ISSN: 0976-3104
SPECIAL ISSUE: Emerging Technologies in Networking and Security (ETNS)

THE IIOAB JOURNAL

ARTICLE OPEN ACCESS

AN INTEGRATED APPROACH FOR SUPERVISED LEARNING OF ONLINE USER REVIEWS USING OPINION MINING

Shobana[1] and Anny Leema[2*]

[1]*Dept. of Computer Sc. and Engineering, B.S.Abdur Rahman University,vandalur, Chennai-600100, Tamilnadu, INDIA*
[2]*Department of MCA, B.S.Abdur Rahman University,vandalur, Chennai-600100, Tamilnadu, INDIA*

ABSTRACT

Abstract - Internet is a place where people stores and shares information among the other entities reside in the network. In recent years, the amazing development of web technologies, lead to huge quantity of system and user generated information in online systems. This huge amount of information on web platforms enables them to use as data sources, in review making applications based on opinion mining and sentiment analysis. The paper put forward an algorithm for detecting sentiments on user reviews on movie, based on Naive Bayes classifier. We considered the opinion mining domain for the analysis of user reviews and the techniques used in sentimental analysis. We implemented the proposed algorithm to the user reviews and tested its performance, and suggested directions of development.

Published on: 08th– August-2016

KEY WORDS

Opinion Mining; Web Content; Mining; Sentiment Analysis; Naïve Bayes

***Corresponding author: Email:** annyleema@gmail.com; **Tel:** +91 99 44238480

INTRODUCTION

The development of web technology and its information sharing among the web evolved as exponential increase in amount of information in online system. The data stored in online consists of various levels and types. Sometimes data analysts and researchers also feel tricky to vary the data that are residing on web. Because of huge volume of information it becomes difficult to process the information by individuals which leads to information overload and affects decision making processes in organizations.

Hence it is important to incorporate a new data analyzing and processing techniques for creation of knowledge. Knowledge discovery and feature extraction from data warehouse is a primary task of organizations in order to develop their products and improve their business strategies. Most of the data stored in online is in the form of text. Storing and transferring of text based data is more convenient and also easier to read by the people. In this context, applying data mining techniques is more reliable to handle data.

Difference between web mining and data mining are important in terms of data collection. In data mining, it is assumed that the data is already collected and stored in databases. In case of web mining, it uses special mechanisms, such as information retrieval - *IR (Information Retrieval)* and information extraction - *IE (Information Extraction)*, to obtain data and to pre-process them to apply data mining techniques. The web mining mechanism is divided into three categories based on its applications

Web structure mining - Discovering knowledge from hyperlinks to maximize relative information about the relations between web pages.

Web usage mining - Extracting patterns of users and models, from web logs, which is the repository of data access and activities of each visitor to a website.

Web content mining - Extracting knowledge and information from web page content.

ISSN: 0976-3104

Text mining technique is mainly applied to discover knowledge from the unstructured text. The outcomes of the text mining on unstructured text are used in research areas like Artificial Intelligence, Natural Language Processing (NLP) and Machine learning. In addition this technique is also applied for content spam detection, document classification and trend analysis.

Users often using web sites will always wishes to post their opinion, ratings and reviews of products. Most of this information is in unstructured format. Hence it is important to impart the knowledge discovering techniques for detection and extraction of opinions or sentiments from textual information.

Analyzing of customer sentiment and their opinion on a newly launched product, based on feedback from web pages is vital for evaluation of impact and making decision on business development. *Opinion mining* is a kind of knowledge discovery deals with habitual methods of detection ,extraction of opinions and classifying the sentiments presented in a text which are given by the users. Application of opinion mining to the raw text data result in creation of effective referral systems, trend analysis, financial analysis, strategic management, market research and product development.

TECHNIQUES OF OPINIONS MINING

Opinion mining and its uses created a dramatic change in e-commerce. The opinion and review made by the visitors and guests on products enables the organizations to improve their marketing strategy. It has increased big e-commerce sites and recommendations of products and services sites. The large number of reviews on a product promotes easy access to useful and reasonable information to visitors. It can be used to compare offers from different competitors on the market of similar product and make an informed decision about buying a certain offer. It is very difficult for a visitor to read all of lengthy reviews and to form an opinion on a product because:

In some cases the reviews already posted by other existing customers can be very long and only a few sentences may express opinions. Read through of only part of the review may create a false impression about the topic or may give inaccurate result;

The user is not aware about the various metrics used in comparing offers in a certain specialized field. Also, the large number of lengthy reviews makes it difficult for producers and analysis to follow the real expectations of customers. In addition with the lengthy reviews, they face difficulties in follow wide range of products, traded on a variety of web sites. So, it is useful to make a system to detect indicators of performance of a product, and domain specific metrics, to recapitulate the opinions obtained from the large amount of reviews, in several positive and negative aspects.

