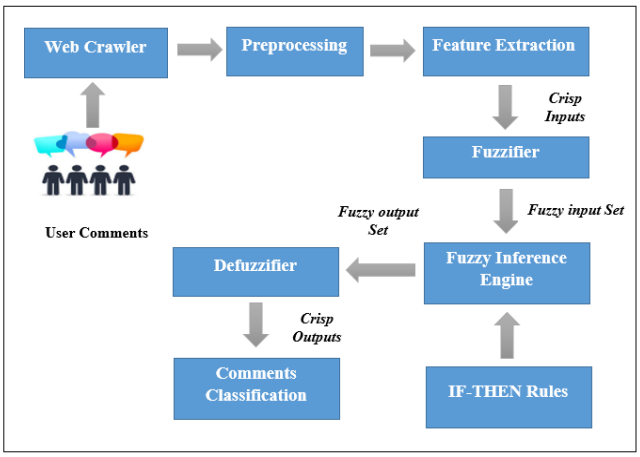

Fig. 1: Overview of Fuzzy based Classification System

B. *Algorithm*

The input to algorithm is seed URL which is T_n= {x | x ϵ Ti, where i=1, 2, 3 …n} to receive the comments from websites. The output generated by the proposed algorithm is classification of comments into positive and negative.

C. *Data Preprocessing*

1) Read input data (I) with the help of web crawler Where I= $T_{i...n}$ i.e. different URL for extracting comments.
2) Collect the contents of web pages and parse that web page by using efficient parser Remove special symbols like #, &, @, etc.
3) Remove all the stop words from comments like is, are, but, etc.
4) Convert the words into stem form i.e. studied to study, where ied is replaced with y. Output is the preprocessed string (R_s). Repeat these steps for each comment.

D. *Feature Extraction*

1) Get the vector from keyword database and calculate the repeated words in the comment. Calculate term weight (C_t) for each repeated word by following equation (1).

$$C_t = Tf_i * Isf_i \quad \text{(1)}$$

Where,
Tf_i = term frequency of each repeated word.
Isf_i = inverse sentence frequency of word.

$$Tf_i = \frac{nj}{\sum nk}$$

Where,
n_j = the number of occurrences of the term j.
n_k = the number of occurrences of all terms in the comment.

$$Isf_i = \log \frac{N}{ni}$$

Where,
N = Total number of comments.
n_i = number of comments in which word i arises.

Now, divide string (R_s) into words and store in a vector V. Identify the duplicate words in the vector and remove them. For each word check for its occurrence in Dictionary and calculate the score (Pn).

Next step is thematic words (Tw) identification. It can be calculated as the proportion of the number of thematic words that arise in the text over the maximum summary of thematic words in the text.

Bag words, here database is maintained for good words and negative words. If the words from comments are matched with these database words, then calculate the scores (Gw) and (Bw) for that comment. Repeat these steps for each comment.

E. Fuzzy Classification

In order to implement comment classification based on fuzzy logic, first, the features extracted in the feature extraction step are used as input to the fuzzifier. Triangular membership functions and fuzzy logic is used to summarize the document.

The input triangular membership function for each feature is divided into five fuzzy sets which are composed of unimportant values low (L) and very low (VL), average values (medium (M)) and important values high (H) and very high (VH). Fuzzy crisp values are created as shown below as example:

Very Low (Vl) - 0.0 To 0.2
Low (L) - 0.2 To 0.4
Medium (M) - 0.4 To 0.6
High (H) - 0.6 To 0.8
Very High (Vh) - 0.8 To 1.0

In inference engine the most important part is the definition of fuzzy IF-THEN rules. The important sentences are extracted from these rules according to features criteria. Sample of IF-THEN rules shows as the following rule.

IF (Term weight is H) and (No Proper Noun is VH) and (No Thematic Word is H)
THEN (Comment is Important)
IF (Term weight is VL) and (No Proper Noun is M) and (No Thematic Word is L)
THEN (Comment is Unimportant)

Likewise, the last step in fuzzy logic system is the defuzzfication. The output membership function which is used to convert the fuzzy results from the inference engine into a crisp output for the final score of each sentence. Then after defuzzification, the classification of comments is done.

