# Mining Opinion Targets from Text Documents: A Review

Khairullah Khan [1,2]
[1] Computer and Information Sciences Department, Universiti Teknologi PETRONAS, Malaysia
[2.] Institute of Engineering and Computing Sciences, University of Science & Technology Bannu Pakistan
khairullah_k@yahoo.com

Baharum B. Baharudin
Computer and Information Sciences Department, Universiti Teknologi PETRONAS, Malaysia
baharbh@petronas.com.my

Aurangzeb Khan
Aurangzebb_khan@yahoo.com
[2.] Institute of Engineering and Computing Sciences, University of Science & Technology Bannu Pakistan

***Abstract*—Opinion targets identification is an important task of the opinion mining problem. Several approaches have been employed for this task, which can be broadly divided into two major categories: supervised and unsupervised. The supervised approaches require training data, which need manual work and are mostly domain dependent. The unsupervised technique is most popularly used due to its two main advantages: domain independent and no need for training data. This paper presents a review of the state of the art unsupervised approaches for opinion target identification due to its potential applications in opinion mining from user discourse. This study compares the existing approaches that might be helpful in the future research work of opinion mining and features extraction.**

***Index Terms*—Opinion Mining; Sentiment Analysis; Opinion Targets; Machine Learning**

## I. Introduction

What other people think is naturally important for human guidance. Through opinions, humans can flux together diverse approaches, experiences, wisdom and knowledge of people for decision making. Humans like to take part in discussions and present their points of view. People often ask their friends, family members, and field experts for information during the decision making process. They use opinions to express their points of view based on experience, observation, concept, beliefs, and perceptions. The point of view about something can either be positive (shows goodness) or negative (shows badness), which is called the polarity of the opinion.

Opinions can be expressed in different ways. The following example sentences show different ways of opinion representations.

*Shahid is a good Cricket player.*
*The meal was quite good.*
*The hotel was expensive.*
*Terrorists deserve no mercy!*
*Hotel A is more expensive than B.*
*Coffee is expensive but tea is cheap.*
*This player is not worth any price and I recommend that you don't purchase it.*

An opinion has three main components i.e. the opinion holder or source of opinion, the object about which the opinion is expressed and the evaluation, view or appraisal which is called the opinion. For opinion identification, all these components are important.

Opinion can be collected from different sources e.g. individual interaction, newspapers, television, internet etc.; however, the internet is the richest source of opinion collection. Before the World Wide Web (WWW), people collected opinions manually. If an individual was to make a decision, he/she typically asked for opinions from friends and family members. Organizations conducted surveys through focused groups for collecting public opinion. This type of survey was expensive and laborious. Now, the internet provides this information with a single click and a very little cost.

With the advent of web 2.0, the internet allows web users to generate web content online and post their information independently. Due to this facility of the internet, web users can participate in a collaborative environment around the globe. Hence, the internet has become a rich source for social networks, customer feedback, online shopping etc. According to a survey, more than 45,000 new blogs are created daily along with 1.2 million new posts each day [1]. The information collected through these services is used for various types of decision making e.g. social network for: political, religious, security, and policy making; customer feedback for: products sales, purchases, and manufacturing. The

doi:10.4304/jetwi.5.4.343-353

trend of online shopping portals is increasing day by day. The vendors collect customer feedback for future trend prediction and product improvement through these portals. Opinion is the key element which has provided the inspiration for this work.

Although the internet is a rich source of opinions, having millions of blogs, forums and social websites with a large volume of updated information, unfortunately the web data is typically unstructured text which cannot be directly used for knowledge representation. Moreover, such a huge volume of data cannot be processed manually. Hence, efficient tools and potential techniques are needed to extract and summarize opinions. Research communities are trying for efficient utilization of the web information for knowledge requisition; this is in order to present it to the user in a well understandable and summarized manner. With the emergence of web 2.0, the task of posting and collecting opinions through the Web has become easy; however, the quality control, processing, compilation, and summarization have become potential research problems.