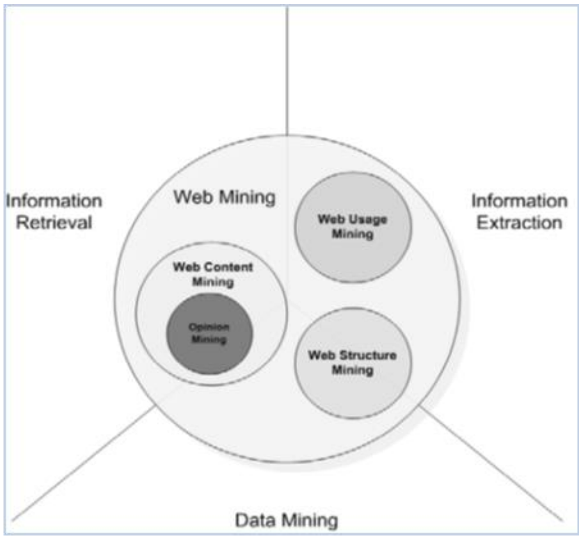

Fig: 1. Data mining

………..

Classifying entire reviews posted by the visitors according to the opinions towards certain product or on object is a sentiment classification. To produce a feature summary, product features are first identified, and positive and negative opinions on them are aggregated. Features are such as components, product attributes like price, availability, warranty, manufacturer, etc,. The effective summarizing of opinions and, grouping of feature expressions is critical. It is very time consuming and tedious for business analysts and organizations to group typically hundreds of feature expressions and reviews that can be identified from the text for an opinion mining application into feature categories. Opinion summarization does not summarize the reviews by selecting a subset or rewrite some of the original sentences from the reviews to capture the main points as the classic text summarization. [2]

For some analyzed entity, we determine the words or phrase that state user opinion. The process is comprised of three stages [10]:

Entity determination – identification of texts i.e., user reviews that contain information about the entity/attributes;

Determination of sentiments – for the text of the previous stage is considered the content of opinions and sentiments, by searching a set of words carrying sentiments, or by prior training a classifier;

Determination of entity - sentiment relationship - at this stage opinions are analyzed and addressed to the entity under review. Usually this is done through a predefined list of patterns.

SENTIMENT CLASSIFICATION

Polarity classification or Sentiment classification is the binary classification task of expressing either an overall positive or an overall negative opinion by labeling an opinionated document. A typical approach for sentiment classification is to use machine learning algorithms.

Machine Learning

A system which is capable of automatically acquiring and integrating the knowledge is referred as machine learning. The systems that learn from analytical observation, experience, training, and other means, results in a system with self-improvement, effectiveness and efficiency. Machine learning systems uses knowledge and a corresponding knowledge organization to test the knowledge acquired, interpret and analyze. The machine learning system can be classified into three methods based on the labeling the data. They are supervised learning and unsupervised learning.

Supervised learning: Supervised learning produces a function which maps inputs values to desired outputs is called as labels. These labels are assigned by the human experts. Since it is a text classification problem, any supervised learning method can be applied, e.g., Naive Bayes classification, and support vector machines (SVM).

Unsupervised learning*:* Unsupervised learning models a set of inputs, like clustering, and labels are not known during training. Classification of data is performed using some fixed syntactic patterns which are used to express opinions. These mechanisms also have to be done by the human experts. The part-of-speech (POS) tags are used to compose syntactic patterns.
Semi-supervised learning: Semi-supervised learning generates an appropriate function or classifier in which both labeled and unlabelled examples are combined. [2, 5]

Sentiment Analysis Tasks

Sentiment analysis includes a task of classifying the polarity of a given text at the document, sentence or feature level expressing of opinion as positive, negative or neutral. The sentiment analysis can be performed at one of the three levels: the document level, sentence level, feature level.

