Journal of Network Communications and Emerging Technologies (JNCET) www.jncet.org
Volume 5, Special Issue 2, December (2015)

Sentimental Analysis of Product Based Reviews Using Machine Learning Approaches

Manvee Chauhan
Department of Computer Science & Engg., Jaypee Institute of Information technology Noida, India

Divakar Yadav
Department of Computer Science & Engg., Jaypee Institute of Information technology Noida, India

Abstract – With the fast growth of e-commerce, large number of products is sold online, and a lot more people are purchasing products online. People while buying also give feedback of product purchased in form of reviews. The user generated reviews for products and services are largely available on internet. Since information available on internet is so widespread we need to extract the needful information for which we make use of sentimental analysis. Sentiment analysis extracts abstract and to the point information required for source materials by applying concept of Natural language processing. It is used to deal with identification and aggregation of the opinions given by the customers. These reviews play vital role in determining potential customer for the products as well as market trend for product. This paper provides summary of reviews for products by classifying these reviews as positive, negative or neutral. Information on internet is highly Since reviews are highly unstructured, machine learning approaches are applied including naïve Bayes and support vector machine algorithms by first taking inputs as unstructured product reviews, performs preprocessing, calculates polarity of reviews, extracts features on to which comments are made and also plots graph for the result. The algorithms precision, recall and accuracy are measured finally.

Index Terms – Machine Learning, Semantic Orientation, Sentiment Analysis, Support Vector Machine, Naïve Bayes.

This paper is presented at International Conference on Recent Trends in Computer and information Technology Research on 25th & 26th September (2015) conducted by B. S. Anangpuria Institute of Technology & Management, Village-Alampur, Ballabgarh-Sohna Road, Faridabad.

1. INTRODUCTION

In past days, purchasing of products was more based on getting product review from nearby neighbors, relatives etc. as products were purchased directly from merchants. People believed relatives, and friends review about product helpful. But with change in technology, we saw development of E-commerce industry with sites flooded by products from different brands made available to customers at the touch of one click. The availability of product based sites with doorstep delivery has made it convenient for customers to shop online. It provides one stop shop for all needs of customers. With so much change in shopping pattern, we see merchants providing customers with feedback option about the product. Customers write reviews from all parts of the world. There are thousands, millions of reviews being written. So a question arises on how to get fundamental judgment about product without going through each of them separately.

A lot of reviews are very long, making it difficult for a potential customer to review them to make an informed decision on whether the customer should purchase the product or not. A vast number of reviews also make it difficult for product manufacturers to keep log of customer opinions and sentiments expressed on their products and services. It thus becomes necessity to produce a summary of reviews. Summarization of reviews is done using sentiment analysis.

Sentiment analysis tends to extract subjective information required for source materials by applying natural concept of natural language processing [4]. The main task lies in identifying whether the opinion stated is positive or negative. Since customers usually do not express opinions in simple manner, sometimes it becomes tedious task to judge an opinion stated. Some opinions are comparative ones while others are direct.

Sentimental analysis helps customer visualize satisfaction while purchasing by simple summarization of these reviews into positive or negative- two broader classified classes. Feedbacks are mainly used for helping customers purchase online and for knowing current market trends about products which is helpful for developing market strategies by merchants.

In this paper, we examine the effectiveness of applying machine learning techniques to the sentiment classification

problem. Machine learning is divided into: supervised and unsupervised approaches [1]. Supervised learning tends to be more accurate because each of the classifiers is trained on a collection of representative data known as corpus in contrast to unsupervised learning which does not require prior training. In order to mine the data instead; it measures how far a word is inclined towards positive and negative. To understand sentiment analysis this paper focuses on supervised machine learning approaches.

The rest of the paper is organized as follows. Second section discusses in brief about the work carried out for sentiment analysis in different domain by various researchers. Third section is about the approach we followed for sentiment analysis. Section four is about implementation details and results followed by conclusion and future work discussion in the last section.

