Comparative Analysis of Sentiment Analysis Techniques

Chetan Kaushik, Atul Mishra

Computer Engg. Department, YMCA University of Science & Technology, Faridabad
Email: chetankaushik8407@gmail.com, mish.atul@gmail.com

Abstract—With the increase in the volume of sentiment rich social media websites, an increasing interest among researchers can be seen regarding Sentiment Analysis and opinion mining. The requirement is to develop a technique that can differentiate between the positive, negative or neutral sentiment underlying an electronic text. By devising an accurate method to identify the sentiments behind any text, one can predict the mood of the people regarding a particular product or service. However, there are various challenges involved in identifying the correct sentiment of a user which are discussed in our research. In this paper, we will discuss various techniques of sentiment analysis and the challenges associated with it.

I. INTRODUCTION

Sentiment Analysis can be described as a type of Natural Language Processing which includes obtaining the feeling of a user or a group of users expressed in various comments, requests or questions posted by them on the internet. It involves building a system that can collect the user opinions and then examine and classify them according to the polarity of the post. In other words, sentiment analysis aims to determine the view of a speaker or writer on particular subject. Liu [1] defined a sentiment as a quintuple “<o, f, soijkl, hi, tl>, where o is a target object, f is a feature of the object o, soijkl is the sentiment value of the opinion of the opinion holder hi on feature f of object o at time tl, soijkl is +ve, -ve, or neutral, or a more granular rating, hi is an opinion holder, tl is the time when the opinion is expressed.”

Sentiment Analysis has applications in various fields. For example, in marketing it helps in determining the success or failure of a new product launch or any new commercial campaign or determining that which version of a product is liked more in which part of the world. Various companies can use this data to determine their future strategies regarding a particular product or service.

Sentimental Analysis can be based on a document, sentence or a phrase. In document based sentimental analysis, sentiment of the whole document is calculated as a whole and summarized according to the polarity. In sentence based sentiment analysis, individual sentences are classified as positive, negative or neutral whereas phrase based sentimental analysis assigns a polarity to the individual phrases contained in a sentence.

The first requirement of sentimental Analysis is to find the subject towards which the opinion is expressed. After that the sentiment is classified as positive (which denotes satisfaction or happiness on behalf of user), negative (which shows rejection or disappointment) or neutral (which denotes no strong sentiment involved). Then the sentiment can be given a score which denotes the degree of positive or negative response from the user.

There are various challenges involved with sentimental analysis. Subabrata [2] categorized these challenges as following:

A. Implicit Sentiment
Sometimes a sentence may carry a strong sentiment without containing any sentiment bearing word in it. For e.g. One has to be on a lot of medications to make such a documentary

B. Domain Dependency
Some words have different polarity when used in different domains. For e.g.
The movie was inspired from a Hollywood movie.
I got inspired by this book.

C. Thwarted Expectations
Sometimes the writer builds up a positive context and refute it in the end. For e.g. Excellent performances, very good music, stunning cinematography, all in vain because of lack of imagination of the writer/director.

D. Pragmatics
The pragmatics of the user needs to be identified. For e.g.
It was good to see India destroy Australia in final.
The match destroyed my interest in sports.

E. World Knowledge
Sometimes the knowledge of an entity which is used in the sentence is required to identify the sentiment. For e.g.
He is just as good a person as Dracula
One has to know about Dracula to understand the correct sentiment behind this sentence.

F. Subjectivity Detection

It is important to differentiate between sentiment rich and neutral sentences from each other. For e.g.

I love Tokyo.
I hate the movie ‘love in Tokyo’.

G. Entity Identification

There may be multiple entities in a sentence it is important to identify that the sentiment is directed towards which entity.

Chelsea is better than Man. Utd.
This statement is +ve for Chelsea and –ve for Man. Utd.

H. Negation

Handling negation is very difficult. One method is to reverse the polarity of every word that comes after a negative word (e.g. not). For e.g.

I do not like this movie.
However this method will fail for-
Not only was the food delicious, the service was excellent.

II. FEATURES FOR SEMANTIC ANALYSIS

Feature engineering is a basic task in performing sentiment analysis. It includes converting a piece of text into a feature vector. This section includes some commonly used features for sentiment analysis.

A. Term frequency and term presence

Term frequency refers to the number of times a term is repeated in a piece of text. It is considered to be very important in conventional text classification tasks. But in sentiment analysis it is observed that term presence bears more importance then term frequency because sometimes the presence of a single term can reverse the polarity of the whole sentence.

B. Term position

Sometimes the terms appearing at one section of the text contains more weight than the terms appearing at a different section. For example a negation at the beginning of the sentence can change the meaning of the entire sentence. Generally terms at the first and last few sentences in a text are given more weightage then the terms which appear in the middle.