With the growing need of opinion analysis a new area called Opinion Mining is gradually emerged in the field of Natural Language Processing (NLP) and Text Mining. OM is a procedure used to extract opinion from a text. “OM is a recent discipline at the crossroads of information retrieval, text mining and computational linguistics which tries to detect the opinions expressed in natural language texts” [1]. OM is a field of knowledge discovery and data mining (KDD) which uses NLP and statistical machine learning techniques to differentiate opinionated text from factual text. OM tasks involve opinion identification, opinion classification (positive, negative, and neutral), target identification, source identification and opinion summarization. Hence, OM tasks require techniques from the field of NLP, Information Retrieval (IR); and Text Mining. The main issue is how to automatically identify opinion components from unstructured text and summarize the opinion about an entity from a huge volume of unstructured text. An overview of the OM concept is shown in the Figure 1.

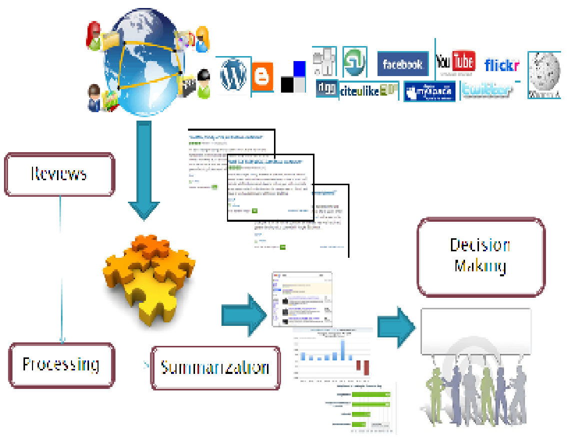

Figure 1. Overview of opinion mining process

The focus of this study is opinion target identification for the opinion mining process. The problem of opinion target identification is related to the question: “opinion about what?’. Opinion target identification is essential for opinion mining. For example, the in-depth analysis of every aspect of a product based on consumer opinion is equally important for consumers, merchants and manufacturers. In order to compare the reviews, it is required to automatically identify and extract those features which are discussed in the reviews. Furthermore, analysis of a product at feature level is more important e.g. which features of the product are liked and which are disliked by consumers [2] . Hence, feature mining of products is important for opinion mining and summarization. The task of feature mining provides a base for opinion summarization[3]. There are various problems related to opinion target extraction. Generally speaking, if a system is capable of identifying a target feature in a sentence or document, then it must be able to identify opinionated terms or evaluative expressions in that sentence or document. Thus in order to identify opinion targets at sentence or document level, the system should be able to identify evaluative expressions. Also, some features are not explicitly presented and are predicted from term semantics called implicit features. The focus of this paper is on explicit feature.

Opinion target identification is basically a classification problem which is defined as: to classify noun phrase or term as opinion target or not [4]. There are two widely used classification methods i.e. supervised and unsupervised. The supervised method needs prior knowledge annotated through manual process. Unsupervised classification depends on heuristics procedures and rules which do not need previous knowledge. Hence there are two main advantages of unsupervised method over supervised: Supervised technique need training data which manually labeled while unsupervised do not need hand-crafted training datasets, moreover supervised techniques are generally domain dependent as training data are manually labeled for specific domain [5, 6]. This paper provides a review of existing unsupervised approaches which has been popularly employed for opinion targets extraction within the past few years. The main goal of this work is to identify potential techniques for opinion targets extraction that might be helpful in the future research work in opinion mining. Hence the main contribution of this paper is the analysis of the factors that affect the existing unsupervised learning techniques of the opinion target extraction.

The entire paper is organized as follows: Section II explains related work and existing unsupervised approaches for opinion target extraction from unstructured reviews. Section III provides comparative analysis of the existing approaches and Section IV Concludes the paper.

## II. Unsupervised Approaches for Opinion Targets Identification

The unsupervised techniques has been popularly used for opinion target identification [6-17].

