METHODOLOGICAL STUDY OF OPINION MINING AND SENTIMENT ANALYSIS TECHNIQUES

Pravesh Kumar Singh[1], Mohd Shahid Husain[2]

[1]M.Tech, Department of Computer Science and Engineering, Integral University, Lucknow, India
[2]Assistant Professor, Department of Computer Science and Engineering, Integral University, Lucknow, India

ABSTRACT

Decision making both on individual and organizational level is always accompanied by the search of other's opinion on the same. With tremendous establishment of opinion rich resources like, reviews, forum discussions, blogs, micro-blogs, Twitter etc provide a rich anthology of sentiments. This user generated content can serve as a benefaction to market if the semantic orientations are deliberated. Opinion mining and sentiment analysis are the formalization for studying and construing opinions and sentiments. The digital ecosystem has itself paved way for use of huge volume of opinionated data recorded. This paper is an attempt to review and evaluate the various techniques used for opinion and sentiment analysis.

KEYWORDS

Opinion Mining, Sentiment Analysis, Feature Extraction Techniques, Naïve Bayes Classifiers, Clustering, Support Vector Machines

1. INTRODUCTION

Generally individuals and companies are always interested in other's opinion like if someone wants to purchase a new product, then firstly, he/she tries to know the reviews i.e., what other people think about the product and based on those reviews, he/she takes the decision.

Similarly, companies also excavate deep for consumer reviews. Digital ecosystem has a plethora for same in the form of blogs, reviews etc.

A very basic step of opinion mining and sentiment analysis is feature extraction. Figure 1 shows the process of opinion mining and sentiment analysis

.

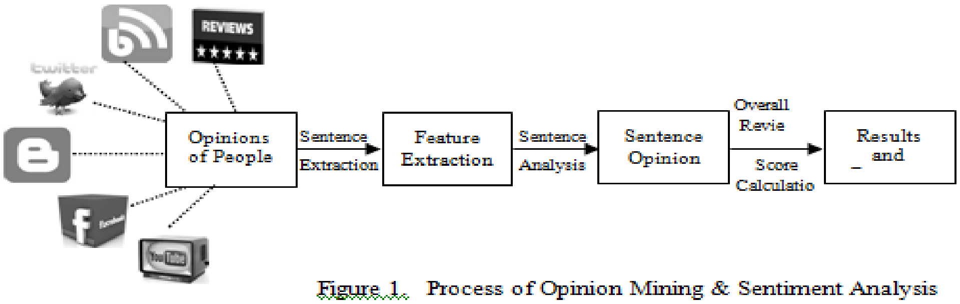

Figure 1. Process of Opinion Mining & Sentiment Analysis

DOI: 10.5121/ijsc.2014.5102

There are various methods used for opinion mining and sentiment analysis among which following are the important ones:

1) Naïve Bays Classifier.
2) Support Vector Machine (SVM).
3) Multilayer Perceptron.
4) Clustering.

In this paper, categorization of work done for feature extraction and classification in opinion mining and sentiment analysis is done. In addition to this, performance analysis, advantages and disadvantages of different techniques are appraised.

2. DATA SETS

This section provides brief details of datasets used in experiments.

2.1. Product Review Dataset

Blitzer takes the review of products from amazon.com which belong to a total of 25 categories like videos, toys etc. He randomly selected 4000 +ve and 4000 –ve reviews.

2.2. Movie Review Dataset

The movie review dataset is taken from the Pang and Lee (2004) works. It contains movie review with feature of 1000 +ve and 1000 –ve processed movie reviews.

