A SURVEY ON SENTIMENT ANALYSIS AND OPINION MINING

Raisa Varghese[1], Jayasree M[2]

[1]*PG Scholar, Govt. Engineering College, Thrissur, Kerala, India*
[2]*Asst. Professor, Govt. Engineering College, Thrissur, Kerala, India*

Abstract

Sentiment analysis is a machine learning approach in which machines analyze and classify the human's sentiments, emotions, opinions etc about some topic which are expressed in the form of either text or speech. The textual data available in the web is increasing day by day. In order to enhance the sales of a product and to improve the customer satisfaction, most of the on-line shopping sites provide the opportunity to customers to write reviews about products. These reviews are large in number and to mine the overall sentiment or opinion polarity from all of them, sentiment analysis can be used. Manual analysis of such large number of reviews is practically impossible. Therefore automated approach of a machine has significant role in solving this hard problem. The major challenge of the area of Sentiment analysis and Opinion mining lies in identifying the emotions expressed in these texts. This literature survey is done to study the sentiment analysis problem in-depth and to familiarize with other works done on the subject.

Index Terms: *Sentiment Analysis, Opinion Mining, Cross Domain Sentiment Analysis*

----------***----------

1. INTRODUCTION

Sentiment analysis and opinion mining are subfields of machine learning. They are very important in the current scenario because, lots of user opinionated texts are available in the web now. This is a hard problem to be solved because natural language is highly unstructured in nature. The interpretation of the meaning of a particular sentence by a machine is tiresome. But the usefulness of the sentiment analysis is increasing day by day. Machines must be made reliable and efficient in its ability to interpret and understand human emotions and feelings. Sentiment analysis and opinion mining are approaches to implement the same.

The sentiment analysis problem can be solved to a satisfactory level by manual training. But a fully automated system for sentiment analysis which needs no manual intervention has not been introduced yet. This is mainly because of the challenges in this field. This paper aims at a literature survey on the problem of sentiment analysis and opinion mining. Many relevant studies have emerged in this field and this paper is a peep into some of them.

2. DIFFERENT LEVELS OF SENTIMENT ANALYSIS

2.1. Document level sentiment analysis

The basic information unit is a single document of opinionated text. In this document level classification, a single review about a single topic is considered. But in the case of forums or blogs, comparative sentences may appear. Customers may compare one product with another that has similar characteristics and hence document level analysis is not desirable in forums and blogs. The challenge in the document level classification is that all the sentence in a document may not be relevant in expressing the opinion about an entity. Therefore subjectivity/objectivity classification is very important in this type of classification. The irrelevant sentences must be eliminated from the processing works.

Both supervised and unsupervised learning methods can be used for the document level classification. Any supervised learning algorithm like naïve Bayesian, Support Vector Machine, can be used to train the system. For training and testing data, the reviewer rating (in the form of 1-5 stars), can be used. The features that can be used for the machine learning are term frequency, adjectives from Part of speech tagging, Opinion words and phrases, negations, dependencies etc. Labeling the polarities of the document manually is time consuming and hence the user rating available can be made use of. The unsupervised learning can be done by extracting the opinion words inside a document. The point-wise mutual information can be made use of to find the semantics of the extracted words. Thus the document level sentiment classification has its own advantages and disadvantages. Advantage is that we get an overall polarity of opinion text about a particular entity from a document. Disadvantage is that the different emotions about different features of an entity could not be extracted separately.

2.2. Sentence level sentiment analysis

In the sentence level sentiment analysis, the polarity of each sentence is calculated. The same document level classification methods can be applied to the sentence level classification problem. Objective and subjective sentences must be found out. The subjective sentences contain opinion words which help in determining the sentiment about the entity. After which the polarity classification is done into positive and negative classes. In case of simple sentences, a

single sentence bears a single opinion about an entity. But there will be complex sentences also in the opinionated text. In such cases, sentence level sentiment classification is not desirable. Knowing that a sentence is positive or negative is of lesser use than knowing the polarity of a particular feature of a product. The advantage of sentence level analysis lies in the subjectivity/ objectivity classification. The traditional algorithms can be used for the training processes.