IV. RESULTS AND DISCUSSION

The Fuzzy based comment classification system uses customer reviews about some restaurants effectively. A review is a subjective text containing a sequence of words describing opinions of reviewer regarding a specific food, services, etc. Review text may contain complete sentences, short comments, or both. Restaurants reviews are collected from websites like Woodland, Orchid, Sapana and Mathura hotels from pune. The fuzzy based comment classification system is compared with the aspect based classification using frequent item set mining [15].

In the context of classification, True Positives (TP), True Negatives (TN), False Negatives (FN) and False Positives (FP) are used to compare the class labels assigned to documents by a classifier with the classes the items actually belong to.

True positive means, which are truly classified as the positive terms. True positives (TP), the classifier correctly labeled as belonging to the positive class.

False positive (FP) means which were not labeled by the classifier as belonging to the positive class but should have been.

True Negative (TN) is that the classifier correctly labeled as belonging to the negative class. True Negative means, which are truly classified as the Negative terms.

False Negative (FN), which is nothing but example which was not labeled by the classifier as belonging to the negative class but should have been. Evaluation measures like precision, recall, F-measure can easily be calculated from these four variables.

1) Precision: Precision and recall are two widely used metrics for evaluating performance in text mining. Precision is used to measure exactness. Precision is the number of examples correctly labeled as positive divided on the total number that are classified as positive. This is shown in the following formula.

$$\text{Precision: } \frac{TP}{TP+FN}$$

2) Recall: Recall is a measure of completeness. while recall is the number of examples correctly labeled as positive divided on the total number of examples that truly are positive. This is shown in the following formula.

$$\text{Recall: } \frac{TP}{TP+FP}$$

3) F-Measure: F-Measure is the harmonic mean of precision and recall. This gives a score that is a balance between precision and recall. F-Measure combines them into one score for easier usage.

$$F= \frac{2\times precision\times recall}{precision+recall}$$

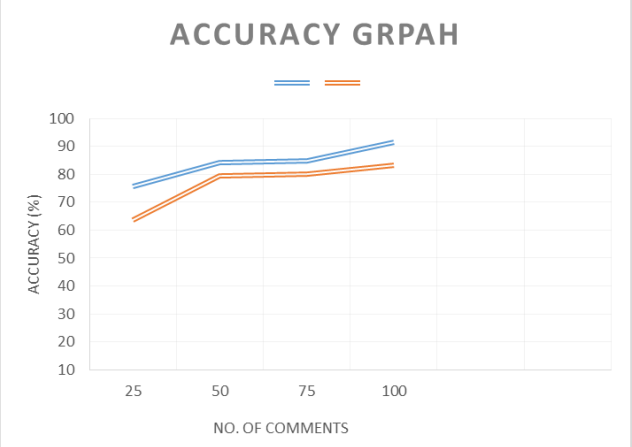

Fig. 2: F-measure of Fuzzy based Classification System

The experiments are performed on web pages containing comments in the range of 25,50,75,100 and F-measure is calculated for each web pages as shown in Fig.2.

The system [15] implements aspect extraction using frequent item set mining in customer product reviews and gives the precision 75%, recall 85.71% and F-measure 80.36%. The fuzzy based Comments classification system gives better results than the aspect based opinion Mining such as precision 84.18%, F-measure 82.08% as shown in Fig.3.

