IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Vol. 08, Issue 6 (June. 2018), ||V (VII) || PP 01-07

Feature Based Sentimental Analysis on Mobile Web Domain

Shoieb Ahamed[1] Alit Danti[2]
1. Government First Grade College, Sorab(T), Shimoga(D), Karnataka, India
2. NES Research Foundation, JNN College of Engineering, Shimoga, Karnataka, India.
Corresponding Author: ShoiebAhamed

Abstract -Sentimental analysis refers to study of analyzing the online reviews in a scientific and structured way. In this work, the proposed is based on SentiWordNet, which produces count of score words into seven categories such as strong-positive, positive, weak-positive, neutral, weak-negative, negative and strong-negative words for the sentimental analysis task and assess performance of various machine learning algorithms like K-Nearest Neighbour(IBK), AdaBoost, Naïve Bayes (NB) and Support Vector Machine (SVM)/SMO algorithms.Online data is collected using web crawler applied with different pre-processing techniques like stop-words removal followed by stemming process, and then reviews are tagged using POS tagger. The proposed approach is experimented on different mobile product web domains and obtained higher success rate in terms of accuracy measured. The experimental results are tested using Ten-Fold cross validation on the training data. The results demonstrate that the proposed approach has higher efficacy and it can be successfully used in Sentimental analysis for the task of online decision.

Keyword - Sentiment Analysis, Opinion mining, POStagging, Mobile Domain.

Date of Submission: 15-06-2018 Date of acceptance: 29-06-2018

I. INTRODUCTION

Opinion mining is an art of tracking the mood of the public about a particular product or topic from a huge set of opinions or reviews publically available in web. So analysis or mining of opinion is necessary. Opinion is nothing but the person's feeling or sentiment or attitude towards certain topic. Suppose a person is interested to purchase a product, collect information in terms of opinions from people. But from huge collection of opinions it is difficult to derive a conclusion whether the product is good or bad. So mining of opinion is developed and from opinion mining goodness and badness about the product can be concluded which will help all the online customers for buying best products and also benefit online product suppliers to create a benchmark and increase their sales.

Sentiment analysis denotes to the usage of natural language processing (NLP), text analysis and computational linguistics to identify and extract subjective information from web data. There are also different synonyms and slightly different tasks, e.g., sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc. Bing Liu [4], in his work related sentiment analysis with opinion mining by stating "Sentiment analysis is the field of study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes".

SentiWordNet is a publicly available lexical resource for research purposes providing a semi-supervised form of sentiment classificationbased on the annotation of all the synsets of WordNet according to the notions of "positivity", "negativity", and "neutrality". Fig. 1 shows the sample output of SentiWordNet[1].

```
not---r---0.625
yet---r---0.01615646258503402
quality---n---0.3521897878
students---n---null
here---r---0.0
certainly---r---0.25
meet---v---0.065183317001
standard---n---0.012755102
```

Fig. 1: Sample Output of SentiWordNet.

II. LITERATURE SURVEY

In the area of sentiment analysis various research work is been carried out. Pang B et al., [6] have implemented the machine learning techniques to classify movie reviews according to the sentiments. They have used POS Tagger for taggingand SentiWordNet for scoring and analysis of movie reviews. In this work, only number of positive scored words, number of negative scored words, and number of neutrally scored words are considered and further classified is done using SentiWordNet.

Shoiab Ahmed and Ajit Danti[1] have developed a novel approach for doing sentimental analysis of web various domains and have concluded that efficacy of various machine learning algorithms on various statistical measures. They have used a unique seven category based scoring method.PreetiRoutrayet al.,[7] have made a survey on various approaches followed for sentimental analysis. This method includes using of WordNet, Support Vector Machine, Naive Bayes, Maximum Entropy, language Model for performing sentimental analysis of data.Dave et al., [3] developed a document level opinion classifier that uses statistical techniques and POS tagging information for sifting through and synthesizing product reviews, essentially automating the sort of work done by aggregation sites or clipping services.

III. PROPOSED METHODOLOGY

In the proposed methodology, first we create the data sets which are available online from various web forums. In this work the datasets from Tourism Domain and from Product domains are collected. The dataset created refers to the opinions of the web users regarding the various tourism places based on their visit experiences and the reviews of the web users regarding buying and usage of various mobile phones coming from different mobile brands.