2. RELATED WORK

Many researchers have worked in the field of sentiment analysis, each one proposing new way of getting better efficiency from machine learning approaches. A LSA to identify product feature opinion words which are required to choose correct sentences to become a summarization of review, with allowing only selected features to show the end results, thereby, reducing actual size of summary [1]. In [2] author talks about the specific problems within sentiment analysis field which includes; document level, sentence level, feature level, comparative opinion and sentiment lexicon problem. Bo pang [4] considers classifying documents not by topic, but by overall sentiment, concluding whether a review is positive or negative. Reviews are converted to simple decision by making use of approaches such as naïve bayes, support vector machine by initially counting the number of positive and negative words in a document.

Since opinions are not always direct e.g. “the nokia phone is good” but also it can be a comparative opinion like “nokia phone has better battery life than samsung”. There exists three levels at which opinions are classified: sentence level, document level, and feature level [8]. At sentence level, subjective and objective opinions exist, at document level, a document is classified based on overall sentiment expressed by opinion holder. However, at feature level, attributes of products are taken into consideration, which provides classification in depth.

In paper [5], a holistic lexicon-based approach is proposed that allows the system to handle opinion words that are context dependent. It takes into account the counting of the number of positive and negative opinion words near the product feature in each sentence. If count of positive opinion words are more than that of negative opinion words, the final prediction on the feature is positive else negative.

Author in [6] makes use of higher n-gram model using three classifiers. The first one being language model which is a generative method that computes the probability of generation of a word sequence. The Passive-Aggressive algorithms are second which consists of a family of margin based online learning algorithms for binary classification. Third, to predict the polarity of a review. Apart from classifying reviews in two broader categories, there also exists a term polarity degree to measure the strength of opinions, as in is the opinion strongly positive, mildly positive, highly negative etc [8].

Author in [9] says product of sentiment value and occurring frequency gives measurement of sentiments. Psenti approach calculates the overall sentiment of stated opinionated text like customer reviews and scales them as a real score between -1 and +1, which can then be easily transformed positive/negative classification or into a scale of 1-5 stars. Creating candidate list using POS tagging with removal of stop word leads to aspect identification. The aspect having less than 5 comments on it, is removed from the candidate list.

In paper [10], the product review is translated into a Vector of Feature Intensities (VFI). A VFI is a vector of N+ 1 value, each one representing a different product feature and the other features. Snyder and Barzilay [11] addressed the problem of analyzing multiple related opinions in a text and presented an algorithm that jointly learns ranking models for individual aspects by modeling the dependencies between assigned ranks. Tabular comparison of various existing approaches is shown in table1.

3. PROPOSED MODELLING

The approach involves use of collection of product based dataset from different E-commerce sites like amazon.com, epinion.com etc. The reviews are collected on products like phone, ipod etc. The objective of the work is to analyze and predict product based reviews by classifying them as positive, negative and neutral by using algorithms like naïve baye’s and SVM. Since input is about product reviews that are unstructured, we perform pre-processing, extracts features on to which comments are made, then calculates polarity of reviews, and also plots graph for the result. The results also cover dealing with negation part. For example- “the nokia phone is not bad” gives positive review though it contains a negative word “not”. The flow diagram for approach is as given below and subsections are explained in details in next subsections.

3.1 Dataset

The dataset was collected from different product sites related to mobile domain product reviews like .cnet.com, downlaod.com, reviewcentre.com, zdnet.com, epinions.com and consumereview.com.

3.2 Pre-processing

The dataset is unstructured; it may contain repetitive words, large number of words that are not at all needed in summarizing of opinions. Pre-processing involves removal of stop words such as 'and', 'or', 'that' etc. followed by porter stemming which involves simplifying target words to base words by removal of suffixes such as – ed, ate, ion, ional, ment, ator, ssess, es, ance or conversion from ator to ate etc. For example, "replacement" is stemmed to replac; "troubled" to trouble ; "happy" to happi ; "operator" to operate. The raw data is pre-processed to improve quality.

3.3 Feature Extraction

Features in reviews are extracted so that it helps customer to know which feature has positive comment and which one has negative. Since, overall conclusion about product is much needed but there is also situation where customer requirements come into the scenario. Use of adjectives is done to classify opinions as positive or negative using unigram model. For example, "the Samsung camera I bought was good; it has got great touch screen, awesome flashlight." The feature extracted out of it would be like: Domain:
Mobile; Product: Samsung; Feature: Camera; Adjective: Good.

