ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

A Survey on Sentiment Analysis Techniques and Methods

Maitri Patel[1], Jayna Shah[2]

M.E Student, Department of Computer Engineering, Sardar Vallabhbhai Patel Institute of Technology, VASAD, India[1]

Assistant Professor, Department of Computer Engineering, Sardar Vallabhbhai Patel Institute of Technology, VASAD, India [2]

ABSTRACT: Sentiment Analysis is a task of extracting information from people opinions towards different entities. Nowadays in era of internet, every person shares their information in social network sites, blogs, product review websites, web forums etc. Thus, the thoughts of other people provide opinion about different entities that helps in decision making process to manufacturer. Sentiment analysis is the process in text mining that helps finding the opinions, sentiment and subjectivity of text. This survey paper provides a comprehensive overview of such sentiment analysis feature selection and classification techniques. The main aim to this survey is to provide existingsentiment analysis techniques and methods that help in sentiment analysis researchers.

KEYWORDS: Sentiment Analysis, Text mining, Feature selection, Classification.

I.INTRODUCTION

Sentiment analysis is a process of information gathering task to attain user's feelings. By analyzing a large numbers of documents, these feelings can be expressed in positive or negative, good or bad, happy or unhappy etc. like in different ways in the form of comments, questions and requests. Generally, sentiment analysis helps to find the attitude of a writer about any topic or the overall sentiment of a document or text that's helps to know what kind of people actually think.

Due to the exponential enhancement in the Internet usage and replacement of public opinions over the internet rather than to ask friends or relatives, sentiment analysis becomes an important process in today's life. The Web is a huge depository of ordered and amorphous data. The analysis of this data to extract hidden public opinions and sentiment is not an easy task. Machine learning technique applied to the dataset gives better result than the human generated result.

Sentiment analysis has been classified at three levels: Document Level, in these whole documents expresses a positive or negative sentiment, Sentence Level, this level goes to the sentences and determines whether each sentence express a positive, negative or neutral opinion, and Aspect Level analysis first locates an opinion content about an aspect in a review rather than focusing on document, paragraph, Claus etc.

II.RELATED WORKS

A. DATA CLEANING

In order to fit our model to the dataset we need to clean or process our data. For that, Data cleaning process consider as the preprocessing step in text analysis that involves Stopword removal, stemming, Tokenization divides given text into token, Conjunction rule, Negation rule, Part of Speech tagging by POS tagger.

Stemming and POS tagging is done in[1]. Special character removal and hyperlink removed from the tweets performed in[3]. A number of steps to be performed for sentiment analysis task before the actual analysis of a comment;these steps are called the preprocessing steps and arelisted below:

- To do sentiment analysis at sentence level, breaking down the comment by punctuation character like '!',).? And further breaking sentence by conjunction like (but, and, or, for, nor, so).
- Removing remaining punctuation characters and the special characters from the sub sentences. Symbols like @, >, <, #, :and many more have been removed.
- In the next, the articles, pronouns, prepositions, auxiliary verbs and the 'Wh' words have been removed from sentences.

B. SUBJECTIVITY ANALYSIS

Generally, Textual data may express positive or negative opinion about things. To determine whether the text is Subjective or Objective is called Subjectivity Analysis. In analysis of movie review subjectivity is the main task for opining mining.

Online reviews may consist both subjective and objective sentences. From these objective sentences consists only factual information about sentiment or opinion while subjective sentences consists opinion so for the further analysis only subjective sentences are consider. For example, given an product review, it determines whether the reviewer is positive or negative about the product this classification done at sentence or Claus. So, feature and opinion pair can be easily found from subjective sentences. Here, in this paper for the subjectivity analysis they used SentiWordNet approach and find out opinion and feature pair based on the rule-based system[1].

C. FEATURE SELECTION

Sentiment Analysis task is considered a sentiment classification problem. The first step in sentiment classification problem is to extract and select text feature. Feature selection technique treat documents either group of words, or as a string which retains the sequence of words in the document. Some feature selection methods describe below:

1. **Document frequency:**

Document Frequency denotes the number of documents in which a term occurs. DF thresholding is the simplest technique for vocabulary reduction in text classification. The document frequency of each term in the training corpus will be computed and the terms with a document frequency less than a predefined threshold will be discarded from the vocabulary. The DF thresholding method assumes that the terms with higher document frequency are more informative for classification. But this assumption may sometimes lead to bad classification accuracy if a term occurs in most of the documents in each class (e.g., stop words). The computational complexity of this method is approximately linear in the number of training documents. Hence it is scalable to any large corpus and usually considered as an ad-hoc approach for feature selection[6].