Document Level Sentiment Classification: In document level sentiment analysis, the primary task is to extract informative text for deducing sentiment of the whole document. Because the objective statements are rendered by subjective statements and complicate further for document categorization task with conflicting sentiment, the learning method creates uncertainty. [6]

Sentence Level Sentiment Classification: The sentence level sentiment classification is a fine-grained level than document level sentiment classification. In sentence level sentiment classification polarity of the sentence can be given by three categories as positive, negative and neutral. The challenge faced by sentence level sentiment classification is the identification features indicating whether sentences are on-topic which is kind of co-reference problem [6]

Feature Level Sentiment Classification*:* Product features are defined as product attributes or components. Analysis of such features for identifying sentiment of the document is called as feature based sentiment analysis. In this approach positive or negative opinion is identified from the already extracted features. It is a fine grained analysis model among all other models [2]

PROPOSED OPINION MINING METHOD

In this case for training, the user comments are collected and extracted from at http://www.cs.cornell.edu/people/pabo/ movie-review-data/ [12]. This collection contains 5331 sentences already classified as positive and 5331 negative opinions from 2000 comments processed and classified in two categories. Comments usually contain several sentences, but opinion will be determined at sentence level, then later determining overall comment opinion. Obtained collection consists of two files, one for each set of positive and negative opinions, containing one sentence per line, making it easy to process. To extract opinions we will use a Naive Bayesian classifier. This type of classifier has the advantage that it is easy to implement, quickly and generate good results.

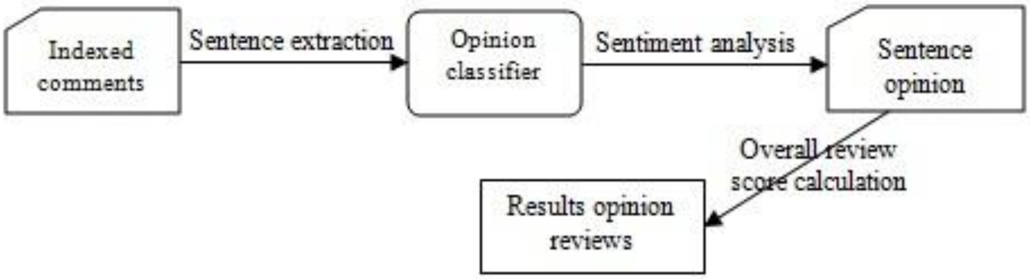

Fig 2. The overall Opinion mining process

Using Naive Bayes algorithm

The Naive Bayes classifier is a probability classifier which is based on Bayes' theorem. It shows the relation between probability of two events A and B, P (A) and P (B) and conditional probability of event A conditioned by B and event B conditioned by A, P (A | B) and P (B | A). It is given as [13]:

$$P\ (A\ /\ B)=\frac{P\ (B/A)\ P(A)}{P(B)}$$

Using Bayes' theorem, we can estimate the probability of an event based on the examples of its occurrence. In this case, we estimate probability that a document is positive or negative, in a certain context, or the likelihood that an event to take place if it was predetermined to be positive or negative. This is facilitated by the collection of positive and negative examples chosen. The process is naive Bayesian because of how we calculate the probability of occurrence of an event - is the product of probability of occurrence of each word in the document. This presumes that there is no connection between the words. This assumption of independence is introduced to facilitate the construction of classifier, it is not entirely true, and there are words that appear together more frequently than individual.

We estimate the probability of a word with positive or negative meaning by analyzing a series of positive and negative examples and calculating the frequency of each of the classes. This learning process is supervised, requiring the existence of pre-classification examples for training. Starting from:

$$P(sentiment/sentence) = \frac{P(sentiment)P(sentence|sentiment)}{P(sentence)}$$

we assume that *P(sentence|sentiment)* is the product of *P(word|sentiment)* for all words in a sentence. We estimate P(word|sentiment) as:

$$P(word|sentiment) = \frac{no.of\ word\ occurance\ in\ class + 1}{no.of\ words\ belonging\ to\ a\ class + total\ no.of\ words}$$

The steps in the classification method proposed in the paper are presented in **Figure- 3**, below

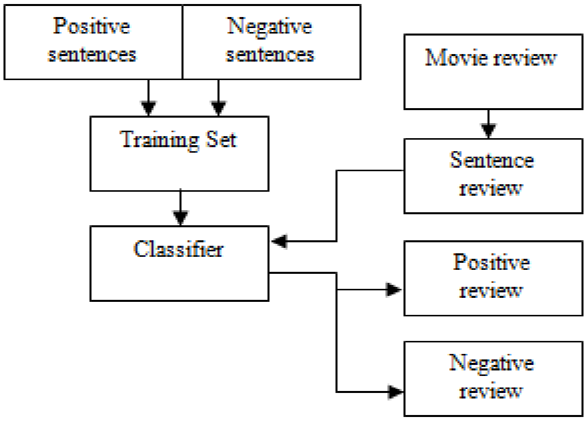

Fig: 3. Stages of classification process

………..