Popescu & Etzioni [9] used an unsupervised technique to extract product features and opinions from unstructured reviews. This paper introduces the OPINE system based on the unsupervised information extraction approach to mine product features from reviews. OPINE uses syntactic patterns for semantic orientation of words for identification of opinion phrases and their polarity.

Carenini, Ng et al. [15] developed a model based on user defined knowledge to create a taxonomy of product features. This paper introduces an improved unsupervised method for feature extraction that uses the taxonomy of the product features. The results of the combined approach are higher than the existing unsupervised technique; however, the pre-knowledge base mechanism makes the approach domain dependent.

Holzinger, Krüpl, & Herzog [10] use domain ontologies based on tabular data from web content to bootstrap a knowledge acquisition process for extraction of product features. This method creates a wrapper for data extraction from Web tables and ontology building. The model uses logical rules and data integration to reason about product specific properties and the higher-order knowledge of product features.

Bloom, Garg, & Argamon [14] describe an unsupervised technique for features and appraisal extraction. The authors believe that appraisal expression is a fundamental task in sentiment analysis. The appraisal expression is a textual unit expressing an evaluative attitude towards some target. Their paper proposed evaluative expressions to extract opinion targets. The system effectively exploited the adjectival appraisal expressions for target identification.

Ben-David, Blitzer et al. [16] proposed a structural correspondence learning (SCL) algorithm for domain classification. The idea depends on perception to get a prediction of new domain features based on training domain features; in other words, the author describes under what conditions a classifier trained on the source domain can be adapted for use in the target domain? This model is inspired by feature based domain classification. Blitzer, Dredze et al. [17] extended the structural SCL algorithm for opinion target identification.

Lu and Zhai [18] proposed automatic integration of opinions expressed in a well-written expert review with opinions scattered in various sources such as blogs and forums. The paper proposes a semi-supervised topic model to solve the problem in a principled way. The author performed experiments on integrating opinions about two quite different topics, i.e. a product and political reviews. The focus of this paper is to develop a generalized model that should be effective on multiple domains for extraction of opinion targets.

Ferreira, Jakob et al. [11] describe an extended pattern based feature extraction using a modified Log Likelihood Ratio Test (LRT), which was initially employed by [7] for target identification. This paper also presented an extended annotated scheme for product features, which was initially presented by [8] and a comparative analysis between feature extraction through Association Mining and LRT techniques.

The association rule mining for target extraction is initially implemented by [8] for target extraction, and extended by Chen et al. [12] using semantic based patterns for frequent feature refinement and identification of infrequent features.

One of the latest work on feature level analysis of opinion is reported by [6]. This paper describes a semi-supervised technique for feature grouping. Feature grouping is an important task for summarization of opinion. Same features can be expressed by different synonyms, words or phrases. To produce a useful summary, these words and phrases are grouped. For feature grouping the process generate an initial list to bootstrap the process using lexical characteristics of terms. This method empirically showed good results.

Goujon [4] presents a text mining approach based on linguistic knowledge to automatically detect opinion targets in relation to topic elements. This paper focuses on identification of opinion targets related to the specific topic. This approach exploits linguistic patterns for target identification.

The two most frequently reported unsupervised approaches for target and opinion identification are Association Mining (AM) [19] and Likelihood Ratio Test (LRT) approach [20]. The following sub sections provide a detail overview these two approaches.

### A. *Association Mining Approach*

The Association Mining approach for product features extraction (AME ) was employed by [8] for the first time. In this work, they extract frequent features through association rule mining technique [19]. This algorithm was originally used for market basket analysis which predicts dependency of an item sale on another item. Based on the analogy of the market basket analysis the authors in [8] assume that the words in a sentence can be considered as bought items. Hence the association between terms can predict features and opinion words association. The implementation of this technique was very successful in features extraction. Later on this approach is extended by [12] for the same task with semantic based pruning for frequent features refinement and identification of infrequent features. The subsequent approach improved the results of opinion target identification through association rule mining algorithm.

