**Literature Review**

**1.Introduction**

Customers opinions can be crucial when it’s time to make a decision in a company for improving the services provided by them. Until recently, the main sources of information were by speaking with customers or as written feedback. Now, the IT field provides new tools to efficiently extract customer feedback and reviews upon the services they receive from a particular company. But the increase in the amount of feedback receives make it impossible to process them manually to identify the opinion of the customers. So as a solution to this problem resulting in emerging fields are opinion mining and sentiment analysis.

The basic task of opinion mining is polarity classification. Polarity classification occurs when a piece of text stating an opinion on a single issue is classified as a positive or negative sentiment. But this basic concept doesn’t give much information from the data in hand. To get the most out of the data one of the emerging trends is aspect-based sentiment analysis. It is a text analysis technique that breaks down the text into aspects(components of a service or a product) and allocates each aspect a sentiment level. This technique can help a business to focus on each aspect of their service/product separately based on customer feedback.

**2. Sentiment Analysis**

**2.1 Different Levels of Analysis:**

* Document-level: The task at this level is to classify whether a whole opinion document expresses a positive or negative sentiment
* Sentence level: The task at this level goes to the sentences and determines whether each sentence expressed a positive, negative, or neutral opinion.
* Entity and Aspect level: Instead of looking at language constructs (documents, paragraphs, sentences, clauses or phrases), aspect level directly looks at the opinion itself. It is based on the idea that an opinion consists of sentiment (positive or negative) and a target (of opinion).

Both the document level and sentence level classifications are already Sentiment Analysis and Opinion Mining 12 highly challenging. The aspect-level is even more difficult (Bing Liu,2012)

**3. Aspect Based Sentiment Analysis**

**The major tasks in aspect-based opinion mining are aspect extraction and aspect opinion classification.**

**3.1. Aspect Extraction**

Aspect extraction: This task extracts aspects that have been evaluated. For example, in the sentence, “The voice quality of this phone is amazing,” the aspect is “voice quality” of the entity represented by “this phone.” Note that “this phone” does not indicate the aspect GENERAL here because the evaluation is not about the phone as a whole, but only about its voice quality. However, the sentence “I love this phone”

evaluates the phone as a whole, i.e., the GENERAL aspect of the entity represented by “this phone.” Bear in mind whenever we talk about an aspect, we must know which entity it belongs to.

There are four main approaches for Aspect Extraction:

1. Extraction based on frequent nouns and noun phrases

2. Extraction by exploiting opinion and target relations

3. Extraction using supervised learning

4. Extraction using topic modeling

**3.1.1 Extraction based on frequent nouns and noun phrases**

This technique involves finding frequent nouns and noun phrases from the review as aspects. This method finds explicit aspects that are nouns and noun phrases from reviews in a given domain. This method finds explicit aspect expressions that are nouns and noun phrases from a large number of reviews in a given domain.

Hu and Liu (2004) used a data mining algorithm. Nouns and noun phrases (or groups) were identified by a part-of-speech (POS) tagger. Their occurrence frequencies are counted, and only the frequent ones are kept. A frequency threshold can be decided experimentally. The reason that this approach works is that when people comment on different aspects of an entity, the vocabulary that they use usually converges. Thus, those nouns that are frequently talked about are usually genuine and important aspects. Irrelevant contents in reviews are often diverse, i.e., they are quite different in different reviews. Hence, those infrequent nouns are likely to be non-aspects or less important aspects. Although this method is very simple, it is actually quite effective. Some commercial companies are using this method with several improvements.

The precision of this algorithm was improved in (Popescu and Etzioni, 2005). Their algorithm tried to remove those noun phrases that may not be aspects of entities.

**3.1.2 Extraction by exploiting opinion and target relations**

Since opinions have targets, they are obviously related. Their relationships can be exploited to extract aspects which are opinion targets because sentiment words are often known.

This method was used in (Hu and Liu, 2004) for extracting infrequent aspects. The idea is as follows: The same sentiment word can be used to describe or modify different aspects. If a sentence does not have a frequent aspect but has some sentiment words, the nearest noun or noun phrase to each sentiment word is extracted. Since no parser was used in (Hu and Liu, 2004), the “nearest” function approximates the dependency relation between sentiment word and noun or noun phrase that it modifies, which usually works quite well.