3.4 Training and classification

Supervised learning generates a function which maps inputs to desired outputs also called as labels because they are training examples labeled by human experts. We apply naïve bayes and support vector machine techniques to carry out supervised
learning on the dataset fetched.

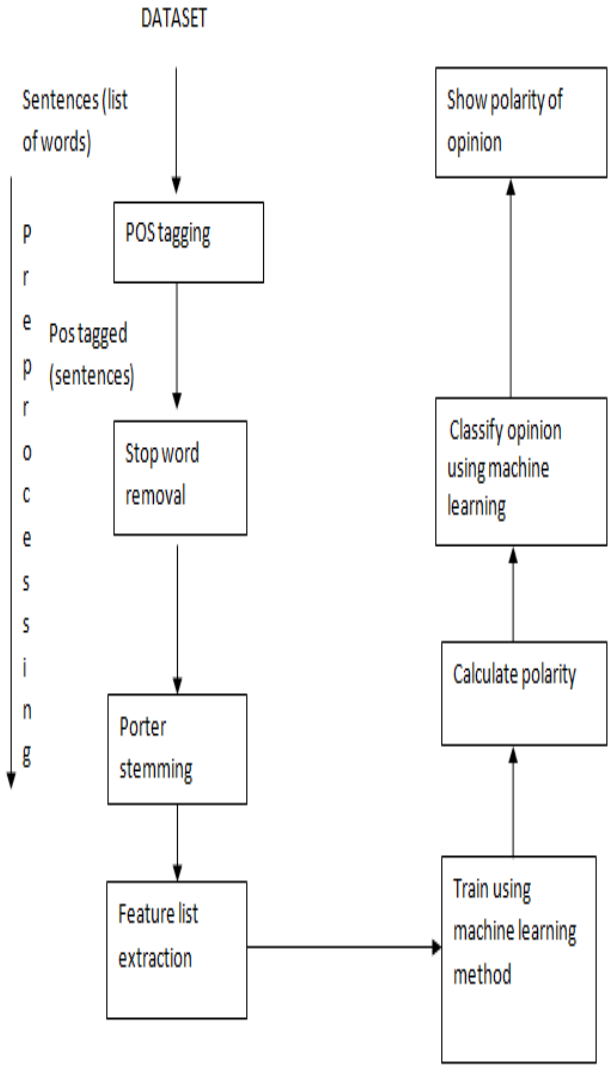

Fig. 1 Flowchart for approach

Table1 : Comparison against existing approaches

S.No	Studies/method	Feature selection	Data source	Performance (accuracy)	Precision	Recall	F1 measure	Disadvantages/ Advantages
1	Synder and Barzilay[3]	Multi aspect sentences	English restaurant review	7%	-	-	-	Segmentation into single multi aspect units.
2	Dotplotting and fragkou [8]	Linear text segmentation	Chinease restaurant review	31%	0.19	0.39	0.25	Works at document level rather than sentence level.
3	maxEnt based classifier[12]	Aspect identification	Dianping.com	86%	-	-	-	Makes use of labeled training data.
4	Hu and Liu[3]	Feature generalization-noun or noun phrases.	Amzaon.com and Cnet.com	84.2%	0.693	0.642	-	Cannot effectively deal with the implicit feature expression problem.
5	Alekh agarwal[3]	Machine learning using graph cut technique	WorldNet	90%	-	-	-	Generic method of using graph cut technique for efficient opinion classification.
6	Anidya[3]	Text mining techniques	Amzon.com	-	-	-	-	Information contained in product description are used in identifying reviews that have most impact.
7	Lie Zhang[12]	Noun product feature , Opinion lexicon	Product review	-	-	-	-	Noun product feature is directly modified into positive and negative opinion words.
8	Xiowen ding[3]	Lexicon based approach	Product review	-	-	-	-	Opinion words which are context dependent are easily used.
9	Yongyong zhail[8]	Opinion Feature Extraction based on Sentiment Patterns	Product review	-	-	-	-	Takes into account structure characteristic of reviews.
10	Gamgarn somprasestri [8]	Maximum entropy	Amazon reviews	-	0.726	0.787	0.754	-
11	Kaiqaunan xu[8]	Multiclass SVM	Amazon reviews	61%	0.619	0.934	0.742	-
12	Rui xia[8]	Naive bayes, maximum entropy, SVM	Movie review, multidomain dataset, amazon	NB-85.8% ME-85.5% SVM-86.4%	-	-	-	Two type of features-POS based and word-relation based are combined.