C. N-gram features

N-grams are used widely in natural language processing tasks for identifying context. However it is not clear that whether higher order N-grams perform better than the lower order N-grams or not.

D. Subsequence kernels

Generally, word or sentence level modes are used for sentiment analysis. Bickel [3] used a method in which subsequence kernels were used to implicitly capture the feature space.

The word subsequence kernel of order n are weighted some of all word sequences of length n that occur in both the strings which are being compared.

The mathematical formula is

\[ K_n(s, t) = \sum_{u \in \Sigma^n} \sum_{i=j}^{i=n|u|} \lambda^{(i[n]-i[1]+1)+(i[n]-j[1]+1)} \]

Where i refers to a vector of length n that consists of the indices of string s that correspond to the subsequence u, \( \lambda \) is a kernel parameter similar to gap penalty and \( i[n] - i[1] + 1 \) is the total length of the span of s that constitutes a particular occurrence of the subsequence u.

E. Adjectives only

Adjectives are the most commonly used features in sentiment analysis. People generally use adjectives to depict their sentiments and high accuracy is observed in sentiment analysis techniques that focus on only adjectives for sentiment analysis.

F. Adjective – Adverb Combination

Adverbs generally has no polarity, but when added to an adjective they can contribute heavily in determining the polarity of a sentence.

Benamara[4] showed how adverbs can alter the sentiment value of a sentence and can be classified as

1. Adverbs of affirmation: certainly, totally
2. Adverbs of doubt: maybe, probably
3. Strongly intensifying adverbs: exceedingly, immensely
4. Weakly intensifying adverbs: barely, slightly
5. Negation and minimizers: never

Two types of AACs were defined by the work:

Unary AAC: contains one adjective and one adverb
Binary AAC: contains more than one adjective and adverb

III. RELATED WORK

A lot of research has been done on Sentiment Analysis or Opinion Mining of data. These studies focus on determining the correct sentiment behind an electronic text. Most of these approaches can be classified under two types – machine learning and semantic orientation. This section discusses the existing work on both of these approaches.
A. Machine learning

A machine learning strategy involves two sets of documents. A training set and a test set. The machine learning algorithm first needs to be trained for both supervised learning tasks (like classification, prediction etc.) and unsupervised learning tasks (clustering etc.). In training phase the algorithm is trained with some particular inputs so that later on it can be tested for unknown inputs. The objective is to train our algorithm in such a way that later it becomes able to classify new unknown inputs. There are several machine learning methods which are being used for Sentiment Analysis. Some of them are discussed in this section.

Naive Bayes is one of the most effective and simple approach amongst them. It is widely used as an algorithm for classification of text (Melville [5], Rui [6], Ziqiong [7], Songho [8], Qiang [9] and Smeureanu[10]). In this approach, first the prior probability of an entity being a class is calculated and the final probability is calculated by multiplying the prior probability with the likelihood. The method is naïve in the sense that it assumes every word in the text to be independent. This assumption makes it easier to implement but less accurate.

Another approach is Support Vector Machines (SVM). It is also used for text classification based on a discriminative classifier (Rui [6], Ziqiong [7], Songho [8], and Rudy [11]). The approach is based on the principle of structural risk minimization. First the training data points are separated into two different classes based on a decided decision criteria or surface. The decision is based on the support vectors selected in the training set. Several different variants of SVM are used, one of them is a multiclass SVM used for Sentiment Analysis [12].

The centroid classification algorithm [8] first calculates the centroid vector for every training class. Then the similarities between a document and all the centroids are calculated and the document is assigned a class based on these similarities values.

The K-Nearest Neighbor (KNN) approach [8] finds the K nearest neighbors of a text document among the training documents. The classification is done on the basis of the similarity score of the class to the neighbor document.

Winnow is another commonly used approach. The system first predicts a class for a particular document and then receives feedback. If a mistake is detected then the system updates its weight vectors accordingly. This process is repeated over a collection of sufficiently large set of training data.

Rudy [11] proposed a method based on a combined approach which included rule based classification, supervised learning and machine learning. A 10 fold cross validation was carried out for each sample set. A hybrid classification method is used in which several classifiers work together. If the first classifier fails to classify then it is passed on to the next classifier. The process continues until the document is classified or there is no other classifier left.

Ensemble technique [6] combines the output of several classification methods into a single integrated output. Zhu [13] proposed an approach based on artificial neural networks to divide the document into positive, negative and fuzzy tone. The approach was based on recursive least squares back propagation training algorithm.

Long-Sheng [14] combined the advantages of machine learning and information retrieval techniques using a neural network based approach.