2.3. Phrase level sentiment analysis

The phrase level sentiment classification is a much more pinpointed approach to opinion mining. The phrases that contain opinion words are found out and a phrase level classification is done. This can be advantageous or disadvantageous. In some cases, the exact opinion about an entity can be correctly extracted.

But in some other cases, where contextual polarity also matters, the result may not be fully accurate. Negation of words can occur locally. In such cases, this level of sentiment analysis suffices. But if there are sentences with negating words which are far apart from the opinion words, phrase level analysis is not desirable. Also long range dependencies are not considered here. The words that appear very near to each other are considered to be in a phrase.

3. SUBJECTIVITY/ OBJECTIVITY CLASSIFICATION

Subjectivity/Objectivity classification is a challenge that should be addressed along with sentiment analysis problem. The text pieces may or may not contain useful opinions or comments. The subjective sentences are the relevant texts, and the objective sentences are the irrelevant texts. So we must sort out the sentences that are useful for us and those which are not. The subjective sentences are those sentences having useful information for the sentiment analysis. Such classification is termed as subjectivity classification. Some works have been done focusing on this particular problem.

In [1], the authors present a method of subjectivity identification for sentiment analysis. This is important because the irrelevant data from the reviews could be eliminated. This eliminates the processing overheads of a large amount of textual data. The method they propose is using minimum cuts to produce subjective extracts from the text. The work has been focused in the sentence level subjectivity extraction.

A classification approach using Naive Bayesian classifier is used in [2]. They present the results of developing subjectivity classifiers using un-annotated texts for training. In this work of learning Subjective and Objective sentences, the method automatically generates training data. This is done by a Rule-based approach. The rule-based subjective classifier classifies a sentence as subjective if it contains two or more strong subjective clues. In contrast, the rule-based objective classifier looks for the absence of clues: it classifies a sentence as objective if there are no strong subjective clues in the current sentence, there is at most one strong subjective clue in the previous and next sentence combined, and at most 2 weak subjective clues in the current, previous, and next sentence combined classifiers. They use Subjective Precision, Subjective Recall, Subjective F measure, Objective Precision, Objective Recall and Objective F measure for the evaluation. They also implement a self training procedure for the system.

4. MAJOR CHALLENGES INVOLVED IN SENTIMENT ANALYSIS

There are several challenges that are to be faced to implement sentiment analysis. Some of them are listed below.

4.1. Named Entity Extraction

Named entities are definite noun phrases that refer to specific types of individuals, such as organizations, persons, dates, and so on. The goal of named entity extraction is to identify all textual mentions of the named entities in a text piece. Named entity recognition is a task that is well suited to the type of classifier-based approach like sentiment analysis. Consider the following example,

EXAMPLE 1: (i) The Canon Power Shot is a great camera for beginners. (ii) It is easy to use and it is very good quality. (iii) The graphics are great and it takes the picture quickly. (iv) It has a wonderful face identification feature which makes the picture even better than it was before. (v) After you take the picture you can also do a red eye correction! (vi) Audio is pretty good but the HD quality is less than desirable.

Here the mention about the brand of camera, 'Canon Power shot' is a named entity. For effective sentiment analysis such mentions should be sorted out.

4.2. Information Extraction

Information comes in many shapes and sizes. The complexity of natural language can make it very difficult to access the information in the opinion text.

The tools in NLP are still not fully capable to build general-purpose representations of meaning from unrestricted text. Regarding information available, one important form is structured data, where there is a regular and predictable organization of entities and relationships. Another is unstructured data which can be found in the Internet in large volume. Information Extraction has many applications, including business intelligence, media analysis, sentiment detection, patent search, and email scanning. In the sentiment analysis application, the information that is to be extracted are the opinions and the corresponding polarity values.