2. **Mutual Information (MI):**

The mutual information between the term x and y is defined as [6]:

$$MI(X;Y) = \sum_{x,y} p(x,y)\log(\frac{p(x,y)}{p(x)p(y)}) \qquad \text{eq. (1)}$$

This method assumes that the term with higher category ratio is more effective for classification. On the other hand the method is biased towards low frequency terms as can be seen from the following form[6] :

$$\text{MI}\ (t, c) = \log(t/c) - \log(t) \qquad \text{eq. (2)}$$

Rare terms will have a higher score than common terms for those terms with equal conditional probability (t/c). Hence MI might perform badly when a classifier gives stress on common terms. The MI of a term over all categories can be computed in two following ways [6]:

$$MI_{avg}(t) = \sum_{i=1}^{m} P(c_i)(t, c_i) \qquad \text{eq. (3)}$$

$$MI_{max}(t) = max_{i=1}^{m} \{\text{MI}\ (t, c_i)\} \qquad \text{eq. (4)}$$

For example, say a discrete random variable X represents visibility at a certain moment in time and random variable Y represents wind speed at that moment.

3. Information Gain (IG):

Information gain measures the number of bits of information obtained for category prediction by knowing the presence or absence of a term in a document. Let there be {i = 1, 2, ..,} classes in the target space, then it is defined as [6]:

$$IG(t) = -\sum_{i=1}^{m} P(c_i)logP(c_i) + P(t)\ \sum_{i=1}^{m} P(c_i/t)logP(c_i/t) + P(\bar{t})\sum_{i=1}^{m} P\ (c_i/\bar{t})logP(c_i/\bar{t}) \qquad \text{eq.(5)}$$

This measure gives more weight to common terms rather than the rare terms. Hence IG might perform badly when there is scarcity of common terms between the documents of the training corpus.

4. $\chi2$ statistic (CHI):

The $\chi2$ statistic is used to measures the association between the term and the category. CHI score between a term t and a class c is defined as[6]:

$$\chi2(t,\, c) = \mathrm{N}\times \frac{(\mathrm{P(t,c)} \times \mathrm{P}((\bar{t},\bar{c}) - \mathrm{P}(\mathrm{t},\bar{c}) \times \mathrm{P}(\bar{t},\mathrm{c}))^2}{\mathrm{P(t)}\times \mathrm{p}(\bar{t})\times P(c)\times P(\bar{c})} \qquad \text{eq. (6)}$$

Here N is the total number of documents. The $\chi2$ statistic of a term over all classes is combined in the following two ways:

$$\chi^2(t) = \sum_{i=1}^{m} P\ (c_i)^2(t,\, c_i) \qquad \text{eq.(7)}$$

$$\chi^2(t) = max_{i=1}^{m}\ \{\chi^2(t,\, c_i)\} \qquad \text{eq. (8)}$$

They have used the maximum CHI score for comparison in experiments. An entropy based feature ranking method, in which feature importance is measured by the contribution to an entropy index based on the data similarity. But the method has huge computational complexity and hence not suitable for a corpus with a large number of terms and documents.

A mutual information based feature selection method presented sequential forward selection methods based on an improved mutual information measure for text classification.

D. SENTIMENT CLASSIFICATION TECHNIQUE

Sentiment classification technique can be roughly divided into machine learning, lexicon based and hybrid approach. The machine learning approach applies machine learning algorithm to the linguistic feature. Supervised Methods make a use of a large number of labeled training documents. Lexicon based approach relies on a sentiment lexicon, a collection of known and precompiled sentiment terms. Lexicon based approach depends on finding opinion lexicon which is used to analyze the text. This approach divided into dictionary based approach and corpus based approach which use statistical or semantic methods to find sentiment polarity. The hybrid approach combines both approaches and is very common with sentiment lexicons playing a key role in the majority of methods.There is a brief explanation of both approaches. The various approaches and classification algorithm describe in Fig.1.