3.4.1 Naïve bayes:

Naive Bayes classifiers work on the principle that the value of a particular feature is independent of the value of any other feature. For example- Samsung phone will be considered as phone if it has basic call function, touch screen and camera. A naive Bayes classifier considers each of these features to contribute independently to the probability that this Samsung phone is a mobile, regardless of any possible correlations between the cameras, call function and touch screen. Assign to a given document d a, the class c * = arg maxc P(c | d)

$$Pnb(c|d) = \frac{P(c)(\prod P(f|c)^{n}i(d))}{P(d)}$$

Class c* is assigned to product review *d*, where, *f* represents a feature and ni(d) represents the count of feature fi found in product review *d*. There are a total of m features. Parameters P(c) and P (f|c) are obtained through maximum likelihood estimates which are incremented by one for smoothing.

Dataset after being preprocessed and after extracting features, is input to train through naïve bayes, hence providing polarity for reviews. For example: "the phone is great" provides positive opinion regarding the product.

3.4.2 Support Vector Machine:

Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. Every data represented as a vector is classified in a particular class.

Now the task is to find a margin between two classes that is far from any document. The distance defines the margin of the classifier, maximizing the margin reduces indecisive decisions. SVM also supports classification and regression which are useful for statistical learning theory and it helps recognizing the factors precisely, that needs to be taken into account, to understand it successfully.

3.4.3 Implementation:

We use Microsoft visual studio as platform to develop an interface where we could test and train datasets, extract features out of it, choose a classifier Naïve bayes or Support vector machine to work upon and thus predict polarity of opinion at the end. Overall we have used files extracted from E-commerce datasets for testing and training both. The dataset used has total size of 13094 product reviews out of which 12094 are used for training and 1000 are used for testing. For training flow is as follows as depicted in fig 2.

The pseudo code for process is as shown:

Input- Dataset of product reviews

Output- Classification of these reviews as positive and negative.

Step1: Preprocess the data
- Removal of special characters
- Removal of stop words
- Stemming the word

Step2: Get feature list
- If word in stop word list
- Removal word
- Return file
- Else append word to file

Step3: Extract feature list
- Match every word in pre processed list
- If word matches adjective in base list
- Display word
- If word matches feature in base list
- Display feature

Step4: combine both feature and preprocessed list

Step5: Use machine learning algorithms
- Compute probability

Step6: classify opinion as positive, negative or neutral.

4. RESULTS AND DISCUSSIONS

We applied algorithm on closed domain of mobile product review. Computation of results from the dataset created to be used as test cases was opted from cnet.com website, amazon.com. The precision, recall and accuracy are calculated as follows:

Accuracy is given by:

$$Accuracy = Tp + Tn / Tp + Tn + Fp + Fn$$

Recall positive and recall negative ratio are computed as follows:

$$Recall\ (positive) = Tp / Tp + Fn$$

$$Recall\ (negative) = Tn / Fp + Tn$$

Precision positive and recall negative ratio are computed as follows:

$$Precision(positive) = Tp / Tp + Fp$$

$$Precision(negative) = Tn / Fn + Tn$$

Table 2 Matrix for Naïve baye:

True Negative(Tn) = 5069	False Negative(Fn) = 978
False Positive(Fp) = 955	True Positive(Tp) = 5092

Table 3 Matrix for Support vector machine

True Negative(Tn) = 4937	False Negative (Fn) = 1110

False Positive (Fp) = 1283	True Positive (Tp) = 4764

Table 4, Table 5 and Table 6 show the performance measures of naive bayes and support vector machine based classifiers respectively in terms of precision, recall and accuracy.