The opinions are collected with the help of Web Crawler. The collected reviews are stored separately in text format. These documents are then given as input for the preprocessing task which includes stop words removal, stemming followed by POS tagging.Even though existing SentiWordNet gives the scores and their count of the words, the major drawback is it may not give a complete solution for the task of sentimental analysis. To overcome this in opinion mining, in this approach count of score words are used based on seven categories.Analyzing of web reviews using the score count of these seven score words on any web domain in addition to applying of various machine learning techniques will give better results. Therefore, this approach will be a more powerful aid for the web users to take decision in their web data mining like selecting a best mobile.

For example, the sample output generated for the online movie data reviewby novelproposed approach using SentiWordNet[2] is shown in Fig. 2. Each line represents a word's score value and its associates score category.

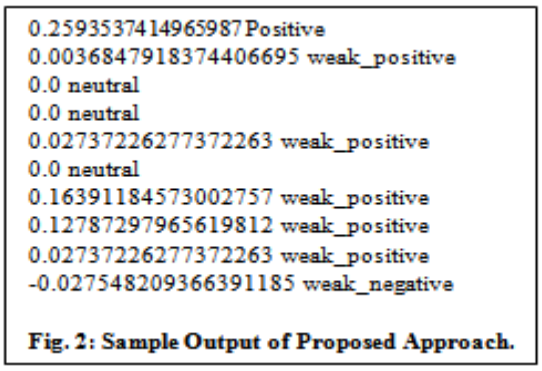


Fig. 2: Sample Output of Proposed Approach.

This approach is implemented on online mobile product reviews datausing the web mining tool Web Crawler. Once the input data reviews are ready, it needs to undergo a process ofpre-processing before using it for the task of opinion mining. The overall architecture of the proposed sentimental analysis and opinion mining model is illustrates in this section. The Fig. 3 shows architecture of proposed method.

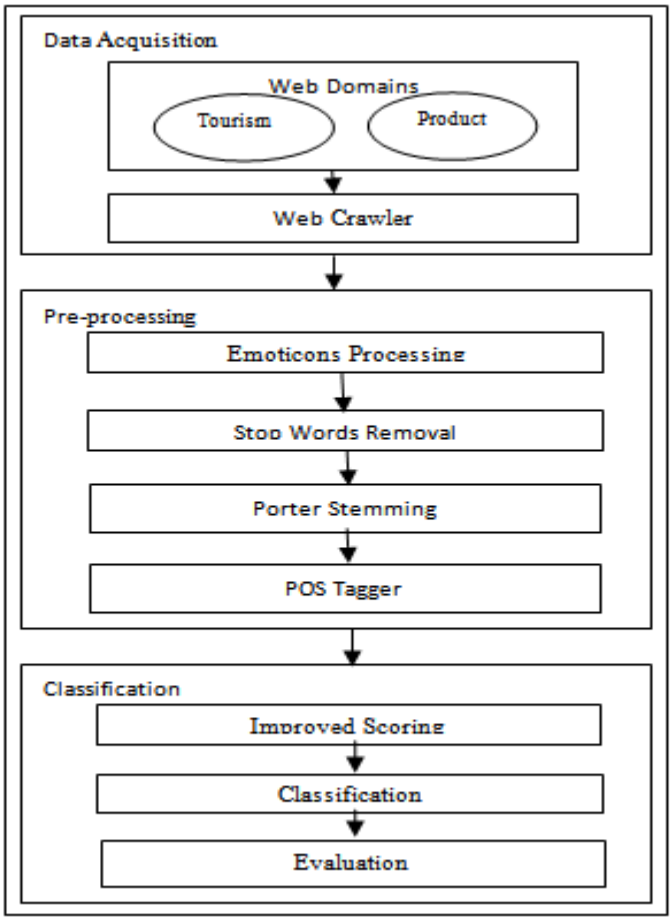


Fig. 3: Architecture of Proposed Approach.

3.1. Data Acquisition:

This is the initial step involved in any opinion mining and sentiment analysis since the quality of datasets is very important measure. Huge amount of data from web iscollectedusing Web Crawler for getting the onlinereviews. The HTML Parser parses each page and eliminates all tag information and generates plain text sentences. The collected reviews are stored in the form of text documents.