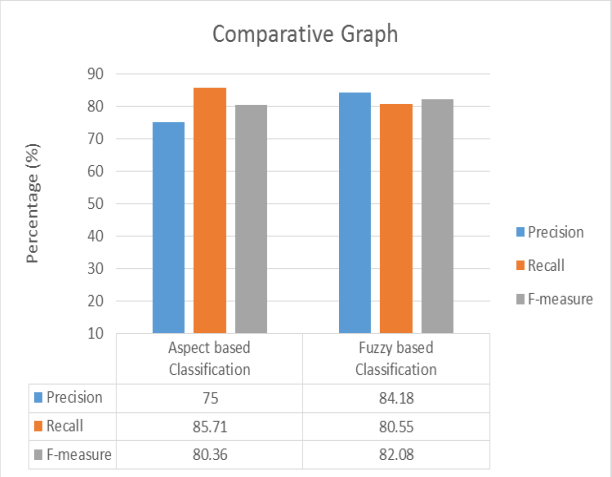

Fig. 3: Comparative Results

V. CONCLUSION AND FUTURE SCOPE

The proposed system gives heuristic and general approach named Opinion Mining of live comments from the website using Fuzzy Logic and NLP. Here the focus is on comments which are used for classification and experiments are performed on some web pages. The accuracy achieved by the proposed system with the help of evaluation parameters like Precision, Recall and F-measure are 84.18%, 80.55% and 82.08%. Here, proposed system uses HTTP protocol based websites for implementation. In future, HTTPs protocol based sites (Yahoo, Twitter, etc.) can be taken into consideration by taking authorization as these are secured sites and also by combining Fuzzy based classifier with other classification techniques, accuracy can be achieved.

REFERENCES

[1] https://en.wikipedia.org/wiki/Sentiment_analysis.
[2] Walaa Medhat, Ahmed Hassan, Hoda Korashy, "Sentiment analysis algorithms and applications: A survey", 2014 Production and hosting by Elsevier, Ain Shams Engineering Journal (2014), Volume 5, Issue 4.
[3] https://en.wikipedia.org/wiki/Machine_learning
[4] Jalel Akaichi, "Social Networks' Facebook' Statutes Updates Mining for Sentiment Classification",SocialCom/PASSAT/BigData/EconCom /BioMedCom 2013, © 2013 IEEE.
[5] Ms. Kranti Vithal Ghag, Dr. Ketan Shah, "Comparative Analysis of Effect of Stopwords Removal on Sentiment Classification", IEEE International Conference on Computer, Communication and Control (IC4-2015).
[6] Keke Cai, Scott Spangler, Ying Chen, Li Zhang, "Leveraging Sentiment Analysis for Topic Detection", 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008 IEEE.
[7] Peifeng Li, Qiaoming Zhu, Wei Zhang, "A Dependency Tree based Approach for Sentence-level Sentiment Classification", 2011 12th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, © 2011 IEEE.
[8] Federica Bisio, Paolo Gastaldo, Chiara Peretti, and Rodolfo Zunino, "Data Intensive Review Mining for Sentiment Classification across Heterogeneous Domains", 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, A SONAM'13, August 25-29, 2013, Niagara, Ontario, CAN.
[9] Shoushan Li, Zhongqing Wang, Sophia Yat Mei Lee, Chu-Ren Huang," Sentiment Classification with Polarity Shifting Detection", 2013 International Conference on Asian Language Processing, © 2013 IEEE.
[10] Junjie Lin, Wenji Mao, "Personality based Public Sentiment Classification in Microblog", ©2015 IEEE.
[11] Chunxi Liu, Li Su, Qingming Huang, Shuqiang Jian," News Video Story Sentiment Classification and Ranking", ©2011 IEEE.
[12] Ms. K. Mouthami, Ms. K. Nirmala, Devi Dr. V. Murali Bhaskaran," Sentiment Analysis and Classification Based On Textual Reviews", © 2013 IEEE.
[13] Yan Su, Shoushan Li, Shengfeng Ju, Guodong Zhou," Multi-view Learning for Semi-Supervised Sentiment Classification", 2012 International Conference on Asian Language Processing, © 2012 IEEE.
[14] Vo Ngoc Phu, Phan Thi Tuoi," Sentiment Classification using Enhanced Contextual Valence Shifters", © 2014 IEEE.
[15] A. Jeyapriya, C.S. Kanimozhi Selvi, "Extracting Aspects and Mining Opinions in Product Reviews using Supervised Learning Algorithm", Ieee Sponsored 2nd International Conference On Electronics And Communication Systems (Icecs '2015).