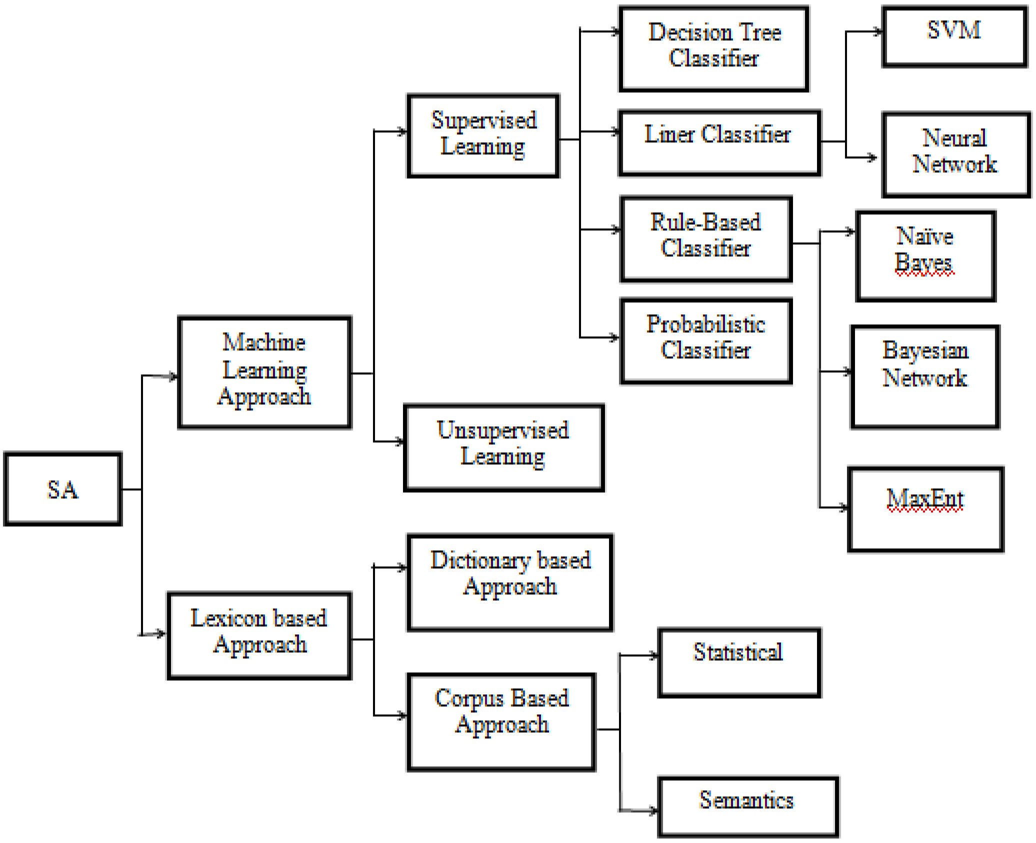

Fig.1. Sentiment Classification Techniques[4]

1. Machine Learning Approach:

Machine Learning Approach is generally relies on ML algorithm for the regular text classification problem. ML approach is divided into Supervised Learning and Unsupervised Learning as shown in Fig.1.

a. Supervised Learning: The supervised learning methods depend on the existence of labeled training documents. Naïve Bayes, Maximum Entropy, Decision Tree etc. are used for the classification.

i. Naïve Bayes:

The Naïve Bayes classifier is the simple and typically used classifier. Naïve Bayes classifier performs well in sentiment analysis these distinguish opinion as positive and negative very well. These classifier model works with the BOWs characteristic extraction that looks the position of the word within the document. These uses Bayes Theorem to predict the probability which might be offer set belong to a select label.Bayes Theorem is as follow:

$$P(Ci|X) = P(X|Ci)\cdot P(Ci) / P(X) \qquad \text{eq.(9)}$$

ii. Maximum Entropy:

The Maximum Entropy Classifier noted as a conditional exponential classifier. In Maximum Entropy Classification the probability of the document belongs to particular class given a context must maximize the entropy of the classification system. It processes same as describe in Naïve Bayes algorithm.

iii. Rule-based classifiers:

In rule based classifiers, the data space is modeled with a set of rules. The left hand side represents a condition on the feature set expressed in disjunctive normal form while the right hand side is the class label. The conditions are on the term presence. Term absence is rarely used because it is not informative in sparse data.