3.2 Pre-Processing

1) Stop words Removal:

These are the words which are filtered out prior to or after processing of natural language data. These stop words can cause problem while searching for phrases that include them. There are some of the most common stop words such as the, is, at, which, on etc. Therefore such words are not necessary to carry our purpose. Those words are removed by using simple coding. In this work more than 50 stop words are removed from our document as they are irrelevant for our purpose.

2) Stemming:

Online reviews are generally used with informal language and they include internet slang and contemporary spellings like use of apostrophes, ing form of words to name a few. So such words must be re-visited and stemmed for correct data retrieval. There are several types of stemming algorithms which differ in respect to performance and accuracy.

In this work,Porter Stemming algorithm is adapted which utilizes suffix stripping. It does not address prefixes. Steps involved in this algorithm are given below.

Step 1: Remove plurals and -ed or -ing suffixes
Step 2: Turns terminal y to i when there is another vowel in the stem
Step 3: Maps double suffixes to single ones: -ization, -ational, etc.
Step 4: Deals with suffixes, -full, -ness etc.
Step 5: Takes off -ant, -ence,etc from the suffixes.
Step 6: Removes a final –e from given suffixes.

3) Parts Of Speech Tagging:

The reviews are tagged by their respective parts of speech using POS Tagger. A POS tagger parses a string of words, (e.g., a sentence) and tags each term with its part of speech using the equation (1) and (2). For example, every term has been associated with a relevant tag indicating its role in the sentence, such as VBZ (verb), NN (noun), JJ (adjective) etc.

$$T(w_{i,n}) = \arg max \max_{t1,n} xe^{-x^2} \prod_{i=1}^{n} P(w_i|w_{i,i-1}) \quad (1)$$

$$T(w_{i,n}) = \arg max \max_{t1,n} xe^{-x^2} \prod_{i=1}^{n} P(t_i|w_i) \quad (2)$$

Where'w'represents the word, 't_i' represents the word tag. The tagging problem is defined as the sequence of tags as 't_i, n'.

In this work Stanford POS Tagger is adaptedand its conversion to SentiWordNet Tags is shown in Table 1.

3.3 Classification

1) SentiWordNet:

Table 1: Conversion of POS tags to SentiWordNet tags

SentiWordNet Tag	**POS tag**
a (adjective)	JJ, JJR, JJS
n (noun)	NN, NNS, NNP, NNPS
v (verb)	VB,VBD, VBG, VBN, VBP, VPZ
r (adverb)	RB, RBR, RBS

The SentiWordNet is a lexical resource, where each WordNet synset 's' is associated to two numerical scores Pos(s) and Neg(s), describing how positive or negative terms contained in the synsets. If the analyzer finds a pre-defined keyword in a sentence of a given blog page for a specific gadget, it looks for the modifiers associated with that keyword. If it finds such a word, it obtains its score from SentiWordNet for further process.

2) Scoring:

For opinion mining,a novelapproach using SentiWordNetis proposed that produces the count of scored words by classifying them into the seven possible categories i.e., strong-positive, positive, weak-positive, neutral, weak-negative, negative and strong-negative words. These scorecounts are used to perform the task of sentimental analysis.

IV. EXPERIMENTAL RESULTS

The source of input for the work is reviews of the users in the opinionated sites. The opinion of people i.e. reviews of users are collected from various domains such as Product. Since the web domains are in the form of HTML, if we want to retrieve the user reviews from the specific domain then we must parse the HTML document specified in the URL. To work with the real-time HTML the java library called Jsoup is used. After parsing the HTML documents the collected reviews are stored in text file.

4.1 Dataset:

In this work, the product Review Dataset is prepared by collecting the reviews of the web users ondifferent mobiles like Micromax A116 Canvas HD, Motorola Moto E, Nokia Lumina 920 and Asus Zenfone5 for comparison from the website "www.mouthshut.com". Reviews of the 200 online users are collected forexperimentation.

For this datasets the pre-processing is done. In which stop words are removed and stemming is performed. Then StanfordPOS tagger is employed to tag the dataset with their parts of speech. The tagged documents are fed as input to SentiWordNet for scoring and count of each scored word is determined.

Experimental results reveal the efficiency of the proposed approach.The Table 2 shows the count of scored words for online reviewsof mobile product datasets.Table 3 shows the Precision Values for mobile product datasets obtained by different machine learning algorithms.