The AME approach formulates the process of opinion target identification into two steps. In the first step, it extracts frequent features through the Apriori algorithm and in the second step it employs a pruning algorithm to refine the candidate features from irrelevant features. The overall process is shown in a block diagram Figure 2.

The Apriori algorithm is called the king of data mining techniques as it was introduced in the early stages of the data mining field and has been potentially exploited for data mining and knowledge discovery. This algorithm has two steps: in step 1, it generates frequent item sets from a set of transactions that satisfies a user's specified minimum support, and in the second step, it discovers association rules from the frequent item sets discovered in step 1.

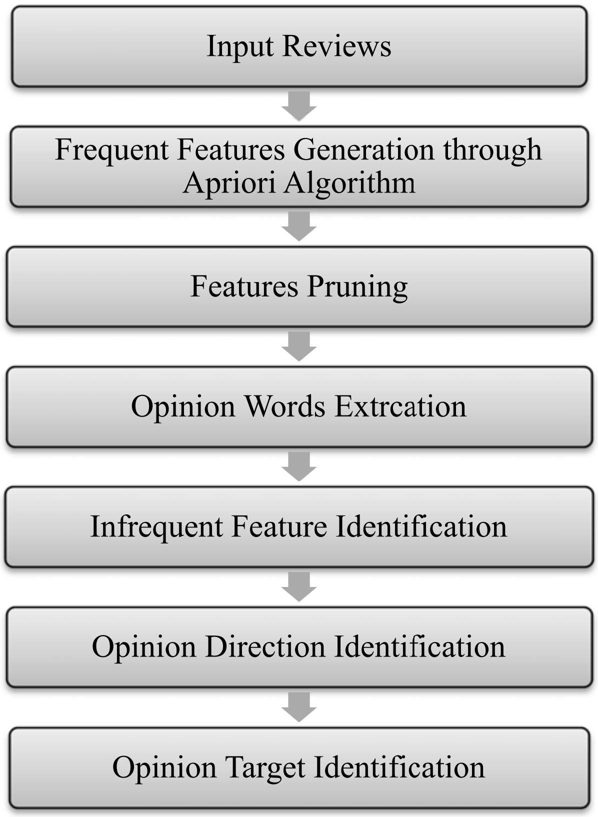
Input Reviews
Frequent Features Generation through Apriori Algorithm
Features Pruning
Opinion Words Extrcation
Infrequent Feature Identification
Opinion Direction Identification
Opinion Target Identification

Figure 2: Association Mining approach for opinion target extraction[8]

The association mining approaches uses the first step of the Apriori algorithm for extraction of product features that are frequently discussed in the review documents. The Apriori algorithm generates frequent feature sets from nouns in the reviews. This approach formulates the process of frequent feature identification as presented below.

***Frequent Features Identification***

The algorithm searches for frequently occurring product features in the input documents using the following steps.

- Each sentence is considered as a transaction.
- Each noun phrase in the sentence is considered as an item. Feature sets are created from the items.
- The algorithm then iterates through all the feature sets and counts the frequencies of each individual feature.

Based on the total number of candidate features a threshold value is calculated which is called the minimum support. Any feature having a frequency less than the minimum support threshold are discarded from the features' list. The authors in this work consider a feature set as frequent if it appears in more than 1% (minimum support) of the review sentences.

***Features Pruning***

The second step of this approach is pruning, which is used to refine the features obtained in step 1. The following two pruning steps are described.

- *Compactness Pruning*

Compactness is used to check features that contain two or three words and remove those features which are not co-occurring more than at least two times. For example, having the phrase "battery life" if it appears in two or more sentences at a distance of at most three words in between them then it is a compact feature. However, if it does not co-occur at least two times then it is removed from the feature list.