4.3. Sentiment Determination

The sentiment determination is a task that assigns a sentiment polarity to a word, a sentence or a document. A

traditional way for sentiment polarity assignment is to use the sentiment lexicon. The adjectives of a sentence are given importance in opinion mining because they have more probability to carry information while sentiment analysis problem is considered. The presence of any of the words in the opinion lexicon can be helpful while finding the sentiment polarity. There are approaches like dictionary based approach and Corpus based approaches to build up the opinion lexicon.

4.4. Co-reference Resolution

Co-reference resolution is to be done in aspect level and entity level. In the case of opinionated text, we can see comparative texts. These comparative texts may contain co-references. These references must be effectively resolved for producing correct results. For example, consider the following opinionated text,

EXAMPLE 2: Comparing Nikon's Coolpix to its main competitor the Canon, it takes excellent photos and is quite compact.

Here two named entities are mentioned and they are Nikon and Canon. The pronoun 'it' in the text refers to 'Nikon's Coolpix'. When the co-referring words are not found out, effective sentiment analysis cannot be carried out. The importance of co-reference resolution lies in the fact that it helps in providing more information in the Information retrieval tasks. There are several anaphora resolution factors that help in the task. Constraints and preferences are considered while carrying out this task. The scope of the resolution task is also to be defined. The scope can be a sentences, nearby sentences or a document etc. The co-reference resolution is important to the sentiment analysis problem and very complex task in itself. The resolution problem itself is not solved yet in NLP.

4.5. Relation Extraction

Relation extraction is the task of finding the syntactic relation between words in a sentence. The semantics of a sentence can be found out by extracting relations between words and this can be done by knowing the word dependencies. This is also a major research area in NLP and serious researches are going on to solve this problem. Textual analysis like POS tagging, shallow parsing, dependency parsing is a pre-requisite for relation extraction. These steps are prone to errors. Many of the problems in NLP are not fully solved because of the unstructured nature of text. Relation extraction also belongs to the group of challenging problems. The place of relation extraction in sentiment analysis is very high and thus this challenge is to be met and solved.

4.6. Domain Dependency

A sentiment classifier that is trained to classify opinion polarities in a domain may produce miserable results when the same classifier is used in another domain. Sentiment is expressed differently in different domains. For instance, consider two domains, digital camera and car. The way in which customers express their thoughts, views and prospective about digital camera will be different from those of cars. But some similarities may also be present. So Sentiment analysis is a problem which has high domain dependency. Therefore cross domain sentiment analysis is a challenging problem that has to be unfolded.

5. OPINION MINING AND SENTIMENT ANALYSIS

The sentiment analysis problem is met using some of the techniques using natural language processing technique, proximity method etc. Following are a brief study on a few of them.

A notable approach in [3] uses a sentence level sentiment analysis. The word level feature extraction is done using Naive Bayesian Classifier. The semantic orientation of the individual sentences is retrieved from the contextual information. This machine learning approach on average claims an accuracy rate of 83%. For classifying and analyzing of the sentiment from the reviews, machine learning and lexical contextual information are used. The paper focuses on sentence level to check whether the sentences are objective or subjective and to classify the polarity of the sentences to positive or negative opinion.

The naive bayes approach is used to annotate each sentence as positive and negative on the bases of useful word level feature. SVM classifier is trained on the annotated sentences for the positive and negative classification. Contextual information is used to calculate the polarity of sentence and mark it as either negative or positive. The paper[4] presents experiments for sentiment analysis to automatically distinguish prior and contextual polarity. Beginning with a large stable of clues marked with prior polarity, method identifies the contextual polarity of the phrases that contain instances of those clues in the corpus.

