ANALYTICAL STUDY OF FEATURE EXTRACTION TECHNIQUES IN OPINION MINING

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

[1]M.Tech, Department of Computer Science and Engineering, Integral University, Lucknow, India
erpraveshkumar@gmail.com

[2]Assistant Professor, Department of Computer Science and Engineering, Integral University, Lucknow, India
siddiquisahil@gmail.com

ABSTRACT

Although opinion mining is in a nascent stage of development but still the ground is set for dense growth of researches in the field. One of the important activities of opinion mining is to extract opinions of people based on characteristics of the object under study. Feature extraction in opinion mining can be done by various ways like that of clustering, support vector machines etc. This paper is an attempt to appraise the various techniques of feature extraction. The first part discusses various techniques and second part makes a detailed appraisal of the major techniques used for feature extraction.

KEYWORDS

Opinion Mining, Feature Extraction, Clustering, Support Vector Machines.

1. INTRODUCTION

People are generally more interested in other's opinion and when it comes to company product then it becomes even more important. So, the information gathering behaviour makes us to collect and understand other's views. Increasing networking capabilities provide a way for surprisingly large multitude of resources that contains options like blogs, reviews etc. As a result, now businesses are trying to analyse these floating opinions for competitive advantages. Now with the burst of activities in the area of sentiment analysis they deal with opinions in the first class object as there is increasing interest in the systems of opinion mining.

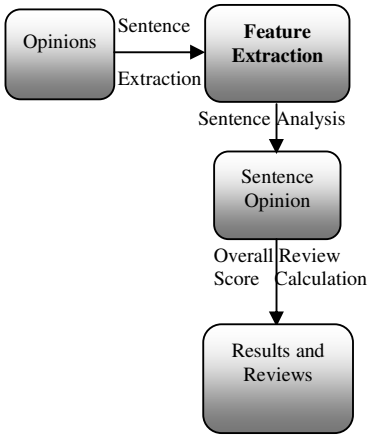

Figure 1: Opinion Mining Process

A fundamental step in opinion-mining and sentiment-analysis applications is feature extraction. The combined opinion mining procedure is illustrated in figure 1. Several methods are used for feature extraction among which, following are the important ones:

1) Naïve –Bayes Classifier (NB)
2) Support Vector Machine (SVM)

Sundarapandian et al. (Eds) : ACITY, AIAA, CNSA, DPPR, NeCoM, WeST, DMS, P2PTM, VLSI - 2013
pp. 85–94, 2013. © CS & IT-CSCP 2013 DOI : 10.5121/csit.2013.3410

3) Multi Layer Perceptron (MLP)
4) Clustering Classifier

In this paper, I have categorized the work done for feature extraction and classification in opinion mining and sentiment analysis. We have also tried to find out the performance, advantages and disadvantages of different techniques.

2. DATA SETS

This section provides brief details of datasets used by us in our experiments.

2.1 Product Review Dataset

The following dataset is taken from the work of blitzer and is of multi-domain characteristics. It consists of product reviews taken from amazon.com which belongs to a total of 25 categories like toys, videos, reviews etc. From the randomly selected five domains 4000 +ve and 4000-ve reviews are randomly samples.

2.2 Movie Review Dataset

The second dataset is taken from the work of pang & lee (2004). It contains movie review with of the feature of 1000+ve and 1000-ve processed movie reviews.

Table 1: Result of Simple n-gram

	Movie Review	Product Reviews
Unigram only	64.1	42.91
Bigram	76.15	69.62
Trigram	76.1	71.37
(uni+bi) gram	77.15	72.94
(uni+bi+tri) gram	80.15	78.67

3. CLASSIFICATION TECHNIQUES

3.1 Naïve Bayes Classifier

This classifier is based on the probability statement that was given by Bayes. This theorem provides conditional probability of occurrence of event E_1 when E_2 has already occurred, the vice versa can also be calculated by following mathematical statement.

$$P(E_1 \mid E_2)=\frac{P(E_2 \mid E_1)P(E_1)}{E_2}$$

This basically helps in deciding the polarity of data in which opinions / reviews / arguments can be classified as positive or negative which is facilitated by collection of positive or negative examples already fed.