3. CLASSIFICATION TECHNIQUES

3.1. Naïve Bayes Classifier

It's a probabilistic and supervised classifier given by Thomas Bayes. According to this theorem, if there are two events say, e_1 and e_2 then the conditional probability of occurrence of event e_1 when e_2 has already occurred is given by the following mathematical formula:

$$P(e_1 \mid e_2) = \frac{P(e_2 \mid e_1)P(e_1)}{e_2}$$

This algorithm is implemented to calculate the probability of a data to be positive or negative. So, conditional probability of a sentiment is given as:

$$\text{P(Sentiment} \mid \text{Sentence)} = \frac{\text{P(Sentiment)P(Sentence} \mid \text{Sentiment)}}{\text{P(Sentence)}}$$

And conditional probability of a word is given as:

$$\text{P(Word} \mid \text{Sentiment)} = \frac{\text{Number of word occurence in class} + 1}{\text{Number of words belonging to a class} + \text{Total nos of Word}}$$

Algorithm

S1: Initialize P(positive) ← num – popozitii (positive)/ num_total_propozitii

```
S2: Initialize P(negative) ← num – popozitii (negative) / num_total_propozitii

S3: Convert sentences into words
        for each class of {positive, negative}:

                for each word in {phrase}

                        P(word | class) < num_apartii (word | class) 1 | num_cuv (class) + num_total_cuvinte

        P (class) ←P (class) * P (word | class)

Returns max {P(pos), P(neg)}
```

The above algorithm can be represented using figure 2

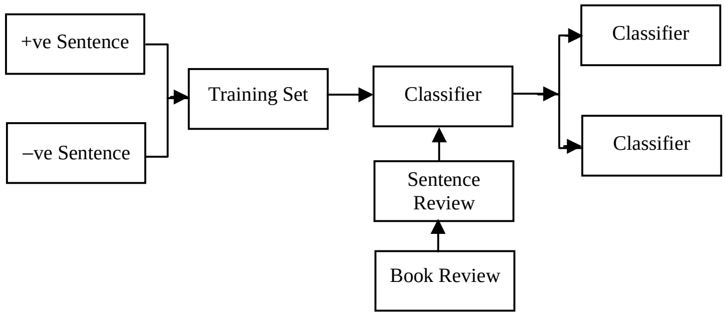

Figure 2. Algorithm of Naïve Bayes

3.1.1. Evaluation of Algorithm

To evaluate the algorithm following measures are used:

- Accuracy
- Precision
- Recall
- Relevance

Following contingency table is used to calculate the various measures.

	Relevant	Irrelevant
Detected Opinions	True Positive (tp)	False Positive (fp)
Undetected Opinions	False Negative (fn)	True Negative (tn)

Now, Precision = $\frac{tp}{tp + fp}$

Accuracy = $\frac{tp + tn}{tp + tn + fp + fn}$, F = $\frac{2 * \mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$; Recall = $\frac{tp}{tp + fn}$

3.1.2. Accuracy

On the 5000 sentences [1] Ion SMEUREANU, Cristian BUCUR train the Naïve Gauss Algorithm and got 0.79939209726444 accuracy; Where number of groups (n) is 2.

3.1.3. Advantages of Naïve Bayes Classification Method

1. Model is easy to interpret.
2. Efficient computation.

3.1.4. Disadvantage of Naïve Bayes Classification Method

Assumptions of attributes being independent, which may not be necessarily valid.

3.2 Support Vector Machine (SVM)

SVM is a supervised learning model. This model is associated with a learning algorithm that analyzes the data and identifies the pattern for classification.

The concept of SVM algorithm is based on **decision plane** that defines decision boundaries. A decision plane separates group of instances having different class memberships.

For example, consider an instance which belongs to either class Circle or Diamond. There is a separating line (figure 3) which defines a boundary. At the right side of boundary all instances are Circle and at the left side all instances are Diamond.