- *Redundancy Pruning*

Redundancy pruning is used to remove redundant features that contain single words. A feature is considered as redundant if it occurs in a compact feature and has a lower frequency then the p-support. The p-support is different from the general support count in association mining. For example, "life" occurs 6 times and "battery life' occurs 5 times then in the candidate features, the feature "life" alone is considered as a redundant feature. This work only considers nouns for the features and this rule does not consider any other lexical categories at all.

### *B. Association Mining by Wei et al. (2010)*

This approach uses a semantic-based refinement of the frequent features obtained through the association mining approach. This work describes a model based on a list of positive and negative subjective adjectives defined in the General Inquirer (GI). The aim of semantic-based refinement is to overcome the following two limitations of the [8] approach:

- Frequent but non Product Features,
- Infrequent but Product Features.

This approach describes the following three semantic-based pruning rules to handle these limitations.

***Co-occurrence-based Pruning***

The previously described association mining approach is based on the frequency of noun phrases to discover frequent features. However, some of the noun phrases in a document may have a high frequency but not be an opinion target. This rule is designed to address this limitation. This rule is defined as:

- For each frequent feature a count is carried out for the number of review sentences in which the feature co-occurred with subjective adjectives.
- If the count obtained in the previous step is less than a prescribed co-occurrence threshold value (this study considers it as 1) then it is removed from the frequent feature list.

The formal representation of this model is given as below.

$$\text{IF} \quad \sum_{i=1}^{|S|} \text{co} - \text{occur}(f, ow, s_i) < \propto \text{ Then } F = F - \{f\} \quad (1)$$

Where

$$\text{co} - \text{occur}(fqf, ow, s_i) = \begin{cases} 1 \text{ if } \exists \text{ op} \in \text{ow such that } f \in s_i \text{ and op} \in s_i \\ 0 \quad \text{otherwise} \end{cases} \quad (2)$$

Here $|S|$ represents the number of sentences, f is a frequent feature, $s_i$ a sentence, ow an opinion word, and F frequent feature sets. In this step, the frequent features are considered as product features.

### *Opinion-based Infrequent Feature Identification*

The earlier approach employs the nearest adjective as opinion words to identify infrequent features in the review sentences that do not contain frequent features. This approach may not be effective for all adjectives e.g. "such/JJ thing/NN", "whole/JJ lot/NN", "simple/JJ point/NN" etc. Similarly in a sentence "The/DT picture/NN is/VBZ not/RB rich/JJ in/IN color/NN", the noun closest to the adjective "rich" is "color" but picture is not the target feature, rather the word color is target. To address this limitation, the author describes the following rule:

If a review sentence contains a subjective adjective, then this rule first examines the word or group of words immediately after the subjective adjective in the sentence. If the word after the adjective is a noun or noun phrase, then it is considered as an infrequent feature and is added to the list of frequent features. If the word after the adjective is not a noun phrase, then the heuristic searches for a noun phrase before the adjective in the sentence. For example, with the sentence "this/WDT camera/NN has/VBZ excellent/JJ picture/NN quality/NN", according to this rule, "picture quality" is the actual feature. Hence, this rule satisfies both conditions of the nearest adjective and is similar to the previous approach; moreover, the situation as described in the previous sentence where the feature is picture and as the word "in/IN" is not a noun after the subjective adjective thus it searches for the nearest noun before the subjective adjective.

### *Conjunction-based Infrequent Feature Identification*

Some of the features rarely occur and thus the frequency based approach fails to identify them. However, based on the conjoined relation with other features they can be easily identified. This rule is described as follows:

For every conjunction of nouns and noun phrases in each review sentence, if one has been identified as a target feature, then this rule includes the remaining nouns and noun phrases in the conjunction as a product feature. The mathematical model of this rule is defined as:

$$\text{If } \exists \text{ np}_i \in \text{CN such that np}_i \in \text{PF}, \qquad \text{Then}$$

$$\forall \text{ np}_j \in \text{CN and np}_j \neq \text{np}_i \text{ PF} \cup \{\text{np}_j\} \qquad (3)$$

Where $\text{np}_i$ and $\text{np}_j$ represents a noun or noun phrase in conjunction (CN) with the identified features, and PF represents product features already identified in the previous step.