DOI: 10.15680/IJIRCCE.2017. 0501029

b.Unsupervised Learning: Unsupervised learning is the task of inferring a function to describe hidden structure from unlabeled data. The dataset given to the learner are unlabeled. Clustering is used in the unsupervised learning.

2. Lexicon Based Approach:

The lexicon-based approach depends with reference to find the feeling vocabulary that is use to analyze the content. There are two way that throughout this approach. The dictionary-based approach which depends on discovering opinion words then searches the lexicon of their synonyms and antonyms. The corpus-based approach starts with list of opinion words, and then finds different opinion words throughout a huge corpus to assist in discovering opinion words with context specific orientations. This is done by using applied mathematics or linguistics techniques[4].

This paper describes domain specific lexicon and adverb/adjective score is calculated using SentiWordNET[1]. In this paper classifying the Twitter dataset into positive, negative, neutral SentiWordNet 3.0.0 dictionary is used and itcontains 117659 words[3]. Lexicon Based Approach divided into two category Dictionary Based and Corpus Based Approach as shown in Fig.1.

a. Dictionary Based Approach:

Dictionary-based approaches begin with a predefined dictionary of positive and negative words, and then use word counts or other measures of word incidence and frequency to score all the opinions in the data. With a complete dictionary, the cost for automated analysis of texts is extremely low. Start with the small seed of opinion word. Used well known corpora WordNet to find out synonyms and antonym.

b. Corpus Based Approach:

That helps to solve the problem of finding opinion words with context specific orientations. This approach is applied to large corpora. Example: In the dataset the conjunction like AND, OR, BUT, EITHER-OR given. Sometimes conjoin adjectives have same orientation and there are also expressions such as BUT, HOWEVER which indicate opinion changes.

3. SentiWordNet:

SentiWordNet is a lexical resource. It contains positive, negative score for all terms. POS tagging and unique id given to each terms and tagging identified as, n-Noun, a-Adjective, v-Verb, r-Adverb. The value of PosScore and NegScore are the positivity and negativity of the terms. The objectivity score can be calculated as:

$$\text{ObjScore} = 1 - (\text{PosScore} + \text{NegScore}) \qquad \text{eq. (10)}$$

$$\text{PosScore} + \text{NegScore} + \text{ObjScore} = 1 \qquad \text{eq. (11)}$$

III. COMPARSTIVE ANALYSIS

Sr. No.	Author Name	Technique	Dataset	Advantage	Disadvantage	Accuracy
1	PurtataBhoir and ShilpaKolte	ML and Lexicon Based for finding subjectivity/objectivity and Rule-Based system for opinion analyzing	Movie Review	Subjectivity/ Objectivity Analysis is done.	Lexicon based approach does not gives better result in classification.	NB-71.42% SWN-53.33%
2	V.K. Singh, R. Piryani, A. Uddin and P. Waila	Lexicon Based Method	Movie Review	Can be used as an add-on step in movie recommendation system.	Domain Specific.	SWN(AAC)-77.6% SWN(AAAVC)-78.7%

DOI: 10.15680/IJIRCCE.2017. 0501029

3	Anurag P. Jain and Mr. Vijay D. Katkar	Machine Learning	Twitter Dataset	Classification as Naïve Bayes gives better results to sentiment analysis process.	Feature Selection Process is not used for sentiment Analysis.	RandomForest BayesNet KNN-91.0418%
4	Rajdeep Singh, RoshanBagla, HarkiranKaur	Lexicon Based Method	Social Network sites	Handling feature selection that usually appear best result to the system.	Domain specific so this customize system performance for limited corpus.	SWN-90.4%
5	A.jeyapriya and C.S.KanimozhiSelvi	Machine Learniing	Amzon and epinion sites	Sentence and aspect level analysis provide better performance	Refinement at feature level analysis needed.	NB- 80.36%

Table 1: Comparative Analysis

As shown in Table 1. From Literature Survey of different techniques conclude that subjectivity analysis is more important task in a movie review datasets and to find the subjectivity and objectivity using SentiWordNet gives result to analysis. In machine learning analysis before classification of the text or document feature selection techniques to the document helps effective classification process. Along with these effective feature selection techniques gives better results in sentiment classification. And Lexicon Based Approach helps in finding opinion using predefined positive and negative orientation dictionary. Variety of feature selection and classification algorithm gives better performance to the sentiment analysis process.

