CHAPTER 2

LITERATURE REVIEW

2.1 SENTIMENT ANALYSIS

Opinion Mining (OM), sentiment mining, sentiment detection are the commonly used synonyms for Sentiment Analysis (SA). SA aims at detecting the sentiment expressed in customer review by using statistical self-learning techniques or NLP. The objective is to classify the review documents as either positive or negative based on the sentiment expressed by the customers in their reviews (Bing Liu 2008). This kind of classification is carried out at the document level, without discovering about what people liked or did not like.

Research in the field of sentiment analysis is carried out either using machine learning (ML) algorithms or by natural language processing (NLP) techniques. Recently the focus has shifted to combine the best of both ML and NLP techniques in the form of hybrid approaches. Also the soft computing technique like fuzzy logic is gaining popularity in carrying out the task of SA.

2.2 SENTIMENT ANALYSIS USING MACHINE LEARNING ALGORITHMS

Bo Pang et al (2002) attempted to classify the sentiments by treating them as a topic-based text classification problem. He used the standard ML techniques like NB, Maximum Entropy (ME) model, and SVM.

He tried the classification with only unigrams, unigrams and bigrams, and only bigrams etc. For the unigram based classification he achieved a classification accuracy of 81%, 80.4, 82.9% respectively using NB, ME, SVM. He used a review corpus consisting of 753 negative and 1301 positive reviews given by 144 reviewers.

Dave et al (2003) used a scoring function to assign a score to each term, which ranges from -1 to +1. The sum of scores of the words in an unknown document has been used to determine the class of a document. They pointed out the following issues relating to the sentiment analysis:

- Rating inconsistency – Inability of the reviewers to understand the rating system and give a 1 instead of a 5.
- Ambivalence and comparison – Some reviewers start with a lot of negative opinions but finally end up saying that overall they were satisfied. Some reviewers compare the positive experience with one product with a negative experience of another.
- Sparse data – Reviews being very short
- Skewed distribution – sometimes the availability of positive reviews are predominant and less number of negative reviews. Certain products and product types alone have more reviews.

They pointed out the importance of separating the types of reviews and parts within reviews so that they can be treated differently.

Pang and Lee (2004) proposed a novel machine-learning method that applies the text categorization to just the subjective portions of the document. Extracting the subjective portions in a review document was achieved by finding the minimum cut in graphs. They proposed the concept of subjectivity detectors and proved that by extracting the subjective sentences

from the whole document the classification accuracy of the default polarity classifiers like NB and SVM can be greatly improved. They achieved 86.4% and 87.2% of classification accuracy using NB and SVM classifiers respectively for a dataset consisting of 1000 positive and 1000 negative reviews using only the unigrams. They carried out a ten fold cross validation test for reporting this accuracy.

Kim and Hovy (2004) proposed a probabilistic method based on opinion holder identification, opinion region identification, word level sentiment classification and sentence level sentiment classification to assign a class to the review documents. They proposed that if a sentiment region contains more and stronger positive than negative words, then the sentiment will be positive. The algorithm proposed by them first selects the sentences with topic phrase and holder candidates. In the next step, the opinion regions are delimited, followed by the sentence level sentiment classifier, which calculates the polarity of all opinion bearing words individually. Finally, the polarity scores are combined to determine the holder's sentiment for the whole sentence. For the word level sentence classification they used the seed words to start with and with use of WordNet, the expansions were gathered and added to the corresponding seed list. For their experiment, they started with 23 positive verbs and 21 negative verbs along with 15 positive and 19 negative adjectives. After extracting the expansions for each seed word from the WordNet, they were able to obtain 5880 positive adjectives and 6233 negative adjectives along with 2840 positive verbs and 3239 negative verbs. They used a named entity tagger to identify the opinion holder at the sentence level and also they chose the holder that is more close to the opinion phrase in case of more than one identified opinion holder in a sentence. For locating the sentiment region they proposed 4 windows as follows:

- Window 1 – The whole sentence
- Window 2 – The words between holder and topic
- Window 3 – Window 2 ± 2 words
- Window 4 – Window 2 to the end of the sentence

They achieved a maximum of 77.9% of accuracy for their experiments with adjectives, for which they used 231 training data and 231 test data. Also for the experiments with verbs, they achieved a maximum of 81.2% of accuracy, for which, they used 251 training data and 251 test data. They concluded that, presence of word is more important than the sentiment strength and sentiment regions are effective when compared to the analysis on the whole sentence.