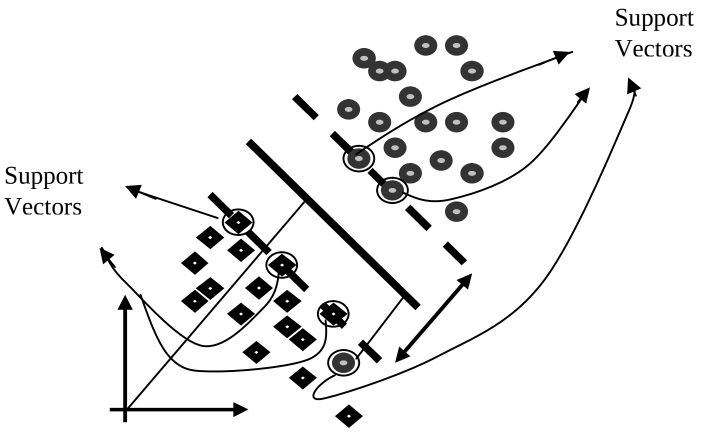

Figure 3. Principle of SVM

Is there is an exercise/training data set D, a set of n points is written as:

$$D=\left\{(x_i,c_i)\,\middle|\,x_i\,\varepsilon\,R^p,\ c_i\,\varepsilon\{\neg 1,1\}\right\}_{i-1}^{n} \qquad ....(1)$$

Where, x_i is a p-dimensional real vector. Find the maximum-margin hyper plane i.e. splits the points having $c_i = 1$ from those having $c_i = -1$. Any hyperplane can be written as the set of points satisfying:

$$w \bullet x - b = 1 \qquad ....(2)$$

Finding a maximum margin hyperplane, reduces to find the pair w and b, such that the distance between the hyperplanes is maximal while still separating the data. These hyperplanes are described by:

$$w \bullet x - \mathrm{b} = 1 \text{ and } \mathrm{w} \bullet \mathrm{x} - \mathrm{b} = -1$$

The distance between two hyperplanes is $\frac{\mathrm{b}}{\|\mathrm{w}\|}$ and therefore $\|\mathrm{w}\|$ needs to be minimized. The minimized $\|\mathrm{w}\|$ in w, b subject to $c_i\,(w.x_i - b) \geq 1$ for any i = 1… n.

Using Lagrange’s multipliers (α_i) this optimization problem can be expressed as:

$$\underset{\mathrm{w,b}}{\min} \quad \underset{\alpha}{\max} \left\{ \left\{ \{ \frac{1}{2}\|\mathrm{w}\|^2 - \sum_{i=1}^{n} \alpha_i [c_i (\mathrm{w.x}_i - \mathrm{b}) - 1] \right\} \right\} \; .....(3)$$

3.2.1. Extensions of SVM

There are some extensions which makes SVM more robust and adaptable to real world problem. These extensions include the following:

1. Soft Margin Classification

In text classification sometimes data are linearly divisible, for very high dimensional problems and for multi-dimensional problems data are also separable linearly. Generally (in maximum cases) the opinion mining solution is one that classifies most of the data and ignores outliers and noisy data. If a training set data say D cannot be separated clearly then the solution is to have fat decision classifiers and make some mistake.

Mathematically, a slack variable ξ_i are introduced that are not equal to zero which allows x_i to not meet the margin requirements with a cost i.e., proportional to ξ.

2. Non-linear Classification

Non-linear classifiers are given by the Bemhard Boser, Isabelle Guyon and Vapnik in 1992 using kernel to max margin hyperplanes.

Aizeman given a kernel trick i.e., every dot product is replaced by non-linear kernel function. When this case is apply then the effectiveness of SVM lies in the selection of kernel and soft margin parameters.

3. Multiclass SVM

Basically SVM relevant for two class tasks but for the multiclass problems there is multiclass SVM is available. In the multi class case labels are designed to objects which are drawn from a finite set of numerous elements. These binary classifiers might be built using two classifiers:

1. Distinguishing one versus all labels and
2. Among each pair of classes one versus one.

3.2.2. Accuracy

When pang take unigrams learning method then it gives the best output in a presence based frequency model run by SVM and he calculated 82.9% accuracy in the process.