TABLE 4 Naïve Bayes measurements

Positive recall	83.89
Negative recall	84.15
Positive precision	84.21
Negative precision	83.83

TABLE 5 Support vector machine measurements

Positive recall	81.10
Negative recall	79.37
Positive precision	78.78
Negative precision	81.64

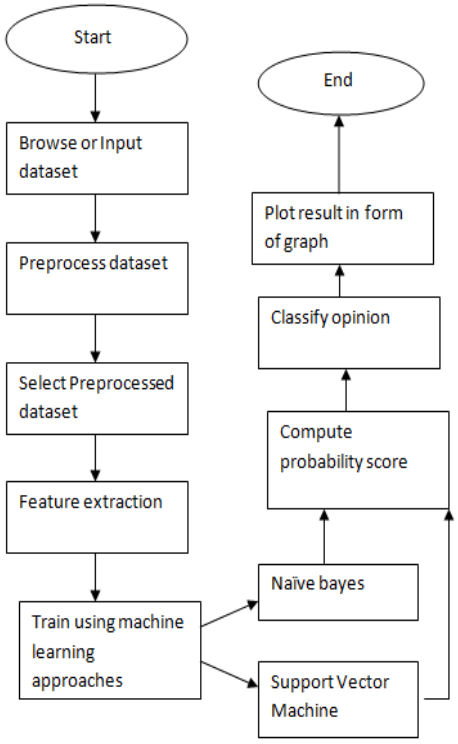

Fig. 2 Proposed methodology flow chart

TABLE 6 Performance measurements in terms of accuracy

Methods	Accuracy
Naïve bayes	84.02
Support vector machine	80.2

The graphs are as follows:

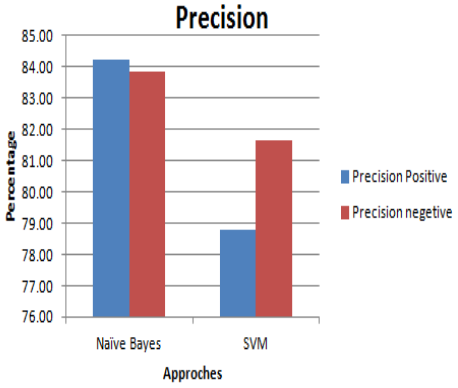

Fig. 3 Precision for Naïve Bayes and SVM

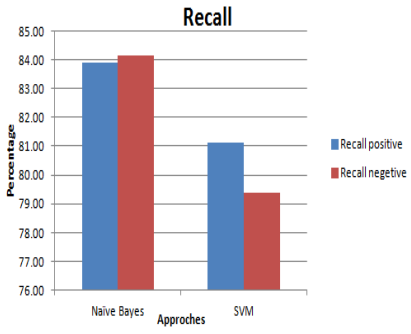

Fig. 4 Recall for Naïve Bayes and SVM

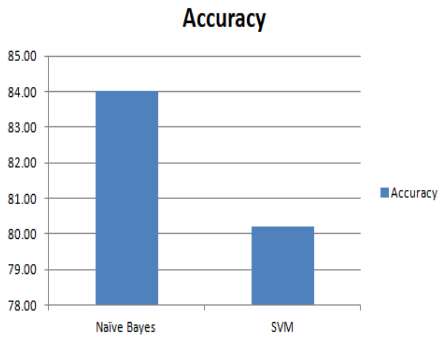

Fig. 5 Performance measurement in terms of accuracy

5. CONCLUSION

Sentiment analysis deals with identifying and aggregating the sentiment or opinions expressed by the users. Sentiment analysis is to classify the polarity of text in document or sentence whether the opinion expressed is positive, negative, or neutral. We see here that two approaches have been compared and a result for both approaches respectively on the product review dataset has been done. Naïve Bayes is found to give better accuracy that is 84.02% as compared to SVM approach which is 80.2%. We see that for text files that are too large in size take much more computation time. Automatic sentimental analysis is very useful to identify and predict current and future trends.

Till now opinion at feature level has been taken up but many limitations still exist which can be further taken up.