Gamon (2004) illustrated the use of large initial feature vectors combined with feature reduction based on the log likelihood ratio for the sentiment classification in the very noisy domain of customer data. For the linguistic features he used “NLPWin”, an NLP based tool developed by Microsoft Research to obtain the phrase structure tree and logical form for each string. He used a linear SVM classifier trained by the Sequential Minimal Optimization tool (SMO). He achieved an accuracy of 77.5% for the dataset containing 40884 very noisy feedback items. He also established that the use of abstract linguistic analysis features consistently contributes to the classification accuracy.

Pang and Lee (2005) proposed a meta-algorithm based on a metric labeling system to assign sentiments using a 3-point or a 4-point scale rather than just classifying them as thumbs up or thumbs down. They succeeded in identifying the author’s evaluation on a multi-point scale. They demonstrated the use of Positive Sentence Percentage (PSP) in a given review to assign the star rating. They created a sentence polarity dataset containing 10662 movie

review snippets for establishing the fact that PSP can be used for computing the document similarity to accomplish the rating inference task.

2.3 SENTIMENT ANALYSIS USING NATURAL LANGUAGE PROCESSING APPROACH

Turney (2002) illustrated a simple unsupervised learning algorithm that uses the semantic orientation of each opinion phrase to decide on the overall sentiment expressed in a document as either positive or negative. He used the PMI-IR algorithm, which uses the Point-wise Mutual Information (PMI) score to measure the similarity of words or phrases. He achieved an average classification accuracy of 74% for reviews of four different domains. The accuracy ranged from 84% for automobile reviews to 66% for movie reviews.

Tong (2001) classified the online movie reviews by developing a lexicon specific to movie reviews, in which he included the opinion phrases that indicate the sentiment of the author towards the movie. He proposed the idea of developing domain specific lexicons to classify the online reviews.

Yi et al (2003) used NLP based sentiment analyzer for assigning sentiments to each of the references corresponding to the given subject. They developed a sentiment lexicon and a sentiment pattern database for this purpose. The feature extraction algorithm used in their research was able to identify the topic related feature terms, which enabled a finer level of sentiment analysis with an accuracy of 87% for review articles with a precision ranging from 86% to 91%. The sentiment pattern database consisted of sentiment extraction patterns. As the experiments were carried out using a shallow parser, they felt that full parsing of the sentences would facilitate a better sentence structure analysis and stressed the need for automating the process of creating

advanced sentiment patterns, which at present requires a considerable amount of human effort to handle the semantics.

Nasukawa et al (2003) used a pattern based approach based on NLP techniques for assigning sentiments at topic level. They demonstrated the advantage of extracting the sentiments associated with polarity of specific subjects rather than classifying a whole document. For the purpose of identifying the sentiment expressions and their semantic relationship with the subject, a Markov model based POS tagger has been used. They reported a precision score of 94.5% and a recall score of 24% for their experiment involving 255 camera reviews.

Hiroshi et al (2004) used a pattern based approach for assigning sentiments at topic level. They equated the task of sentiment analysis to language translation based on which, they proposed a system for extracting the sentiment units from the text. They used the transfer based machine translation engine and replaced the translation patterns and bilingual patterns with sentiment patterns and sentiment polarity lexicon. A precision score of as high as 100% was achieved using their method based on machine translation engine. The sentiment units created by the machine translation engine were less redundant more informative than the naïve predicate-argument structures.

Minqing Hu and Bing Liu (2006) studied the three common types of review formats viz. 1. Pros and Cons 2. Pros and Cons with detailed review 3. Free format. They used the association rule mining to extract the product feature if the review belongs to either format 2 or format 3. They employed a supervised mining method for extracting the product feature if the review is in Pros and Cons format. The polarity of the identified opinion words was determined using the WordNet. The sentence level opinion orientation was decided based on the dominant orientation.

Wilson et al (2005) developed a method for phrase-level sentiment analysis. They first decided whether an expression is neutral or polar and then disambiguated the polarity of the polar expressions. They demonstrated the concept of prior polarity and contextual polarity of positive and negative words. For example, the word “reason” has a positive prior polarity but if the word is preceded by “no”, then the negation happens and the contextual polarity of the word “reason” becomes negative. They used the “BoosTexter AdaBoost” machine learning algorithm with 5000 round of boosting for disambiguating the contextual polarity.

Guang Qiu et al (2010) emphasized the importance of opinion lexicon expansion and opinion target extraction for the purpose effective sentiment analysis. They proposed a propagation algorithm for identifying opinion words and the associated targets. The algorithm starts with a seed word and identifies the associated target and continues to identify the other available opinion words and targets until no more new opinion words or targets can be added. For the purpose of part-of-speech (POS) tagging, they used the Stanford POS tagging tool and for parsing the sentences, they used the “Minipar” sentence parser. They considered opinion words to be adjectives and targets to be nouns or noun phrases. They accomplished the propagation i.e. the identification of opinion words and targets iteratively by carrying out the following tasks:

1. Extracting targets using the opinion words
2. Extracting targets using the extracted targets
3. Extracting opinion words using the extracted targets
4. Extracting opinion words using both the given and extracted opinion words.