3.2.3. Advantages of Support Vector Machine Method

1. Very good performance on experimental results.
2. Low dependency on data set dimensionality.

3.2.4. Disadvantages of Support Vector Machine Method

1. One disadvantages of SVM is i.e. in case of categorical or missing value it needs pre-processed.
2. Difficult interpretation of resulting model.

3.3. Multi-Layer Perceptron (MLP)

Multi-Layer perceptron is a feed forward neural network, with one or N layers among inputs and output. Feed forward means i.e, uni-direction flow of data such as from input layer to output layer. This ANN which multilayer perceptron begin with input layer where every node means a predicator variable. Input nodes or neurons are connected with every neuron in next layer (named as hidden layers). The hidden layer neurons are connected to other hidden layer neuron.
Output layer is made up as follows:

1. When prediction is binary output layer made up of one neuron and
2. When prediction is non-binary then output layer made up of N neuron.

This arrangement makes an efficient flow of information from input layer to output layer.
Figure 4 shows the structure of MLP. In figure 4 there is input layer and an output layer like single layer perceptron but there is also a hidden layer work in this algorithm.

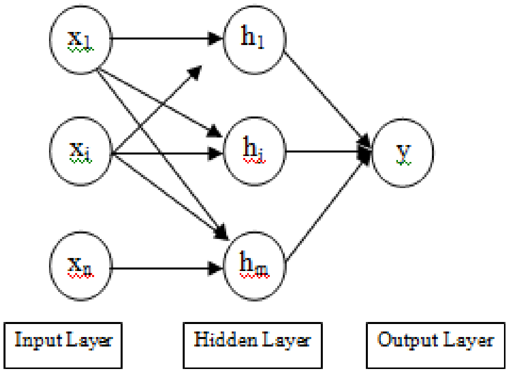

Figure 4. Structure of MLP

MLP is a back propagation algorithm and has two phases:

Phase I: It is the forward phase where activation are propagated from the input layer to output layer.

Phase II: In this phase to change the weight and bias value errors among practical & real values and the requested nominal value in the output layer is propagate in the backward direction.

MLP is popular technique due to the fact i.e. it can act as universal function approximator. MLP is a general, flexible and non-linear tool because a “back propagation” network has minimum one hidden layer with various non-linear entities that can learn every function or relationship between group of input and output variable (whether variables are discrete or continuous).

An advantage of MLP, compare to classical modeling method is that it does not enforce any sort of constraint with respect to the initial data neither does it generally start from specific assumptions.

Another benefit of the method lies in its capability to evaluation good models even despite the presence of noise in the analyzed information, as arises when there is an existence of omitted and outlier values in the spreading of the variables. Hence, it is a robust method when dealing with problems of noise in the given information.

3.3.1. Accuracy

On the health care data Ludmila I. Kuncheva, (IEEE Member) calculate accuracy of MLP as 84.25%-89.50%.

3.3.2. Advantages of MLP

1. It acts as a universal function approximator.
2. MLP can learn each and every relationship among input and output variables.

3.3.3. Disadvantages of MLP

1. MLP needs more time for execution compare to other technique because flexibility lies in the need to have enough training data.
2. It is considered as complex "black box".

3.4 Clustering Classifier

Clustering is an unsupervised learning method and has no labels on any point. Clustering technique recognizes the structure in data and group, based on how nearby they are to one another.

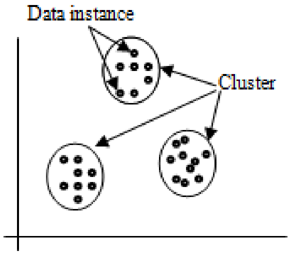

Figure 5. Clustering

So, clustering is process of organizing objects and instances in a class or group whose members are similar in some way and members of class or cluster is not similar to those are in the other cluster

This method is an unsupervised method, so one does not know that how many clusters or groups are existing in the data.

Using this method one can organize the data set into different clusters based on the similarities and distance among data points.