The future scope of improvement -

- Reviewing product based opinions in multiple languages.
- Dealing with problem of mapping slangs.
- Dealing with sarcastic opinions.
- Identifying comparative opinions and finding which among two product compared is best one.
- Dealing with anaphora resolution like what is actually being referred to in the opinion.

REFERENCES

[1] Liu, Chien-Liang, et al. "Movie rating and review summarization in mobile environment." Systems, Man, and Cybernetics, Part C: IEEE Transactions on Applications and Reviews,Volume 42 issue 3, pp.397-407,2012.

[2] Feldman, Ronen. "Techniques and applications for sentiment analysis."Communications of the ACM 5,Volume 56 Issue4, pp.82-89,2013.

[3] Hearst, Marti A., Susan T. Dumais, Edgar Osman, John Platt, and Bernhard Scholkopf. "Support vector machines." Intelligent Systems and their Applications, IEEE 13,Volume. 4 ,pp.18-28, 1998.

[4] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." In Proceedings of the ACL-02 conference on Empirical methods in natural language processing- Association for Computational Linguistics, Volume 10, pp. 79-86,2002.

[5] Ding, Xiaowen, Bing Liu, and Philip S. Yu. "A holistic lexicon-based approach to opinion mining." Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, pp. 231-240, 2008.

[6] Cui, Hang, Vibhu Mittal, and Mayur Datar. "Comparative experiments on sentiment classification for online product reviews." Proceedings of the 21st national conference on Artificial intelligence (AAAI). Volume 2, pp 1265-1270, 2006.

[7] Jebaseeli, A. Nisha, and E. Kirubakaran. "A Survey on Sentiment Analysis of (Product) Reviews." International Journal of Computer Applications 47, Volume 11, pp36-39, 2012.

[8] Wang, Min, and Hanxiao Shi. "Research on sentiment analysis technology and polarity computation of sentiment words." Progress in Informatics and Computing (PIC), 2010 IEEE International Conference on. Vol. 1. IEEE, 2010

[9] Mudinas, Andrius, Dell Zhang, and Mark Levene. "Combining lexicon and learning based approaches for concept-level sentiment analysis." Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining. ACM, pp330-333, 2012.

[10] de Albornoz, Jorge Carrillo, et al. "A joint model of feature mining and sentiment analysis for product review rating." Advances in information retrieval. Springer Berlin Heidelberg, pp.55-66, 2011.

[11] Mattosinho, F. J. A. P. "Thesis on Mining Product Opinions and Reviews on the Web."Technische Universitat Dresden ,2010.

[12] Pang, Bo, and Lillian Lee. "Thesis on Opinion mining and sentiment analysis."Foundations and trends in information retrieval Volume1-2,pp 1-135,2008

[13] Zhu, Jingbo, et al. "Aspect-based opinion polling from customer reviews." IEEE Transactions on Affective Computing, Volume 2.1,pp.37-49, 2011.

[14] Na, Jin-Cheon, Haiyang Sui, Christopher Khoo, Syin Chan, and Yunyun Zhou. "Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews." Advances in Knowledge Organization Volume9, pp. 49-54, 2004.

[15] Nasukawa, Tetsuya, and Jeonghee Yi. "Sentiment analysis: Capturing favorability using natural language processing." In Proceedings of the 2nd international conference on Knowledge capture, ACM, pp. 70-77, 2003.

[16] Li, Shoushan, Zhongqing Wang, Sophia Yat Mei Lee, and Chu-Ren Huang. "Sentiment Classification with Polarity Shifting Detection." In Asian Language Processing (IALP), 2013 International Conference on, pp. 129-132. IEEE, 2013.

[17] M.Karamibekr,A.A.Ghorbani,"Verb Oriented Sentiment Classification," Processed of the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT), Vol (1): pp. 327-331, 2012.

[18] Melville, Prem, Wojciech Gryc, and Richard D. Lawrence. "Sentiment analysis of blogs by combining lexical knowledge with text classification." InProceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1275-1284. ACM, 2009.

[19] Abbasi, Ahmed, Hsinchun Chen, and Arab Salem. "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums."ACM Transactions on Information Systems (TOIS) Volume 26 issue 3, pp.1- 12, 2008.