This approach resulted in a considerable improvement in precision, recall and F-score over the previously available approaches for opinion words and opinion target identification.

Sheng Gao and Haizhou Li (2011) proposed a Cross Domain Topic Indexing (CDTI) method for using the classifiers trained on one domain (source domain) to classify a document from another domain (target domain). CDTI can be used to find the common topic space among various domains to facilitate knowledge transfer. They demonstrated the use of probabilistic learning algorithm and inference algorithm for carrying out the sentiment analysis in the cross-domain. They proposed four versions of CDTI. They are:

Version A: Dependent CDTI – Source and target terms are dependent

Version B: Term independent CDTI – source and target terms are independent

Version C: Document independent CDTI - source and target documents are independent but source and target terms are dependent

Version D: Both term and document independent CDTI – Source and target terms are independent and also the source and target documents are independent.

They demonstrated that CDTI performs better than the methods like Probabilistic Latent Semantic Analysis (PSLA) and Spectral Feature Alignment (SFA) used for the purpose of cross domain topic indexing.

Danushka et al (2011) proposed a sentiment sensitive thesaurus to facilitate the feature vector expansion to train a binary classifier. They proposed a method for sentiment classification when labeled data for a target domain is not available but some labeled data for multiple other domains

designated as source domains are available. Cross domain sentiment classification has been achieved by effectively learning from multiple source domains. They used both unigrams and bigrams in their research and named them as lexical elements. POS tagging and lemmatization have been used for identifying the lexical elements from both the source domain labeled reviews and unlabeled reviews in source and target domains. After identifying a lexical unit, the co-occurring lexical elements in the same review sentence are selected as features and after that from other source domains wherever that lexical element appears, the co-occurring elements are added to the list of features. In this way, a particular lexical element can be represented as a vector with all possible features that co-occur with it. They established the fact that the expansion of feature vectors using the sentiment sensitive thesaurus bridges the gap between source and target domains in the cross-domain sentiment classification.

Yohan Jo and Alice oh (2011) proposed a new method to classify the online reviews based on the automatic detection of aspects on which the sentiments are expressed and combining the aspects and sentiments together as “senti-aspects” pair. They proposed a sentence latent Dirichlet allocation (SLDA), a probabilistic generative model, which assumes that all words in a single sentence are generated from one aspect. Then SLDA was extended to Aspect and Sentiment Unification Model (ASUM). The advantage of this ASUM is that it does not require labeled reviews. They applied this method on the reviews pertaining to electronic items and restaurants and the classification performance of ASUM was better than the baseline classifiers.

Gamgarn and Pattarachai (2010) used a syntactic information based dependency analysis for identifying the product feature and the opinion. They used the Stanford lexicalized parser generating the syntactic parse trees, which are then converted as dependency trees. From the dependency trees, the

noun phrases and adjective pairs are identified to find the polarity of the sentence. They have defined the dependency path as the shortest path between feature word and opinion word in a dependence graph. The identification of syntactic relationship between the product feature and opinion word along with dependency path resulted in a 0.71, 0.76, and 0.73 precision, recall, and F-score respectively for the experiments conducted using 1250 sentences. They used 80% of the total data for training and 20% of the total data for testing.

Hu and Li (2011) proposed the topical term description model (TTDM) to mine the content structure of topical terms in sentence-level contexts. They separated out the sentences containing at least one topical term and termed them as “sensitive sentences”. They created the maximum spanning trees (MST) for each topical term and its context words. Nodes of the MST represent the opinion words or phrases and the edges between the nodes indicate their association. They used the point-wise mutual information (PMI) to measure the weight of the link between a pair in the MST. The sentence level sentiment classification is decided based on the fact that if a MST is generated by more positive polarity than negative, it has a higher chance of providing the positive perspective and vice versa. Finally the document level sentiment classification is carried out by comparing generation probabilities of MSTs of sensitive sentences using a log-ratio decision function.

Adnan Duric and Fei Song (2011) described a method based on long-range and short range dependencies. They used the Hidden Markov Model- latent Dirichlet allocation (HMM-LDA) topic model. They modeled the review documents based on the combination of syntactic and semantic classes and proposed a method for separating the topical content of the entity

from the opinion context of the entity. They described the following criteria for selecting the salient features in a review:

- The features should be highly expressive and provide useful information to the classification process.
- All the identified features together should provide a comprehensive viewpoint of the whole corpus.
- Features should be domain dependent and frequent enough
- Feature should be discriminative enough

They have not included the aspect identification module in their research but they suggested that the HMM-LDA topic model can be used for extracting the most relevant aspects in a review.