Clustering organization is denoted as a set of subsets C = C_1 . . . C_k of S, such that: $S = \bigcup_{i=1}^{k} C_i$ and $C_i \cap C_j = \phi$ for $i \neq j$. Therefore, any object in S related to exactly one and only one subset.

For example, consider figure 5 where data set has three normal clusters.

Now consider the some real-life examples for illustrating clustering:

Example 1: Consider the people having similar size together to make small and large shirts.
1. Tailor-made for each person: expensive
2. One-size-fits-all: does not fit all.

Example 2: In advertising, segment consumers according to their similarities: To do targeted advertising.

Example 3: To create a topic hierarchy, we can take a group of text and organize those texts according to their content matches.

Basically there are two types of measures used to estimate the relation: Distance measures and similarity measures.

Basically following are two kinds of measures used to guesstimate this relation:

1. Distance measures and
2. Similarity measures

Distance Measures

To get the similarity and difference between the group of objects distance measures uses the various clustering methods.

It is convenient to represent the distance between two instances let say x_i and x_j as: d (x_i, x_j). A valid distance measure should be symmetric and gains its minimum value (usually zero) in case of identical vectors.

If distance measure follows the following properties then it is known as metric distance measure:

1. Triangle inequality $d(x_i, x_k) \leq d(x_i, x_j) + d(x_j, x_k)$

$$x_i, x_j, x_k \quad S$$

2. $d(x_i, x_j) = 0 \quad x_i = x_j$

$$x_i, x_j \quad S$$

There are variations in distance measures depending upon the attribute in question.

3.4.1. Clustering Algorithms

A number of clustering algorithms are getting popular. The basic reason of a number of clustering methods is that “cluster” is not accurately defined (Estivill-Castro, 2000). As a result many clustering methods have been developed, using a different induction principle.

1. Exclusive Clustering

In this clustering algorithm, data are clusters in an exclusive way, so that a data fits to only one certain cluster. Example of exclusive clustering is K-means clustering.

2. Overlapping Clustering

This clustering algorithm uses fuzzy sets to grouped data, so each point may fit to two or more groups or cluster with various degree of membership.

3. Hierarchical Clustering

Hierarchical clustering has two variations: agglomerative and divisive clustering

Agglomerative clustering is based on the union among the two nearest groups. The start state is realized by setting every data as a group or cluster. After some iteration it gets the final clusters needed. It is a bottom-up version.

Divisive clustering begins from one group or cluster containing all data items. At every step, clusters are successively fragmented into smaller groups or clusters according to some difference. It is a top-down version.

4. Probabilistic Clustering

It is a mix of Gaussian, and uses totally a probabilistic approach.

3.4.2. Evaluation Criteria Measures for Clustering Technique

Basically, it is divided into two group's internal quality criteria and external quality criteria.

1. Internal Quality Criteria

Using similarity measure it measures the compactness if clusters. It generally takes into consideration intra-cluster homogeneity, the inter-cluster separability or a combination of these two. It doesn't use any exterior information beside the data itself.

2. External Quality Criteria

External quality criteria are important for observing the structure of the cluster match to some previously defined classification of the instance or objects.

3.4.3. Accuracy

Depending on the data accuracy of the clustering techniques varied from 65.33% to 99.57%.

3.4.4. Advantages of Clustering Method

The most important benefit of this technique is that it offers the classes or groups that fulfill (approximately) an optimality measure.

3.4.5. Disadvantages of Clustering Method

1. There is no learning set of labeled observations.
2. Number of groups is usually unknown.
3. Implicitly, users already choose the appropriate features and distance measure.

4. Conclusion

The important part of gathering information always seems as, what the people think. The rising accessibility of opinion rich resources such as online analysis websites and blogs means that, one can simply search and recognize the opinions of others. One can precise his/her ideas and opinions concerning goods and facilities. These views and thoughts are subjective figures which signify opinions, sentiments, emotional state or evaluation of someone.