A two-step process is used that employs machine learning and a variety of features. Firstly the method classifies each phrase containing a clue as neutral or polar. Secondly it takes all phrases marked in previous step as polar and disambiguates their contextual polarity (positive, negative, both, or neutral). The method describes a system that automatically identifies the contextual polarity for a large subset of sentiment expressions, achieving reliable results. Another significant work is the implementation of both Natural Language understanding and Generation in Sentiment analysis [5]. A couple of algorithms to search and predict the orientation of opinions are specified in this research work. In their system there is a review database that stores the opinionated texts. The method then finds frequent features that many people have expressed their opinions on. After that, the opinion words are extracted using the resulting frequent features, and semantic orientations of the opinion words are identified with the help of WordNet. The system then finds those infrequent features.

The orientation of each opinion sentence is identified and a final text summary is generated in this work. The part of

speech tagging from natural language processing is used to find opinion features. The output of the above paper is a text summary of opinions. Thus Summarization of text is also done as a subsystem. But this summarization work is truly dependent on the features and hence is far from the automatic summarization work in the field of NLP. The paper proposes a method by utilizing the adjective synonym set and antonym set in WordNet to predict the semantic orientations of adjectives. The paper also describes the need of pronoun resolution in opinion mining even though it is not addressed.

A method of sentiment analysis which does not use conventional natural language rules is specified in [6]. The work uses a machine learning approach (Naive Bayesian)for classification. The class association rules is used to extract the associations between term features appearing in consumer review opinions and product features for a particular consumer product.

A set of pre-classified opinion sentences is utilized as training data to develop class association rules. Each sentence is labeled with one or more product features, fj , or no product feature, none. The f-measure is used as metric for evaluation, and claims efficiency up to 70%. In the paper, the review sentences are divided into various classes according to the association rules. The classification of the opinionated text is done using both class association rules and naive Bayesian classifier. After which the experiments done proves that Class association rules perform better than the traditional naïve Bayesian classifiers. In [7], the authors present an approach for opinion mining which relies on natural language processing techniques. The work is accomplished by the sentiment lexicon and a pattern database. The two feature selection algorithms discussed in this work are based on mixture model and the likelihood ratio. They propose a sentiment pattern based analysis for the sentiment classification work.

In [8], an in-depth study of dependency relations among the words of a sentence is discussed. In their work, the dependencies are classified as short range and long range dependencies. They use a clustering approach after the parsing is done. In the paper [9] a combined model of sentiment analysis is done. Considering every levels of analysis like phrase level, sentence level and document level have their own advantages. But a combination model including all the three may achieve better performance.

A combined model based on phrase and sentence level analyses and a description on the implementation of different levels of analyses are presented. For the phrase-level sentiment analysis, a template is used. The newly defined template is Left-Middle-Right template. The Conditional Random Fields are used to extract the sentiment words. The Maximum Entropy model is used in the sentence-level sentiment analysis. The combination model with specific combination of features performs slightly better than the traditional single level models. Another paper which studies the mining of on-line reviews in the movie domain is [10]. In the paper they come up with a proposal of a model called S-PLSA(Sentiment Probabilistic Latent Semantic Analysis). This is a generative model for sentiment analysis that does a deeper comprehension of the sentiments in blogs.

The model S-PLSA is used for summarizing sentiment information from reviews. From the S-PLSA model, they developed ARSA(Autoregressive Sentiment-Aware model), a model for predicting sales performance based on the sentiment information and the product's past sales performance. They have considered the role of review quality in sales performance prediction. The model predicts the quality rating of a review. The quality factor is then incorporated into a another model called ARSQA (Autoregressive Sentiment and Quality Aware model). Two models, ARSA and ARSQA models are designed for product sales prediction. These models reflect the effect of sentiments, and past sales performance on future sales performance. Sentiment analysis problem is attempted to be solved using a clustering approach in [11]. This paper also discusses application of TF-IDF weighting method, voting mechanism and importing term scores and claims almost stable results. A feature level Sentiment analysis is discussed in [12]. Here the work has been concentrated on Chinese product reviews.

The feature selection process is based on an apriori algorithm. The Apriori association mining rules is used to extract the candidate product features. Then the orders of some candidate product feature words are adjusted. Finally, point-wise mutual information (PMI) methods are used to filter feature words so as to obtain the meaningful product feature words. The work is very simple and not upto satisfaction. But the feature extraction done in this work is mentionable. A very distinguishable approach to opinion mining is put forward in [13]. The model is based on nouns and adverb-adjective-noun (AAN) combinations in sentiment analysis .