Nikos Engonopoulos et al (2011) proposed a method based on sequential modeling with conditional random fields (CRF) for the entity-level sentiment classification. They achieved a significant improvement in performance on small pieces of text. A Word-level sentiment classifiers based on CRF was used to analyze the entity-level sentiment. They generated a flow of sentiment labels for each document using the CRF model and based on that the sentiment patterns are discovered. They concluded that the sentiment classification of word sequences can be extended to provide the complete sentiment flow over the document being analyzed.

Louis-philippe and Rada (2011) suggested a new method for extracting opinions from the videos uploaded by the customers. The transcribed version of the video content, and factors like smile, look away time, pauses and pitch of the voice were also taken into account while deciding the sentiment expressed. An automatic sentiment classifier built-in this method would first transcribe the video content then by using the image

processing techniques the features like smile and look away time will be gathered along with the audio data like pauses and pitch. The tri-modal approach is a novel idea in the field of sentiment analysis by which, without actually watching the video the sentiment summary can be generated. They used a hidden Markov model (HMM) based classifier for their analysis.

Andrew L.Mass et al (2011) used a vector space model that learns word representations capturing semantic and sentiment information. First, this method produces the representations of the words that occur together in the documents and then the word sentiments are captured. They compared their model with the other popular vector space models like latent semantic analysis (LSA), latent Dirichlet analysis (LDA) using the IMDB movie review dataset containing 50000 movie reviews. They achieved 88.9%, 88.89%, and 88.13% on Pang and Lee dataset, IMDB dataset, and sentence subjectivity dataset respectively.

Andrea Esuli and Fabrizio Sebastiani (2005) proposed a method for determining the orientation of the subjective terms. They used a semi-supervised learning on the term representations obtained from the term glosses. The term glosses were taken from the machine-readable dictionary available freely online. Their method is less data-intensive and less computation-intensive compared to the other methods available for determining the orientation of subjective terms.

Anindya and Panagiotis (2007) proposed a ranking system to predict the usefulness of the reviews. They developed two ranking mechanisms viz. 1. consumer-oriented ranking mechanism 2. manufacture-oriented ranking mechanism. Since the web contains huge number of reviews, identifying the reviews that are very useful for decision making becomes very important. They determined the usefulness of a review by subjectivity analysis. For this purpose, a separate subjectivity classifier was constructed

and a dynamic language model classifier with n-grams was used to train the classifier. They analyzed the effect subjectivity on the product sales and usefulness using a linear specification model.

Agarwal and Bhattacharyya (2005) established the importance of determining the mutual relationship between documents for sentiment analysis. They used a novel graph-cut technique for the final sentiment classification of a document. They used the combination of NLP and machine learning approach for their analysis. For the subjectivity detection the Stanford Log-Linear model tagger was used and the SVM classifier for the classification of reviews. They achieved above 90% accuracy by analyzing the "about" type sentences from the whole review with distance and context information.

Pimwadee Chaovalit and Lina Zhou (2005) compared the supervised and unsupervised classification approaches for the movie review mining. They concluded that the performance of ML approaches depends on the feature selection method employed and suggested that simply applying a n-grams to the classification would result in poor classification. Applying POS tagger to facilitate the feature selection would result in higher performance. Also, they confirmed the fact that the factual information found in a review changes the polarity of the overall sentiment thus resulting in a misclassification. Separating out the subjective sentences from the factual sentences becomes an important task for achieving superior performance in sentiment analysis.

Ana-Maria Popescu and Oren Etzioni (2005) developed an unsupervised information extraction system named "OPINE". They determined the semantic orientation of words based on a new labeling procedure called "relaxation labeling". The probability of an object label is estimated using an update equation in an iterative manner. The relaxation

labeling algorithm stops when the global label assignment stays constant. The semantic orientation (SO) of words was calculated based on the neighbourhood features. They identified the polarity of the opinion phrases based on the SO label assigned to the head words in them.

Maite Taboada et al (2011) proposed a lexicon-based approach for extracting sentiment from text. They developed a semantic orientation calculator (SO-CAL), which uses dictionaries of words with their polarity and strength. As an additional feature, the SO-CAL incorporates intensification and negation methods for each word in its dictionary. Their dictionary consists of separate sections for adjectives, nouns, verbs and adverbs with their SO label ranging from -5 to +5. They have also succeeded in quantifying the modification potential of the intensifiers and negators. With all the features enabled in the SO-CAL, they were able to achieve an overall average performance of 78.37% for reviews from multiple domains like movie, camera etc.