Table 2: Score count on Mobile Product Reviews Datasets.

ScoreClass	ASUS	Micromax	Motorola	Samsung
Strong Positive	1	3	4	1
Positive	33	16	31	46
Weak Positive	75	98	102	90
Neutral	229	185	159	166
Weak Negative	49	45	46	38
Negative	5	15	6	9
Strong Negative	2	0	2	1

Table 3: Precision Value for Mobile Product reviews.

Algorithm	ASUS	Micromax	Motorola	Samsung
NB	82.6	90.5	80.8	83.3
SMO	44.3	54.7	49.8	51.6
IBK	99	98.9	98.9	97.9
Adaboost	59.9	62.2	62.2	54
Total Instances	394	362	350	351

Figure 4:Score classes of Mobile product Reviews

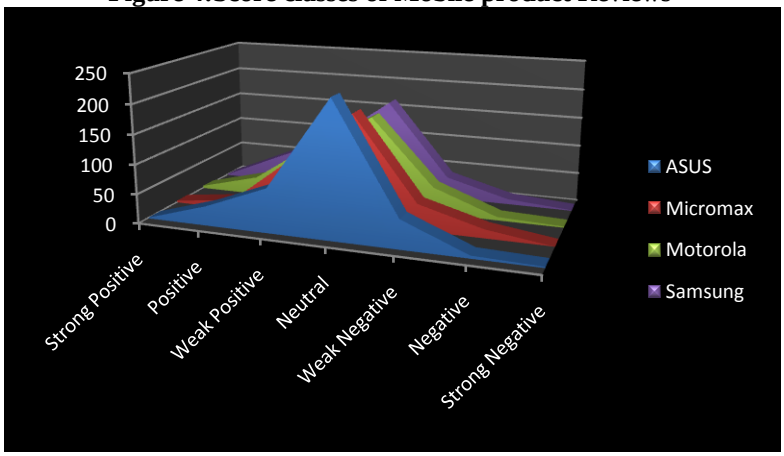
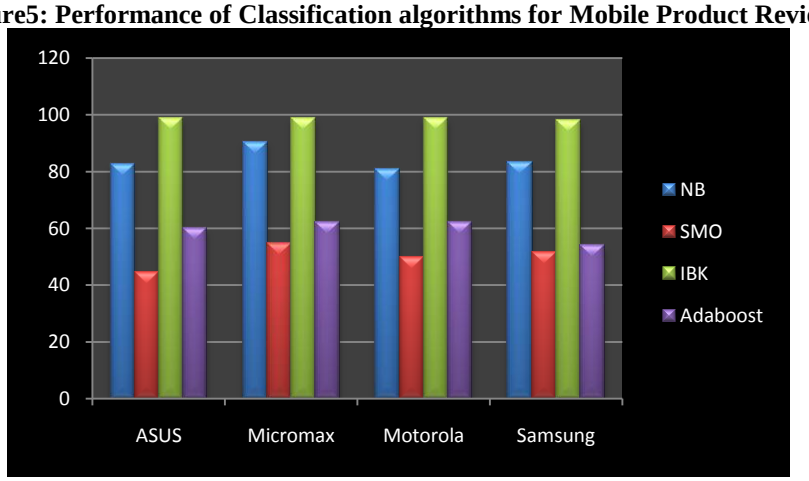


Figure5: Performance of Classification algorithms for Mobile Product Reviews



4.2 Evaluation:
Proposed approach is evaluated by various measures and their accuracy is shown in the Table 4.

Table 4: Different Accuracy Measures for online movie reviews

Sl. No.	Measures	NB	SMO	IBK	AdaBoost
1.	Kappa Statistic	77.7 %	35.37 %	89.3 %	82.2 %
2.	Mean Absolute Error	20.15%	21.6 %	15.8 %	7.21 %
3.	Root Mean Squared Error	27.3 %	32.1 %	17.2 %	19.3 %

Kappa Statistic compares the accuracy of the system to the accuracy of a random system using equation (3).

$$\text{(Kappa) } K = \frac{a_0 - a_e}{1 - a_e} \quad (3)$$

Where 'a_0' is observed accuracy and 'a_e' is expected accuracy.