The AAN based sentiment analysis technique deploys linguistic analysis of adverbs of degree , domain specific adjective and abstract noun. A set of general axioms (based on a classification of adverbs of degree into five categories, classification of adjective into ten specific domain, classification of abstract noun in two categories) for opinion analysis is also defined. The way in which the adjectives and adverbs are found and scored is interesting. Unary and binary AAN algorithms are also mentioned in the work. Another new approach is a proximity based sentiment analysis [14].

The idea is based on the findings about the way in which humans express their thoughts. When a person starts writing positively about a topic or subject they continue with this positive trend for a period of time. Later inflexion words like "however" are used and then start writing in negative sense about the topic. In a paragraph people usually do not repeatedly write one positive and one negative word together. Typically segments of a written text (e.g.

paragraphs or sentences) capture a concept or trend of thought over a short period of time. Such trends could fluctuate as one moves along the written document. The average distance between positive-oriented (or negative-oriented) words is expected to be small for segments bearing positive (negative) sentiments. Consequently, the average distance between positive-oriented (negative-oriented) words is relatively large for segments bearing negative (positive) sentiments. This is the principle on which the model is developed. Three different proximity-based features, proximity distributions, mutual information between proximity types, and proximity patterns are used for sentiment analysis.

Support Vector Machine Classifier is made use of in [15]. The approach emphasizes the use of a variety of diverse information sources, and SVMs provide the ideal tool to bring these sources together. The methods are used to assign values to selected words and phrases, and bring them together to create a model for the classification of texts. In this paper, The sentiment orientation of a phrase is determined based upon the phrase's point wise mutual information (PMI) with the words like excellent and poor. Semantic values of phrases and words within a text are used to add to features for SVM training. Combinations of SVMs using these features in conjunction with SVMs based on uni-grams and lemmatized uni-grams is a diverse method from ours.

In [16], reviews are classified into positive and negative ones. Traditionally the document classification was performed on the topic basis. The three machine learning methods Naive Bayes, maximum entropy classification, and support vector machine are used for sentiment analysis. The traditional ways of document classification based on topic is tried out for sentiment analysis. They consider positive and negative as two topics and classify the reviews according to that. The work concludes that mere usage of the same technique of topic based classification in the sentiment domain fails. Therefore more sophisticated techniques should be used in solving the sentiment analysis problem. The paper [17] describes the use of Passive-Aggressive (PA) Algorithm Based Classifier. The Passive Aggressive algorithms are a family of margin based on-line learning algorithms for binary classification. PA algorithms work similarly to support vector machines (SVM). PA algorithms try to find a hyper plane that separates the instances into two half-spaces. The margin of an example is proportional to the example's distance to the hyper plane. When making errors in predicting examples, PA algorithm utilizes the margin to modify the current classifier. They update the classifier by the constraints.

Another classifier compared with is Language modeling (LM) Based classifier. Language modeling (LM) is a generative method that calculates the probability of generating a given word sequence, or string. The third classifier is the Winnow classifier. Winnow is an on-line learning algorithm for sentiment classification. Winnow learns a linear classifier from bag-of-words of documents to predict the polarity of a review. Instead of uni-grams or bi-grams, n-grams (6-grams) are used as features in their model. The major observation from this paper is the use of high order n-grams as features. In the paper[18], a sentiment analysis approach to extract sentiments associated with polarities of positive or negative for specific subjects from a document is done. This is in contrast of classifying the whole document into positive or negative. In order to identify sentiment expressions and to analyze their semantic relationships with the subject term, natural language processing plays an important role. The method identifies the subjects in the opinion sentences and associate opinions to these subjects.