Additionally, this relation-based method is also a useful method for discovering important or key aspects (or topics) in opinion documents because an aspect or topic is unlikely to be important if nobody expresses any opinion or sentiment about it.

In (Pratima More and Archana Ghotkar,2016) experiment, the experimental results show that the relation-based method yields on an average 10% more accurate results as compared to the frequency-based method.

**3.1.3 Supervised Learning Approach**

Aspect extraction can be seen as a special case of the general information Sentiment Analysis and Opinion Mining 72 extraction problem. Many algorithms based on supervised learning have been proposed in the past for information extraction (Hobbs and Riloff, 2010; Mooney and Bunescu, 2005; Sarawagi, 2008). The most dominant methods are based on sequential learning (or sequential labeling). Since these are supervised techniques, they need manually labeled data for training. That is, one needs to manually annotate aspects and non-aspects in a corpus.

One can also use other supervised methods. For example, the method in (Kobayashi, Inui and Matsumoto, 2007) first finds candidate aspect and opinion word pairs using a dependency tree, and then employs a treestructured classification method to learn and to classify the candidate pairs as being an aspect and evaluation relation or not. Aspects are extracted from the highest scored pairs. The features used in learning include contextual clues, statistical co-occurrence clues, among others.

**3.1.4 Topic Modeling Approach**

In recent years, statistical topic models have emerged as a principled method for discovering topics from a large collection of text documents. Topic modeling is an unsupervised learning method that assumes each document consists of a mixture of topics and each topic is a probability distribution over words. A topic model is basically a document generative model which specifies a probabilistic procedure by which documents can be generated. The output of topic modeling is a set of word clusters. Each cluster forms a topic and is a probability distribution over words in the document collection.

Lin and He (2009) proposed a joint topic-sentiment model by extending LDA, where aspect words and sentiment words were still not explicitly separated. Brody and Elhadad (2010) proposed to first identify aspects using topic models and then identify aspect-specific sentiment words by considering adjectives only. Li, Huang and Zhu (2010) proposed two joint models, Sentiment-LDA and Dependency-sentiment-LDA, to find aspects with positive and negative sentiments. It does not find aspects independently and it does not separate aspect words and sentiment words.

Although topic modeling is a principled approach based on probabilistic inferencing and can be extended to model many types of information, it does have some weaknesses which limit its practical use in real-life sentiment analysis applications. One main issue is that it needs a large volume of data and a significant amount of tuning in order to achieve reasonable results. To make matters worse, most topic modeling methods use Gibbs sampling, which produces slightly different results in different runs

**3.2 Aspect Opinion Classification**

This task determines whether the opinions on different aspects are positive, negative, or neutral. In the first example above, the opinion on the “voice quality” aspect is positive. In the second, the opinion on the aspect GENERAL is also positive.

There are two main approaches to aspect opinion classification.

* Supervised Learning Approach
* Lexicon-based Approach

**3.2.1 Supervised Learning Approach**

In this approach, classification techniques like Naive Bayes Classifier [24] [23], Support Vector Machine (SVM) [24], Decision Tree classifier [24], kNN classifier [25] are used for text categorization and classifying opinion words on aspects into one of the polarity scales. In (Wei and Gulla, 2010), a hierarchical classification model was also proposed. However, the key issue is how to determine the scope of each sentiment expression, i.e., whether it covers the aspect of interest in the sentence. The current main approach is to use parsing to determine the dependency and the other relevant information. For example, in (Jiang et al., 2011), a dependency parser was used to generate a set of aspect dependent features for classification. A related approach was also used in (Boiy and Moens, 2009), which weights each feature based on the position of the feature relative to the target aspect in the parse tree.

**3.2.2 Lexicon-based Approach**

It is an unsupervised technique. The lexicon approach finds opinion orientation of review text using SentiWordNet. Work in [26], [27] proposes the use of SentiWordNet for classification of opinion words in opinion mining.