Naïve Bayes algorithm is implemented to estimate the probability of a data to be negative or positive. The aforesaid probability is calculated by studying positive and negative examples & then calculating the frequency of each pole which is technically termed as learning. This learning is actually supervised as there is an existence of examples[1]. Thus, the conditional probability of a word with positive or negative meaning is calculated in view of a plethora of positive and negative examples & calculating the frequency of each of class.

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

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

Algorithm is shown in figure 2:

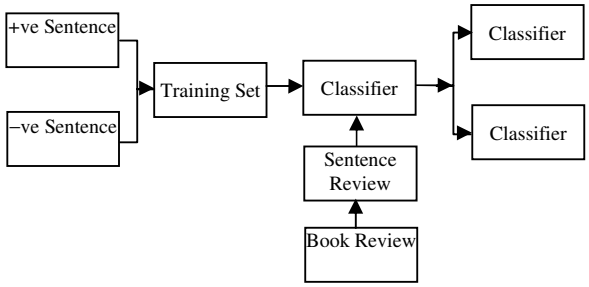

Figure 2

Algorithm has following **steps**

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)}

3.1.1. Evaluation of Algorithm

The measures used for algorithm evaluation are:

- Accuracy
- Precision
- Recall
- Relevance

Contingency table for analysis of algorithm:

	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 = $\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$; Recall = $\frac{tp}{tp+fn}$

[1] The concept is to actually a fact that each & every word in a label is already referred in learning set.

3.1.2. Accuracy

[1] Ion SMEUREANU, Cristian BUCUR, training the naïve gauss algorithm on 5000 sentences and get 0.79939209726444 accuracy where no of groups (n) is 2.

3.1.3. Advantages of Naïve Bayes Classification Methods

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

3.1.4. Disadvantage of Naïve Bayes Classification Methods

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

3.2 Support Vector Machine (SVM)

The basic goal of support vector machine is to search a decision boundary between two classes that is excellently far away from any point in the training data[2].

SVM develops a hyper planes or a set of hyper planes in infinite dimension space. This distance from decision surface to closest data point determines the margin of classifier. So the hyper planes act as decision surface which act as criteria to decide the distance of any data point from it. The margin of classifier is calculated by the distance from the closest data point. This successfully creates a classification but a slight error will not cause a misclassification.

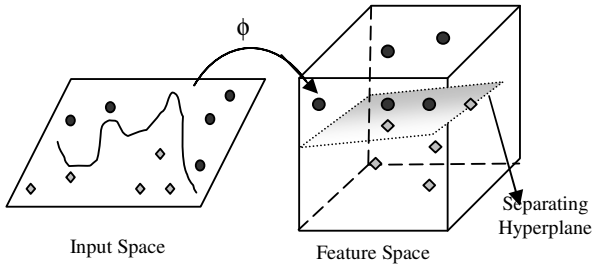

Figure 3: Principle of Support Vector Machine

For a training data set D, a set of n points:

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

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:

$$w \bullet x - b = 1 \quad \text{.....(2)}$$

Finding a maximum margin hyperplane, reduces to finding the pair w and b, such that the distance between the hyperplanes is maximal while still separating the data. These hyperplanes (fig. 1) are described by:

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

The distance between two hyperplanes is $\frac{b}{\|w\|}$ and therefore $\|w\|$ needs to be minimized. Or we can say that minimize $\|w\|$ in w, b subject to $c_i (w.x_i - b) \geq 1$ for any i = 1, … , n.
Using Lagranges multipliers (α_i) this optimization problem can be expressed as:

$$\min_{w,b} \max_{\alpha} \{\frac{1}{2}\|w\|^2 - \sum_{i=1}^{n}\alpha_i[c_i(w.x_i - b) - 1]\} \quad .....(3)$$

3.2.1. Extensions of SVM

To make SVM more robust and also more adaptable to real world problems, it has some extensions, some of which include following:

1. Soft Margin Classification

For very high dimensional problems, common in text classification, sometimes, data are linearly separable. For multidimensional problems like in classification of text, data are linearly separable sometimes. But in most cases, the opinion solution is the one that classifies most of the data and ignore some outliers or noisy data. If the training set D cannot be separated clearly then the solution is to have fat decision classifiers and make some mistake.