6. CROSS DOMAIN SENTIMENT ANALYSIS

Cross domain sentiment analysis is introduced to reduce the manual effort in training the machine using labeled data. Instead the machine learns from a particular domain and analyse the sentiment polarities of texts in another domain. This is a very challenging problem because the kind of words used to express emotions in two different domains may be very different. A paper [19] approaches this topic vastly covering all the difficulties evolved in the problem. A sentiment sensitive distributional thesaurus is created using labeled data for the source domains and unlabelled data for both source and target domains. Sentiment sensitivity is achieved in the thesaurus by incorporating document level sentiment labels in the context vectors used as the basis for measuring the distributional similarity between words. The created thesaurus is used to expand feature vectors during train and test times in a binary classifier.

6. CONCLUSION

Sentiment Analysis problem is a machine learning problem that has been a research interest for recent years. Through this literature survey, the relevant works done to solve this problem could be studied. Although several notable works have come in this field, a fully automated and highly efficient system has not been introduced till now. This is because of the unstructured nature of natural language. The vocabulary of natural language is very large that things become even hard. Several challenges still exist in the field of machine learning and some of them are Named entity Recognition, Coreference Resolution, domain dependency etc. These problems have to be tackled separately and those solutions can be used to improve the methods to do sentiment analysis.

REFERENCES

[1] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, ACL, 2004.

[2] J. Wiebe and E. Riloff, "Creating subjective and objective sentence classifiers from unannotated texts," in Computational Linguistics and Intelligent Text Processing. Springer, 2005, pp. 486–497.

[3] B. B. Khairullah Khan, Aurangzeb Khan, "Sentence based sentiment classification from online customer reviews," ACM, 2010.
[4] P. H. Theresa Wilson, Janyce Wiebe, "Proceedings of human language technology conference and conference on empirical methods in natural language processing," Association for Computational Linguistics, p. 347354, 2005.
[5] M. Hu and B. Liu, "Mining and summarizing customer review," KDD04, ACM, 2004.
[6] C.-p. W. C.C.Yang, Y.C. Wong, "Classifying web review opinions for consumer product analysis," ICEC09, ACM, 2009.
[7] R. B. W. N. Jeonghee Yi, T Nasukawa, "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques," ICDM03, IEEE, 2003.
[8] P. B. Subhabrata Mukherjee, "Feature specific sentiment analysis for product reviews."
[9] W. X. G. C. Si Li, Hao Zhang and J. Guo, "Exploiting combined multi-level model for document sentiment analysis," International Conference on Pattern Recognition IEEE, 2010.
[10] J. X. H. A. A. Xiaohui Yu, Yang Liu, "Mining online reviews for predicting sales performance: A case study in the movie domain," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE, vol. 24,APRIL 2012
[11] F. L. Gang Li, "A clustering-based approach on sentiment analysis," IEEE, 2010.
[12] M. X. Y. S. Haiping Zhang, Zhengang Yu, "Feature-level sentiment analysis for chinese product reviews," IEEE, 2011.
[13] T. K. M. J.K. Sing, Souvik Sarkar, "Development of a novel algorithm for sentiment analysis based on adverb-adjective-noun combinations," IEEE, 2012.
[14] D. A. A. S.M.Shamimul Hasan, "Proximity-based sentiment analysis," IEEE, 2011.
[15] T. Mullen and N. Collier, "Sentiment analysis using support vector machines with diverse information sources," EMNLP, vol. 4, pp. 412–418, 2004.
[16] S. V. Bo Pang, Lillian Lee, "Thumbs up? sentiment classifcation using machine learning techniques," Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), ACL, pp. 79–86, July 2002.
[17] M. D. Hang Cui, Vibhu Mittal, "Comparative experiments on sentiment classication for online product reviews," American Association for Articial Intelligence, 2006.
[18] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in Proceedings of the 2nd international conference on Knowledge capture. ACM, 2003, pp. 70–77.
[19] I. D. W. Danushka Bollegala, Member and J. Carroll, "Cross-domain sentiment classification using a sentiment sensitive thesaurus," IEEE, 2012.