The lexicon-based approach can avoid some of the issues (Ding, Liu and Yu, 2008; Hu and Liu, 2004), and has been shown to perform quite well in a large number of domains. Such methods are typically unsupervised. They use a sentiment lexicon (which contains a list of sentiment words, phrases, and idioms), composite expressions, rules of opinions (Section 5.2), and (possibly) the sentence parse tree to determine the sentiment orientation on each aspect in a sentence. They also consider sentiment shifters, but-clauses (see below) and many other constructs which may affect sentiments.

Blair-Goldensohn et al. (2008) integrated the lexicon-based method with supervised learning. Kessler and Nicolov (2009) experimented with four different strategies of determining the sentiment on each aspect/target (including a ranking method).

**4. Cross-Domain Sentiment Classification**

It has been shown that sentiment classification is highly sensitive to the domain from which the training data is extracted. A classifier trained using opinion documents from one domain often performs poorly on test data from another domain. The reason is that words and even language constructs used in different domains for expressing opinions can be quite different. To make matters worse, the same word in one domain may mean positive but in another domain may mean negative. Thus, domain adaptation or transfer learning is needed. Existing researches are mainly based on two settings.

The first setting needs a small amount of labeled training data for the new domain (Aue and Gamon, 2005). The second needs no labeled data for the new domain (Blitzer, Dredze and Pereira, 2007; Tan et al., 2007). The original domain with labeled training data is often called the source domain.

In (Aue and Gamon, 2005), the authors proposed to transfer sentiment classifiers to new domains in the absence of large amounts of labeled data in these domains. They experimented with four strategies: (1) training on a mixture of labeled reviews from other domains where such data are available and testing on the target domain; (2) training a classifier as above, but limiting the set of features to those only observed in the target domain; (3) using ensembles of classifiers from domains with available labeled data and testing on the target domain; (4) combining small amounts of labeled data with large amounts of unlabeled data in the target domain (this is the traditional semi-supervised learning setting). SVM was used for the first three strategies, and EM for semi-supervised learning (Nigam et al., 2000) was used for the fourth strategy. Their experiments showed that the strategy (4) performed the best because it was able to make use of both the labeled and unlabeled data in the target domain.

In (Yang, Si and Callan, 2006), a simple strategy based on feature selection was proposed for transfer learning for sentence-level classification. Their method first used two fully labeled training set from two domains to select features that were highly ranked in both domains. These selected features were considered domain-independent features. The classifier built using these features was then applied to any target/test domains. Another simple strategy was proposed in (Tan et al., 2007), which first trains a base classifier using the labeled data from the source domain, and then uses the classifier to label some informative examples in the target domain. Based on the selected examples in the target domain, a new classifier is learned, which is finally applied to classify the test cases in the target domain.

**5. Conclusion**

* **Aspect Extraction** 
  + From the 4 main approaches extraction based on frequent nouns, noun phrases and by exploiting opinion and target relations have shown good results when compared to extraction using supervised learning and topic modeling.
* **Aspect Opinion Classification** 
  + From 2 main approaches, models based on Supervised learning which was trained on one domain often performs poorly in another domain.
  + Thus, supervised learning has difficulty to scale up to a large number of application domains when performing aspect opinion classification.

**Therefor the optimal approaches for two main tasks of the experiment will be frequent nouns, noun phrases or exploiting opinion and target relations(Aspect Extraction ) and lexicon based approach(Aspect Opinion Classification)**

**6. References**

* Xiaowen Ding, Bing Liu, Philip S. Yu (2008). *A Holistic Lexicon-Based Approach to Opinion Mining*.
* Wei Wei, Jon Atle Gulla (2010). *Sentiment Learning on Product Reviews via Sentiment Ontology Tree*.
* Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, Christopher Potts(2010). *Learning Word Vectors for Sentiment Analysis.*
* Chenghua Lin, Yulan He (2009). *Joint sentiment/topic model for sentiment analysis.*
* Sunita Sarawagi (2008). *Information Extraction.*
* Razvan C. Bunescu, Raymond J. Mooney(2005). *Learning to Extract Relations from the Web using Minimal Supervision.*
* Jerry R. Hobbs,Ellen Riloff (2010). *Information Extraction.*
* Bing Liu (2012). *Sentiment Analysis and Opinion Mining.*
* Minqing Hu and Bing Liu(2004). *Mining and Summarizing Customer Reviews.*