Dmitriy Bespalov et al (2011) demonstrated the effect of n-gram model combined with latent representation on the document classification task. Their model called as supervised n-gram embedding uses a multi layer perceptron to accomplish the embedding. Their model outperformed the standard bag of words model. The only limitation of this method is the requirement of large training set.

Yue Lu (2011) expressed the sentiment score assignment function as linear programming problem subjected to the following constraints:

- Sentiment priors
- Sentiment ratings
- Similar sentiments
- Opposite sentiments

They solved the linear programming problem using GAMS/CPLEX. Using their optimization framework they were able to pick up domain-specific new sentiment words that are not found in any general purpose sentiment lexicon. Also, they were able to identify new sentiments for the same word in the given domain.

Siamak Faridani (2011) proposed a generalized sentiment analysis method based on canonical correlation analysis (CCA). He demonstrated the use of CCA in the context of sentiment analysis when the user rating depends on multiple dimensions like price, service, value etc. The trained CCA model can infer and extract numerical values for each dimension of the text that is being tested and then runs the K Nearest Neighbor (KNN) algorithm to assign the category.

Taras Zagibalov and John Carrol (2008) described an unsupervised sentiment classification of Chinese product reviews based on automatic seed word selection. Though the method is applied for the reviews in Chinese, the proponents of this method has made it as language independent as possible. Since the method is unsupervised, human annotated data has not been used. They identify lexical unit and sentiment zone in every sentence and a score is calculated based on its polarity. The overall sentiment is determined based on the difference between the number of positive and negative zones. If the result is more than zero, the document is classified as positive and vice versa. They tested this approach on a corpus consisting of 29531 reviews and achieved 83% accuracy, which is almost equal to the accuracy of supervised classifiers like SVM, Naïve Bayes multinomial (NBm) etc.

Nitin Jindal and Bing Liu (2006) studied the problem of identifying the comparative sentences in customer reviews. They proposed a technique, which is a combination of class sequential rule mining (CSR) and machine learning. They initially identified 30 keywords and expanded them using the

WordNet. A total of 83 keywords and phrases were identified for the purpose of comparative sentence identification. Using these keywords and phrases, a sequence database consisting of the keyword and POS tag was built and finally the CSRs were generated. CSR basically expresses the probability that a sentence is a comparison if it contains the pattern X. Then a NB classifier has been used to carry out the classification of a sentence as either comparative or non-comparative.

Murthy Ganapathibhotla and Bing Liu (2008) proposed a system for identifying the preferred entities in comparative sentences (PCS). Their objective was to identify the preferred entity of the author in a comparative sentence like “Camera X is superior to Camera Y”. Their research was based on the observation of comparative and superlative words in the English language. They demonstrated the use of “increasing comparatives” like more, longer etc. and “decreasing comparatives” like less, fewer etc. for the identification of preferred feature in a comparative sentence. They worked on the combinations of comparative words(C) and features (F) in the following cases and corresponding rules were developed.

- C is opinionated – In a comparative sentence, if C has a positive orientation then the entity that comes before C is the preferred or otherwise the entity that comes after C.
- C is not opinionated but F is opinionated
- C and F both are not opinionated
- C is a feature indicator

A total of 837 comparative sentences were tested and with 94.4% of accuracy. They proposed a new measure, the one-side association (OSA) measure for determining the level of association between the comparative

word and the entity feature. They established that the OSA measure is superior to the PMI used traditionally.

Derry Tanti Wijaya and Stephane Bressan (2008) proposed a context-dependant ranking procedure to rank items directly from the review text. They generated sentiment graphs using collocation, negative collocation, and coordination by pivot words such as conjunctions and adverbs. They designed the context-dependent ranking of items using the Page Rank algorithm. They also demonstrated that starting the algorithm with a positive adjective and constructing sentiment graph for individual items results in the superior performance with respect to the context-dependent ranking.

Zhu et al (2009) developed an aspect-based sentence segmentation method for summarizing the sentiment in a review. The issue of multiple aspects in a single sentence has been effectively tackled by them using a two-stage approach. In the first stage, the input sentence is segmented into multiple single-aspect segments. In the second stage, a simple algorithm is used to re-segment the single-aspect segments generated during the first stage based on the polarity changes. A sentiment-lexicon-based method has been employed for the polarity analysis.

Xiaowen Ding et al (2009) proposed a method for “entity discovery” (identifying the entity if the name is available in the review) using sequential pattern mining and “entity assignment” (when pronouns are used to represent the entity) using comparative sentences mining. This method is quite useful when a reviewer is commenting on multiple products in the same review.