The proposed algorithm has following steps:

```
Initialize P (pos) <- nr_P (pos) / nr_total_probability
Initialize P (neg) <- nr_P (neg) / nr_total_probability
Tokenize sentence in words
For each class of {pos, neg}:
For each word in {phrase}
        P (word | class) <- nr_apartii (word | class) 1 / nr_cuv (class) + nr_total_cuvinte
        P (class) <-P (class) * P (word | class) Returns max {P(pos), P(neg)}
```

EVALUATE THE PERFORMANCE OF ALGORITHM

We use two specific measures for information retrieval systems to evaluate the results of algorithm called precision and the recall. The relation between the precision and recall with respect to positive and negative is given as:

Table 1. Contingency table of correctly classified reviews

	Relevant	Irrelevant
Detected opinions	True Positive(TP)	False Positive(FP)
Undetected opinions	False Negative(FN)	True Negative(TN)

ISSN: 0976-3104

Precision: Precision is the ratio of the correctly classified extracted opinions and all extracted opinions, the percentage of correctly classified opinions from classified ones:

$$\text{Precision}=\frac{TP}{TP+FP}$$

Recall: Recall expresses the ratio of correctly classified extracted opinions and classified opinions in data source, the percent of correctly classified opinions from all opinions in a class:

$$\text{Recall}=\frac{TP}{TP+FN}$$

Another evaluation measure for algorithm may be accuracy, expressing the percentage of correct made classifications, and F-measure, a weighted harmonic mean of precision and recall:

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN},\quad F=\frac{2*\text{precizion}*\text{recall}}{\text{Precision}+\text{Recall}}$$

We calculate accuracy of classifier, the recall and precision for the two classes, training the algorithm on 5000 sentences for each class of pre-classification test examples and applying it on the rest of the remaining examples.

Analyzing the algorithm efficiency for the above parameters we achieved a 0.814332247557 value of correct classification of opinions. From the result we conclude that a solution to improve the quality of the algorithm is to eliminate insignificant words for classification. Algorithm originally classified words without lexical content, so that besides nouns, verbs, adverbs and adjectives, are considered articles, prepositions and pronouns without semantic value.

We will eliminate these words called stop words in English that can induce noise in the classification. For this we have built an array of four vectors corresponding to those prepositions, conjunctions, articles and pronouns (e.g. articles - the, the, year, conjunctions - and, now, so, still, only, pronouns - who, whom, which, that, this, me, you, ours, prepositions - about, above, across, after, at, around, with, up). After this step we observe that the algorithm efficiency has improved a little, giving a value of: 0.81699346405229.

Algorithm considers that there is no relationship between words in a sentence, but in reality they are interrelated. So there are certain words that occur frequently together in one of two classes. Usually, we observe that the algorithm efficiency increases for values up to a maximum of three groups of words. In this application we tested the introduction in classification of groups of words for the n = 2 and n = 3. Results are presented in the table below:

Table 2. Algorithm efficiency

	Initial Algorithm, groups of n=1 words	Initial Algorithm, groups of n=1 words, eliminate stop words	Algorithm for groups of n=2 words, eliminate stop words	Algorithm for groups of n=3 words, eliminate stop words
Precision	0.8143	0.8169	0.8208	0.7972
Recall	0.7598	0.7598	0.7659	0.5258
Acurracy	0.7933	0.7948	0.7993	0.6960
F-measure	0.7861	0.7874	0.7924	0.6336
Execution time (s)	1.1535	3.2490	8.0684	14.7787

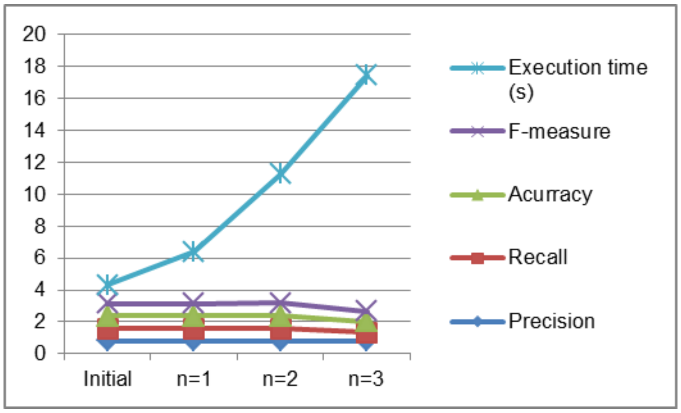

Fig. 4. Algorithm efficiency

………..