Means Absolute Error (MAE) is the average absolute difference between classifier predicted output and actual output as given in the equation (4).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}(Desired_i - Actual_i) \le \varepsilon \quad (4)$$

Root Mean Square Error (RMSE) is a frequently used measure of the differences between value predicted by a model and the values actually observed. RMSD is a good measure of accuracy as given by the equation (5).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Desired_i - Actual_i)^2} \le \varepsilon \quad (5)$$

V. CONCLUSION AND FUTURE ENHANCEMENT

This work showcases a novel approach which will be useful for the users for taking decisions based on the online reviews available in the web. The proposed approach is experimented on Mobile online reviews and experimental results reveal the efficiency of the proposed approach along with accuracy.

This work can be further enhanced by applyingvarious feature selection techniques for classification like Information gain, Categorical Proportional differences,Chi Square, etc.

REFERENCES

[1]. Shoiab Ahmed, Ajit Danti –"Effective Sentimental Analysis and Opinion Mining of Web Reviews Using Rule Based Classifiers"; international conference on Computational Intelligence in Data Mining, Volume1, Pages171-179, SPRINGER publications, DOI10.1007/978-81-322-2734-2_18, December 2015.

[2]. Ajit Danti, Shoiab Ahmed – "A Novel Approach for Sentimental Analysis and Opinion Mining based on SentiWordNet using Web Data", International Conference on Trends in Automation, Communication & Computing Technologies - ITACT 15, pp, 07-11, IEEE Xplore, ISBN:978-1-5090-1887-1, December 2015.

[3]. Dave K., Lawrence S., Pennock D. M., "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews", in Proceedings of the 12th International Conference on World Wide Web, ACM, pp-519–528, 2003.

[4]. Liu B., "Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies", San Rafael, Calif, Morgan & Claypool,5(1): pp-1–167, 2012.

[5]. Ohana B., Tierney B., "Sentiment Classification of Reviews using SentiWordNet", in 9th IT&T Conference, Dublin Institute of Technology, Dublin, Ireland, pp-13, 2009.

[6]. Pang B., Lee L., Vaithyanathan S., "Thumbs up?: Sentiment Classification using Machine Learning Techniques", in Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing-Volume 10, Association for Computational Linguistics, pp-79–86, 2002.

[7]. PreetiRoutray, Chinmaya Kumar Swain and SmitaPrava Mishra, "A Survey on Sentiment Analysis ", International Journal of Computer Applications (0975 – 8887) Volume 76 – No.10, August 2013.
[8]. Turney P. D., "Thumbs up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews", in Proceedings of the 40th AnnualMeeting on Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, pp-417–424, 2002.
[9]. H. A. Ahmed Aqlan, Shoiab Ahmed, Ajit Danti "Death Prediction and Analysis Using Web Mining Techniques". 2017 International Conference on Advanced Computing and Communication Systems (ICACCS -2015), Coimbatore, INDIA, Jan. 06 – 07, 2017.

AUTHOR'S PROFILE

Mr. Shoiab Ahmed
is currently working as Assistant Professor in the Department of Computer Science, Government First Grade College, Sorab(T), Shimoga, Karnataka, India. Area of research includes Web Mining, Sentimental Analysis, and Data Science. He has Completed Master's Degree in Computer Applications(M.C.A) from Visvesvaraya Technological University, Belgaum in the year 2008 and has qualified in Karnataka State Eligibility Test (KSET) - 2014 accredited by UGC.

Dr. Ajit Danti
is currently working as Director and Professor in the Dept. of Computer Applications, Jawaharlal Nehru National College of Engineering, Shimoga, Karnataka, India. He has 25 years of experience in Teaching, Administration and Research. Area of research include image processing, Pattern Recognition and Computer Vision. He has published more than 73 research papers in the international Journals and Conferences. He has authored two books published by Advance Robotics International, Austria (AU) and Lambert Academic Publishing, German which are freely available online. He has Completed Ph.D degree from Gulbarga University in the field of Human Face Detection & Recognition in the year 2006. He has Completed Master's Degree in Computer Management from Shivaji University, Maharashtra in the year 1991 and M.Tech from, Mysore in the year 2011.

IOSR Journal of Engineering (IOSRJEN) is UGC approved Journal with Sl. No. 3240, Journal no. 48995.

ShoiebAhamed "Feature Based Sentimental Analysis on Mobile Web Domain." IOSR Journal of Engineering (IOSRJEN), vol. 08, no. 6, 2018, pp. 01-07.