Wei Jin et al (2009) proposed a robust machine learning algorithm based on the framework of lexicalized-HMM. Their model completely depends on machine learning and integrates multiple linguistic

features like part-of-speech, phrases' internal formation patterns, surrounding contextual clues into automatic learning. Their system could predict new potential product and opinion entities based on the patterns it has learnt. They suggested that pronoun resolution should be carried out on only selected sentences to reduce the false positive rate.

Ramanathan Narayanan et al (2009) proposed a supervised learning model for analyzing the sentiment expressed in conditional sentences. Conditional sentences are unique in structure and they need special handling. They studied the canonical tense patterns and built an SVM classifier to automatically predict the sentiment expressed in a conditional sentence.

Bing Liu (2010) has consolidated the research in the field of sentiment analysis and given the directions for the future research in this area. He argued that the research community likes to focus on the individual sub-problems and that is the reason for the current solutions to be far from perfect. In the conclusion, he urges the importance of conducting more refined and in-depth investigations to build integrated sentiment analysis systems. Also he has suggested that opinion spam identification is the promising area for future research.

Zhongwu et al (2011) proposed a method for clustering the product features using a semi-supervised approach. Lexical similarity was used to identify the initial labeled examples for training. The clustering has been achieved using the proposed L-EM (Labeled Expectation Maximization) algorithm.

Zhang Lei and Bing Liu (2011) proposed a method to identify the nouns product features that are useful for detecting the opinions. They studied the problem of objective nouns and the sentences with implied opinions. They

suggested that, the recall measure of the classification task can be improved by analyzing the noun features. They demonstrated that in addition to the opinion words, the surrounding context can also be used to determine the feature polarity.

Zhang Lei et al (2010) proposed a method overcome the limitations of the double propagation method for feature identification in large and small corpora. The double propagation method results in low precision and low recall for large and small corpora. They have tackled the low recall issue by introducing “part-whole pattern” and “no pattern”. The low precision problem is overcome by using feature relevance and feature ranking. They formulated this problem as a bipartite graph and used the web page rank algorithm to find the important features and also to rank them high.

Zhu Jian et al (2010), proposed a new method based on artificial neural networks (ANN) named as “i-model”. The identified sentiment features are sorted based on a score and feature weight is calculated. The ANN is trained using the prior knowledge base. Their method exhibited a superior performance compared to SVM or HMM classifiers on the Cornell movie review dataset v 2.0.

Qingliang Miao et al (2009) proposed a method for sentiment mining using a ranking factor “temporal opinion quality” (TOQ) developed by them. They expressed every sentence in a review as a tuple including four elements: [title, help, date, R-Content]. The sentences are ranked based on TOQ and Lucene Rank (LR). The final rank depends on both TOQ and LR. They were able to achieve a consistently above 83% of recall and precision.

Huifeng Tang (2009) has thoroughly reviewed the research in the field of sentiment analysis and classified it into four major areas

viz. 1. Subjectivity classification 2. Word sentiment classification 3. Document sentiment classification 4. Opinion extraction. They have identified the interdependency of these areas in their analysis. They have identified the following areas for future research:

- Resolving the problems associated with the reviews containing mixed sentiments i.e. both positive and negative sentiments.
- Aspect level sentiment analysis at a finer level
- Development of Multi-domain sentiment lexicon, which will be useful for multi-class sentiment classification.

2.4 SENTIMENT ANALYSIS USING HYBRID APPROACHES

Konig and Brill (2006) proposed a hybrid model to reduce the human effort required to infer classification rules and labeling of the reviews. In their work, the unlabeled data is first analyzed by a sentiment pattern matching system for a possible labeling if a nearly matching pattern is found, if not the unlabeled data is subsequently analyzed by an ML based classifier to assign the class labels. For the purpose of validating the proposed method, using the pattern discovery algorithm, they identified 300 most frequent text patterns. The first phase of classification was carried out with these text patterns, then during the second phase the unassigned reviews were given to SVM baseline classifier in the form of feature vectors. The classification performance of this hybrid method was 91%.

Rudy Prabowo et al (2009) described a hybrid approach for the sentiment analysis. They established the fact that the use of multiple classifiers sequentially in a hybrid manner can result in better effectiveness

than any individual classifier. They used the following classifiers in some sequence in their experiments:

1. Rule based classifiers (RBC) – Parser based
2. General inquirer based classifier (GIBC) –based on the general inquirer lexicon containing 3672 pre-classified words
3. Statistics based classifier (SBC) – Based on the rule set built with an assumption that bad or good expressions co-occur more frequently with a set of sentiment bearing words. Closeness between the antecedent and the sentiment words decide the polarity of the expression.
4. Support vector machines (SVM)

Their experiments were carried out on four different datasets and the maximum classification accuracy of 90% was for RBC→SBC→GIBC→SVM hybrid classifier.