In this paper, different methods for data (feature or text) extraction are presented. Every method has some benefits and limitations and one can use these methods according to the situation for feature and text extraction. Based on the survey we can find the accuracy of different methods in different data set using N-gram feature shown in table 1.

Table 1: Accuracy of Different Methods

	Movie Reviews			Product Reviews			
N-gram Feature	NB	MLP	SVM		NB	MLP	SVM
	75.50	81.05	81.15		62.50	79.27	79.40

According to the survey, accuracy of SVM is better than other three methods when N-gram feature was used.

The four methods discussed in the paper are actually applicable in different areas like clustering is applied in movie reviews and SVM techniques is applied in biological reviews & analysis. Although the field of opinion mining is new, but still diverse methods available to provide a way to implement these methods in various programming languages like PHP, Python etc. with an outcome of innumerable applications. From a convergent point of view Naïve Bayes is best suitable for textual classification, clustering for consumer services and SVM for biological reading and interpretation.

ACKNOWLEDGEMENTS

Every good writing requires the help and support of many people for it to be truly good. I would take the opportunity of thanking all those who extended a helping hand whenever I needed one. I offer my heartfelt gratitude to Mr. Mohd. Shahid Husain, who encouraged, guided and helped me a lot in the project. I extent my thanks to Miss. Ratna Singh (fiancee) for her incandescent help to complete this paper.

A vote of thanks to my family for their moral and emotional support. Above all utmost thanks to the Almighty God for the divine intervention in this academic endeavor.

REFERENCES

[1] Ion SMEUREANU, Cristian BUCUR, Applying Supervised Opinion Mining Techniques on Online User Reviews, Informatica Economică vol. 16, no. 2/2012.

[2] Bo Pang and Lillian Lee, "Opinion Mining and Sentiment Analysis", Foundations and TrendsR_ in Information Retrieval Vol. 2, Nos. 1–2 (2008).

[3] Abbasi, "Affect intensity analysis of dark web forums," in Proceedings of Intelligence and Security Informatics (ISI), pp. 282–288, 2007.

[4] K. Dave, S. Lawrence & D. Pennock. \Mining the Peanut Gallery: Opinion Extraction and Semantic Classi_cation of Product Reviews." Proceedings of the 12th International Conference on World Wide Web, pp. 519-528, 2003.

[5] B. Liu. \Web Data Mining: Exploring hyperlinks, contents, and usage data," Opinion Mining. Springer, 2007.

[6] B. Pang & L. Lee, \Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales." Proceedings of the Association for Computational Linguistics (ACL), pp. 15124,2005.

[7] Nilesh M. Shelke, Shriniwas Deshpande, Vilas Thakre, Survey of Techniques for Opinion Mining, International Journal of Computer Applications (0975 – 8887) Volume 57– No.13, November 2012.

[8] Nidhi Mishra and C K Jha, Classification of Opinion Mining Techniques, International Journal of Computer Applications 56 (13):1-6, October 2012, Published by Foundation of Computer Science, New York, USA.

[9] Oded Z. Maimon, Lior Rokach, “Data Mining and Knowledge Discovery Handbook” Springer, 2005.

[10] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. “Sentiment classification using machine learning techniques.” In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86.

[11] Towards Enhanced Opinion Classification using NLP Techniques, IJCNLP 2011, pages 101–107, Chiang Mai, Thailand, November 13, 2011

Author

Pravesh Kumar Singh is a fine blend of strong scientific orientation and editing. He is a Computer Science (Bachelor in Technology) graduate from a renowned gurukul in India called Dr. Ram Manohar Lohia Awadh University with excellence not only in academics but also had flagship in choreography. He mastered in Computer Science and Engineering from Integral University, Lucknow, India. Currently he is acting as Head MCA (Master in Computer Applications) department in Thakur Publications and also working in the capacity of Senior Editor.