IV. PERFORMANCE ANALYSIS

Analyzed papers include various dataset those are namely Cornell movie review dataset, Twitter dataset, Customer dataset (amazon.com, epinions.com, cnet.com). From them movie review mining is more challenging reviews than other dataset review because real life opinion terms are mixed in movie review. For example unpredictable terms indicate negative opinion but it gives positive opinion for movie review [1].

The performance of sentiment analysis is calculated by using confusion matrix which is generated when algorithm is already implemented on dataset. For calculating performance various measures are used that are Precision, Recall, F-measure and Accuracy.

$$\text{Precision}=\frac{TP}{TP+FP}\ ,\qquad \text{Recall}=\frac{TP}{TP+FN}\ ,\qquad \text{F- measure}=2\times\frac{Precision\times Recall}{Precision+Recall}$$

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+TN+FN}$$

V. CONCLUSION

There are various research papers on sentiment analysis systems performing analysis on various domains like movie review, product review, news article etc. Some of them uses Machine learning approach and some of them uses lexicon based approach in order to solve challenges of sentiment analysis. Classification techniques and feature selection method affects the accuracy of the system. Machine learning approach will not give accurate result if the feature selection process is not done effectively. And in lexicon based approach if the dictionary contains less word then it leads to decrease the performance. So to improve performance we can use both approach together to maintain the accuracy of supervised machine learning approach and stability of orientation in lexicon based approach.

REFERENCES

[1]PurtataBhaoir and ShilpaKolte,'Sentiment Analysis of Movie Review Using Lexicon Approach',International Conference on Computational Intelligence and Computing Research IEEE, 2015.

[2] V.K. Singh, R. Piryani, A. Uddin and P. Waila,'Sentiment Analysis of Movie Reviews: A new Feature-based Heuristic for Aspect-level Sentiment Classification', IEEE, pp. 712-717, 2013.

[3] Anurag P. Jain and Mr. Vijay D. Katkar,'Sentiment Analysis of Twitter Data Using Data Mining', International conference on Information Processing Vishwakarma Institute of TechnologIEEE, 2015.

[4] WalaaMedhat, Ahmed hassan and HodaKorashy, 'Sentiment Analysis algorithms and applications: A Survey',Ain Shams Engineering Journal Elsevier, pp. 1093-1113, 2014.

[5] Rehab M. Duwairi and Islam Qarqaz, 'Arabic Sentiment Analysis using Supervised Classification', IEEE, 2014.

[6]TanmayBasu and C. A. Murthy, 'Effective Text Classification by a Supervised Feature Selection Approach', IEEE 12th International Conference on Data Mining WorksShops, pp. 918-925, 2012.

[7] J.K. Sing, SouvikSarkar and Tapas Kr. Mitra, 'Development of a Novel Algorithm for Sentiment Analysis Based on Adverb-Adjective-Noun Combinations', IEEE, 2012.

[8] Rajdeep Singh, RoshanBaglaand HarkiranKaur,'Text Analytics of Web Posts' Comments Using Sentiment Analysis', IEEE 2015.

[9] SangitaN. patel and Ms.Jignya B. Choksi, 'A Survey on sentiment Classification techniques', Journal of Research, pp. 15-20, 2015.

[10] Jeyapriya A. and Selvi K. 'Extract and mining opinions in Product Reviews using Supervised Learning Algorithm', IEEE SPONERED 2nd International Conference on Electronics and Communication Systems, pp. 548-552, 2015.

BIOGRAPHY

Maitri D. Patel is a M.E. Student in the Computer Engineering Department, SardarVallabhbhai Patel Institute of Technology, Vasad. Her Research interest in Data Mining.

Jayna B. Shahis an Assistant Professor in the Computer Engineering Department, SardarVallabhbhai Patel Institute of Technology, Vasad.

DOI: 10.15680/IJIRCCE.2017. 0501029