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

2. Non-linear Classification

The basic linear SVM was given by Vapnik (1963) later on Bernhard Boser, Isabelle Guyon and Vapnik in 1992 pavd a way for non-linear classifiers by using kernel to max. margin hyper planes. The only difference from th kernel trick given by Aizerman is that every dot product is replaced by non-linear kernel function.

The effectiveness of SVM in this case lies in the selection of the kernel and soft margin parameters.

3. Multiclass SVM

Generally SVM is applicable for two class tasks. But multiclass problems can be deal with multiclass SVM. In this case labels are designed to instances which are drawn from a finite set of various elements. These binary classifiers can be built by two classifiers like by either distinguish one versus all labels or between every pair of classes one versus one.

3.2.2. Accuracy

Usenet reviews were classified by pang by using numerical ratings which were accompanied as basic truth. Various learning methods are used but unigrams gave best outputs in a presence based frequency model run by SVM. The accuracy undergone is the process was 82.9%.

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. Categorical or missing values need to be pre-processed.
2. Difficult interpretation of resulting model.

[2] Possibly discounting some points as outliers or noise.

3.3 Multi-Layer Perceptron (MLP)

MLP is a neural network which is feed forward with one or more layers between input and output. Feed forward implies that, data flows in one direction i.e., from input layer to output layer (i.e., in forward direction). This ANN which multilayer perceptron starts with input layer where each node or neuron means a predicator variable. Input neurons are connected with each neuron in hidden layers. The neurons in hidden layer are in turn connected to neuron in other hidden layers. The output layer is made up of one neuron in case of binary prediction or more than one neuron in case of non-binary prediction. Such arrangement makes a streamlined flow of information from input layer to output layer.

MLP technique is quite popular owing to the fact that it can act as universal function approximator. A "back propagation" network has at least one hidden layer with many non-linear units that can learn any function or relationship between group of input variable whether discrete and for continuous and output variable whether discrete and for continuous. This makes the technique of MLP quite general, flexible ad non-linear tools.

When output layers is to be classified that has total number of nodes as total number of classes and the node having highest value then it gives the output i.e., estimate of a class for which an input is made. If there is a special case of two classes than, generally there is a node in output layer and classification is carried between two classes and is done by applying cut off point to node value.

An advantages of this technique, compared to classical modelling techniques, is that it does not impose any sort of restriction with respect to the starting data (type of functional relationship between variables), neither does it usually start from specific assumptions (like the type of distribution the data follow). Another virtue of the technique lies in its capacity to estimate good models even despite the existence of noise in the information analyzed, as occurs when there is a presence of omitted values or outlier values in the distribution of the variables. Hence, it is a robust technique when dealing with problems of noise in the information presented; however, this does not mean that the cleaning criteria of the data matrix should be relaxed.

3.3.1. Accuracy

Ludmila I. Kuncheva, Member, IEEE on health care data calculated accuracy of Multilayer perceptron (MLP) as 84.25–89.50%.

3.3.2. Advantages of MLP

1) Capable of acting as a universal function approximator.
2) Capability to learn almost any relationship between input and output variables.

3.3.3. Disadvantages of MLP

1) Flexibility lies in the need to have sufficient training data and that it requires more time for its execution than other techniques.
2) It is somewhat considered as complex "black box".

3.4 Clustering Classifier

To identifying the prominent features of human and objects and recognizing them with a type one can require the object clustering.

Basically clustering is unsupervised learning technique.