CONCLUSION

The expression of opinions of users in specialized sites for evaluation of products and services, and also on social networking platforms, has become one of the main ways of communication, due to spectacular development of web environment in recent years. The large amount of information on these platforms make them viable for use as data sources, in applications based on opinion mining and sentiment analysis. This paper presents a method of sentiment analysis, on the review made by users to movies. Classification of reviews in both positive and negative classes is done based on a naive Bayes algorithm. As training data we used a collection (pre-classified in positive and negative) of sentences taken from the movie reviews. To improve classification we removed insignificant words and introduced in classification groups of words (n-grams). For n = 2 groups we achieved a substantial improvement in classification. As an extension of the research presented in this paper we want to improve the algorithm, enriching the training set of examples, on the way, with examples classified as strong positive or negative, by an established score of classification. We try to determine, in a review, those sentences which do not express opinions, or determine opinions about the film or the film acto addressed strictly on these items. We try to highlight the main aspects on which opinions are expressed and to extract opinions based on aspects identification. rs and identify opinions

The following chart presents the influence of data training volume on the accuracy of classifications in the method used. We detect a critical number of training data, from which point the increase number of the initial training set, will produce very little influence on precision of the algorithm.

CONFLICT OF INTEREST

The authors declare no conflict of interests.

ACKNOWLEDGEMENT

None

FINANCIAL DISCLOSURE

The authors report no financial interests or potential conflicts of interest.

REFERENCES

[1] M Caraciolo (2012, Mar.) Working on sentiment analysis on Twitter with Portuguese language. [Online].http://aimotion.blogspot.com/2010/07/working-on-sentiment-analysis-on.html

[2] Smeureanu A, Diosteanu C, Delcea LA Cotfas.[2011] Busines Ontology for Evaluating Corporate Social Responsibility, *Amfiteatru Economic*, 29: 28-42

[3] L Minqing Hu.[2004] Mining and Summarizing Customer Reviews," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2004, full paper), Seattle, USA,.

[4] Turney P.[2002] Thumbs Up or Thumbs Down? semantic orientation applied to unsupervised classification of reviews. ACL'.

[5] Liu, Web Data Mining - Exploring Hyperlinks, Contents and Usage Data, Secound ed.: Springer, 2011.

[6] Freitag D and McCallum A.[2000] Information extraction with HMM structures learned by stochastic optimization. AAAI-00,

[7] Hatzivassiloglou V and Wiebe J. Effects of adjective orientation and gradability on sentence subjectivity. COLING'00, 2000.

[8] Hearst M.[Direction-based Text Interpretation as an Information Access Refinement. In P Jacobs, editor, Text-Based Intelligent Systems. Lawrence Erlbaum Associates, 1992.

[9] PD Turney and M.L. Littman. Unsupervised learning of semantic orientation from a hundred-billion-word corpus.

[10] Technical Report ERB-1094, National Research Council Canada, Institute for Information Technology, 2002.

[11] Janyce Wiebe.[2000] Learning subjective adjectives from corpora. In AAAI/IAAI, pages 735–740.

[12] M Caraciolo. (2012, Mar.) Working on sentiment analysis on Twitter with Portuguese language. [Online].http://aimotion.blogspot.com/2010/07/working-on-sentiment-analysis-on.html

[13] MF Porter.[1980] An algorithm for suffix stripping. In Program, 14,: 130–137,

[14] Deepak Ravichandran and Eduard Hovy. Learning surface text patterns for a question answering system. In ACL Conference, 2002.

[15] R.Jaya and Rajan.C , [2016]A Study On Data Mining Techniques, Methods, Tools And Applications In Various Industries, International Journal on Concurrent Applied Research in Engineering and Management,vol.1, No.4, pp. 14-24.

[16] Ellen Riloff. Automatically generating extraction patterns from untagged text. In Proceedings of AAAI/IAAI, Vol. 2, pages 1044–1049, 1996.

[17] Nigam, K. and Hurst, M. 2004. Towards a robust metric of opinion. AAAI Spring Symp.on Exploring Attitude and Affect in Text.NLProcessor,2000.

[18] Pang B, Lee L, and Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. EMNLP-02.