Albornoz (2010) proposed a hybrid approach to determine the polarity and sentiment intensity using a combination of NLP and ML methods. They transformed a given sentence into a vector called as vector of emotional occurrences (VEO) and that is fed as input to the ML algorithms like J48, SVM etc. The creation of VEO is done in four steps viz. 1. POS tagging and concept identification 2. Emotion identification 3. Negation and quantifier detection 4. Sentence classification.

2.5 SENTIMENT ANALYSIS USING FUZZY LOGIC

Samaneh Nadali et al (2010) illustrated the application of fuzzy logic for categorizing the customer review documents into five categories

viz. very strong negative (or positive), strong negative(or positive), moderate negative (or positive), very weakly negative(or positive), weak negative(or positive) by using 17 unique rules based on patterns of the commonly found opinion phrases. Opinion words were identified as the combination of adjective, adverb, and verb. They used triangular membership function and Mamdani's defuzzification model to get the final crisp value.

Animesh Kar and Deba Prasad Mandal (2011) proposed a Fuzzy Opinion Miner (FOM), a supervised opinion orientation detection system to rank the products based on the FOM score. Their work is a combination of NLP based sentiment mining and fuzzy logic. After identifying the opinion words and phrases using a standard POS tagger, the fuzzy measures were calculated for the following three distinct cases:

- Adverb-Adjective phrases (Good, Very good etc.)
- "Not/Never" followed by adjective phrases (Not good, Not bad etc.)
- "Not" followed by adverb-adjective phrases (Not very good, Not very bad etc.)

The final ranking is based on the FOM score determined from the fuzzy scores of the identified opinion words and opinion phrases in the reviews.

Slawomir Zadrozny and Janusz Kacprzyk (2006) proposed a fuzzy logic based automatic text document categorization. Their approach is based on the classical calculus of linguistically quantified propositions. They developed a Rocchio type classifier for text document categorization. They

proposed three thresholding strategies based on which the category assignment of a text document is carried out. The thresholding strategies are:

- Rank-based thresholding
- Proportion based thresholding
- Score based local optimization

They used the Reuters corpus for validating the performance of the proposed method and achieved a precision score of 0.831.

Upasana et al (2011) proposed a system for document classification using lexical chaining and fuzzy logic. Keyword extraction is done based on the WordNet lexical database and the Wikipedia has been used to expand the keywords by adding the first level hyperlinks. The triangular membership function has been used to assign fuzzy weight for each keyword or phrase. Two types of lexical chains are created based on the reference source using the neighborhood tokens. Each chain is assigned a strength measure, which can be used for document classification.

Guohong Fu and Xin Wang (2010) proposed a method for sentence level sentiment classification based on fuzzy set theory. They demonstrated the use of fuzzy set theory in modeling the intrinsic fuzziness between sentiment polarity classes. They defined three fuzzy sets namely positive, negative, and neutral to represent the various polarity classes. They used the semi-trapezoid membership function for all the three fuzzy sets. The sentence-level polarity is determined by using the maximum membership principle.

2.6 MAHALANOBIS DISTANCE FOR TEXT CATEGORIZATION

Suli Zhang and Xin Pan (2011) used MD for text classification and proposed a new algorithm, which is a variant of the k-nearest neighbor (KNN) algorithm and categorized news topics into sports, medicine, politics etc. This is the first available paper on text categorization using MD.

Elizabeth A.Cudney et al (2007) compared the performance of Mahalanobis-Taguchi system (MTS) and neural network for multivariate pattern recognition. In the MTS approach, MD is used to measure the degree of abnormality of patterns. Since MD is based on the correlation among the variables, it becomes a handy choice for pattern recognition. They demonstrated that, MD based pattern recognition is superior and economical compared to the neural network, which requires large amount of training data for enhanced classification accuracy.

2.7 SCOPE AND OBJECTIVE

2.7.1 Scope

From the literature survey, it is evident that the field of sentiment analysis is evolving continuously as it has immense potential for business intelligence related applications. Though a lot of sentiment analysis methods have been proposed and practiced using the supervised, unsupervised and semi-supervised learning and classification algorithms, still there is a huge potential for developing new methods for sentiment analysis at the document level. In this thesis, some new methods based on MD and fuzzy logic has been proposed for the binary classification of sentiment expressed in a review at the document level.

2.7.2 Objective

The objective of the present work is to develop a classifier to detect the sentiment expressed in a document as either positive or negative using MD. Also, to compare the performance of the MD-based classifier (MDC) with fuzzy logic based sentiment classifier and the other popular machine-learning classification systems.

2.8 METHODOLOGY

Step 1: Creation of "Representative Term-Document Matrix" (RTDM), a new way of representing a text document.