B. Semantic Orientation

The semantic orientation approach is based on unsupervised learning. It doesn’t require any training in order to classify the sentiment data. It is used to measure how much positive or negative is the word’s polarity.


Andrea [16] proposed a method based on semi supervised learning, which introduced a seed set and expanded it later using Word Net. The assumption was that the words with similar orientation have similar polarity.

Chunxu[17] proposed a method to perform sentiment analysis on content whose contextual information is not known in advance. In this method other related contents were used to extract the required contextual information and then used the information for determining the orientation of the opinion.

Ting-Chun [18] proposed an unsupervised learning algorithm based on part of speech (pos) pattern. They used the sentiment phrase as a query for a search engine and sentiments were predicted based on the search results.

Gang [19] used TF-IDF (term frequency – inverse document frequency) weighing for sentiment analysis. They used K- means clustering on raw data, and then a voting mechanism to further stabilize the clustering. Multiple implementations of the process was applied to classify the documents in to positive and negative groups.

Prabhu [20] used a simple lexicon based technique on twitter data by identifying and extracting sentiments from hashtags and emoticons.

IV. COMPARISON AND EVALUATION

The performance of various sentiment analysis techniques was measured on the basis of accuracy. That is, what percentage of text was accurately classified by the sentiment analysis technique? The performance of different studies discussed earlier are represented in fig 1 and a brief comparison different techniques used in them is shown in table 1. The sources used for evaluation is mostly movie reviews or product reviews.

It was observed that movie review is a more challenging task.
task as compared to product reviews because people use more ironic terms while writing movie reviews hence movie review sentiment analysis is a much more challenging task.

From the performance evaluation, it is difficult to choose one particular technique that stands out, since each method used different sources for training and collection of document with varying text granularity and feature selection methods.

However it is observed that the machine learning approaches show more accuracy than semantic orientation approaches. But machine learning approaches require more time for training. Semantic orientation on the other hand is more useful for real time applications.

![Fig.1 - Performance of different studies](image-url)

### Table I - Summary and comparison of various sentiment analysis techniques

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Technique</th>
<th>Learning Methodology</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Study</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SVM</td>
<td>Supervised</td>
<td>Very high accuracy, Lesser overfitting, Robust to noise</td>
<td>Incapable of multiclass classification, Computationally expensive, Slow</td>
<td>Kaiquan Xu (2011)</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rui Xia (2011)</td>
<td>86.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ziqiong (2011)</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pang and Lee (2004)</td>
<td>86.4</td>
</tr>
<tr>
<td>2</td>
<td>Naïve bayes</td>
<td>Supervised</td>
<td>Faster training and classification, Not sensitive to irrelevant features, Handles streaming data well</td>
<td>Assumes independence of feature, Less accurate than SVM</td>
<td>Rui Xia (2011)</td>
<td>85.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Xue Bai (2011)</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Gamon (2005)</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pang and Lee (2004)</td>
<td>86.4</td>
</tr>
<tr>
<td>3</td>
<td>Centroid classifier</td>
<td>Supervised</td>
<td>Low computation cost, High dimensional data set, Can combine multiple features together</td>
<td>Term dependency within class, Too sensitive to the training data, Large number of features in feature vector</td>
<td>Songhotan (2008)</td>
<td>90</td>
</tr>
<tr>
<td>4</td>
<td>KNN</td>
<td>Supervised</td>
<td>Very fast training, Simple and easy to understand, Robust to noisy training data, Handles large data set well</td>
<td>Biased by value of K, High computation complexity, Gets easily fooled by irrelevant attributes</td>
<td>Songhotan (2008)</td>
<td>90</td>
</tr>
<tr>
<td>5</td>
<td>Winnow classifier</td>
<td>Supervised</td>
<td>Mistake driven approach, More sensitive to relationship among features, Weights of only active features are updated</td>
<td>Less precise than SVM, Tuning not robust on different training collections</td>
<td>Gang Li (2010)</td>
<td>78</td>
</tr>
<tr>
<td>6</td>
<td>K – means clustering</td>
<td>Unsupervised</td>
<td>Faster than supervised learning methods, Easy to implement, Produces tight clusters</td>
<td>Less accurate than supervised learning, Difficult to predict value of K, Doesn’t work well for clusters of different sizes and density</td>
<td>Gang Li (2010)</td>
<td>78</td>
</tr>
</tbody>
</table>
V. CONCLUSION

Sentiment analysis is being used for different applications and can be used for several others in future. It is evident from the above discussion that no classification method outperforms others consistently. It is also observed that different techniques can be combined to overcome each other’s limitation and provide a better classification all around. More work is needed in order to further improve the classification techniques. Several problems such as handling of implicit product features and dealing with negation etc. are still not completely resolved.

REFERENCES