The objective of clustering is to determine a fresh or different set of classes or categories, the new groups are of concern in themselves, and their valuation is intrinsic. In this method, data objects or instances groups into subset in such a method that similar instances are grouped together and various different instances belong to the different groups.

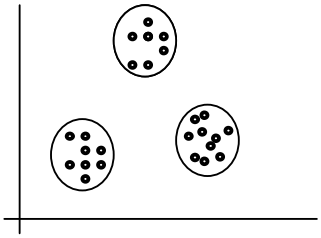

Figure 4

Clustering is an unsupervised learning task, so no class values representing a former combination of the data instances are given (situation of supervised learning).

Clustering basically assembles data objects or instances into subset in such a method that similar objects are assembled together, while different objects belong to different groups. The objects are thereby prepared or organized into efficient representations that characterize the population being sampled.

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 the figure 4 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.

There are two main types of measures used to estimate this 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 conclude the similarity and difference between the pair of instances, many clustering approaches usage the distance measures.

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.

Distance measures method is known as metric distance measure if follow the following properties:

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

$$\forall x_i, x_j, x_k \in S$$

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

$$\forall x_i, x_j \in S$$

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

3.4.1. Clustering Methods

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. Farley and Raftery (1998) gave a classification of clustering methods into two main groups: hierarchical and partitioning methods. Han and Kamber (2001) gave a different categorization: density-based methods, model-based clustering and gridbased methods. A totally different categorization based on the induction principle is given by (Estivill-Castro, 2000).

1. Hierarchical Methods

According to this method clusters are created by recursive partitioning of the instances in either a top-down or bottom-up way.

2. Partitioning Methods

Here, starting from the first (starting) partition, instances are rearranged by changing the position of instances from one cluster to another cluster.

The basic assumption is that the number of clusters will be set by the user before. In order to get global optimality in partitioned-based clustering, an exhaustive enumeration process of all possible partitions is done.

3. Density-based Methods

The basis of this method is probability. It suggests that the instances that belong to each cluster are taken from a specific probability distribution. The entire distribution of the data is supposed to be a combination of numerous disseminations. The objective of these methods is to identify the clusters and their distribution parameters. This particular method is designed in order to discover clusters of arbitrary shape which are not necessarily convex.

4. Grid-based Methods

In grid-based method all the operations for clustering are performed in a grid structure and for performing in the grid structure all the available space is divided into a fixed number of cells.

The basic main advantage of this method is its speed (Han and Kamber, 2001).

5. Soft-computing Methods

This includes techniques like that of neural networks.

3.4.2. Evaluation Criteria Measures for Clustering Technique

Generally, these criteria are split into two groups named Internal and External.

1. Internal Quality Criteria

This criteria generally measures compactness of clusters using similarity measures. 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

They are useful for examining the structure of the clusters match to some already defined classification of the objects.

3.4.3. Accuracy

Several different classifiers were used and the accuracy of the best classifier varied from 99.57% to 65.33%, depending on the data.

3.4.4. Advantages of Clustering Method

The main advantage of this method is that it offers the groups that (approximately) fulfil an optimality measure.

3.4.5. Disadvantages of Clustering Method

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

4. CONCLUSION

The important part to gather information always seems as what the people think. The rising accessibility of opinion rich resources such as online analysis websites and blogs rises as one can simply search and recognize the opinions of some other one. One can precise his/her ideas and opinions concerning goods and facilities. These view and thoughts are subjective figures which signifies someone opinions, sentiments, emotional state or evaluation. In this paper, we present different methods for data (feature or text) extraction and every method have 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 2.

Table 2: Accuracy of Different Methods

	Movie Reviews			Product Reviews		
Ngram Feature	NB	MLP	SVM	NB	MLP	SVM
	75.50	81.05	81.15	62.50	79.27	79.40

According to our survey, accuracy of MLP is better than other three methods when we use Ngram feature.

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 in the field of opinion mining is new, but still the 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 endeavour.

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.