Step 2: Design and development of MDC for sentiment detection.

Step 3: Design and development of fuzzy logic based classifier for sentiment detection.

Step 4: Design and development of Mahalanobis-fuzzy system (MFSC) classifier for sentiment detection.

Step 5: Design and development of hybrid classifier using MDC, MFSC and FLC with various machine learning algorithms.

Step 6: Performance comparison of MDC, MFS classifiers with the fuzzy classifier and the other machine-learning tools like Naïve Bayes, Bayesian Logistic Regression (BLR), Classification and regression tree (CART), Multi Layer Perceptron (MLP) and Sequential Minimal Optimization (SMO) algorithms.

Figure 2.1 shows the flow activities in this study.

2.8.1 The Flow of Activities in this study

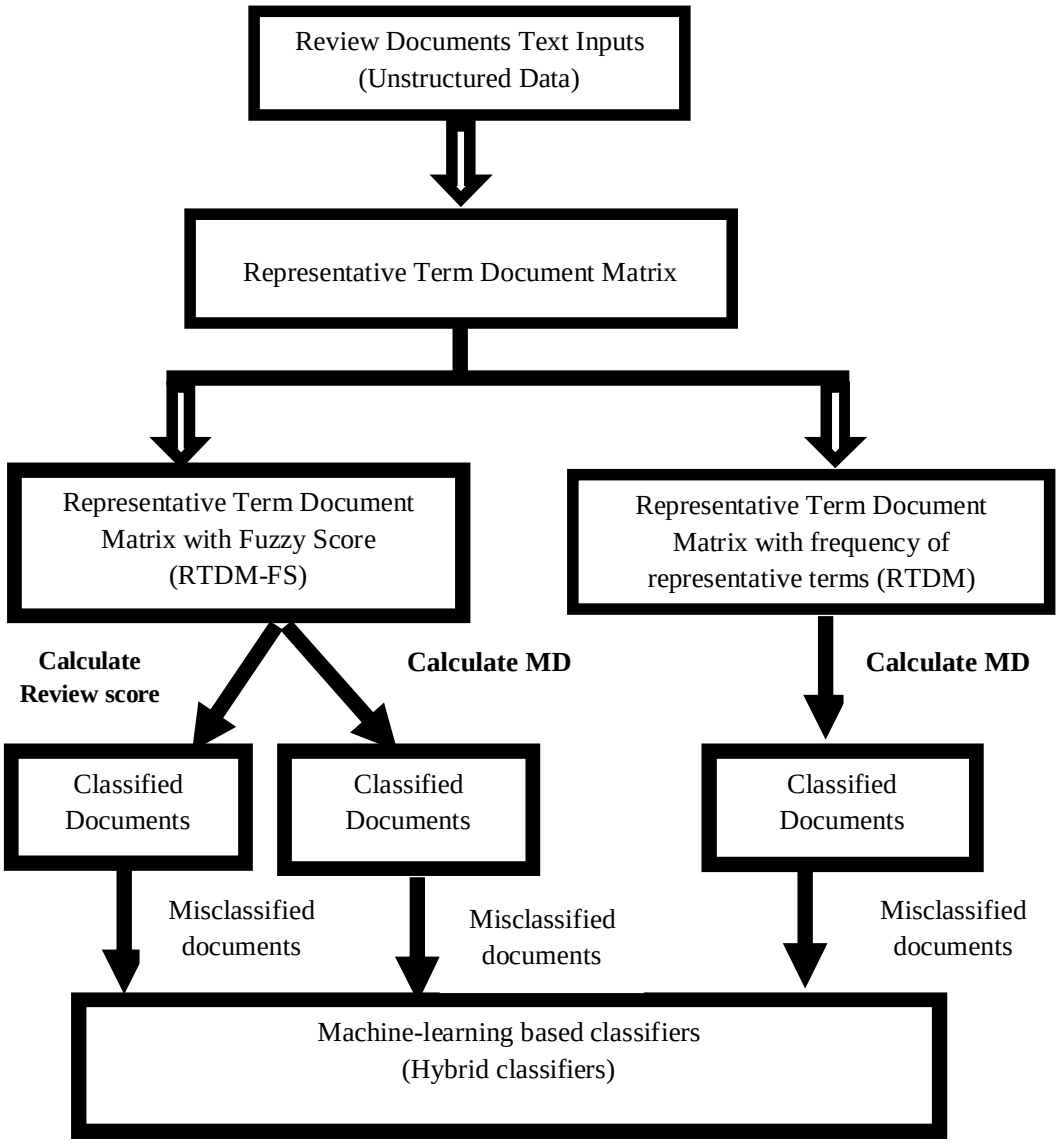

Figure 2.1 The flow of activities in this study