Based on the above three rules, this approach improved both precision and recall of the association mining approach for opinion target identification. This approach reported an average improvement of about 10.7% in recall and 2.5% in precision.

## *C. Likelihood Ratio Test Approach*

The other potentially employed unsupervised classification technique is the Likelihood Ratio Test (LRT). The LRT was introduced by [20] and has been reported in different NLP tasks. The LRT was employed by [7] for product feature extraction and sentiment analysis. One of the latest approaches for product feature identification using the LRT technique is described by [11]. The LRT technique assumes that a feature related to the topic is explicitly presented by a noun phrase in the document using syntactic patterns associated with subjective adjectives. The overall process is explained in the Figure 3.

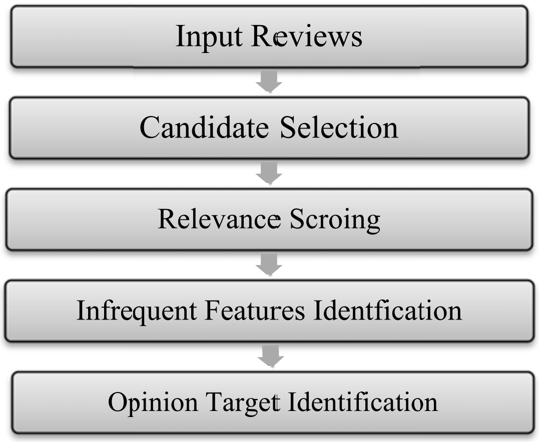
Input Reviews
Candidate Selection
Relevance Scroing
Infrequent Features Identfication
Opinion Target Identification

Figure 3: Opinion targets extraction [7, 11]

Yi, Nasukawa et al. [7] described different linguistic patterns termed as base noun phrases for candidate selection and then employs relevance scoring to refine the candidate features. The overall process of the likelihood ratio test based target extraction is defined as below.

### *Selection of Candidate Feature using Linguistic Patterns*

In this approach the selection process of candidate features is based on noun phrase patterns. The following patterns are employed in this work.

- *Base Noun Phrases (BNP)*

These patterns are used to extract candidate features using the following combination of noun (NN) and adjective (JJ).

NN, NN NN, JJ NN, NN NN NN, JJ NN NN, JJ JJ NN.

- *Definite Base Noun Phrase (dBNP)*

These patterns present noun phrases (BNP) with the definite article "the" before the BNP. The idea behind these patterns is that some proper nouns start with the article "the" therefore these patterns are useful for named entity extraction

- *Beginning Definite Base Noun Phrases (bBNP)*

This pattern presents a sequence of definite noun phrases followed by verbs. This pattern describes that the noun phrase in between the article "the" and a verb are mostly observed as features.

### *Relevance Scoring*

Yi, Nasukawa et al. [7] presented unsupervised technique for relevance scoring of candidate features. This paper employed two unsupervised techniques, i.e. The Mixture Model, and LRT. However, the results show that the LRT performed relatively good. The likelihood ratio test is formulated as:

Let $D_c$ denoted topic relevant collection of documents and $D_n$ represents collection of documents not relevant to the topic. Then a base noun phrases occurring in the $D_c$ are candidate feature to be classified as topic relevant or topic irrelevant using the likelihood ratio test as: if the likelihood score of BNP satisfies the predefined threshold value then BNP is considered as target feature. The LRT value for any BNP x is calculated as:

Let $n_1$ denotes the frequency of a BNP in a Dc, $n_2$ represents sum of frequencies of all BNPs in $D_c$ except x, $n_3$ denoted frequency of x in $D_n$, and $n_4$ represents the sum of frequencies of all BNPs in $D_n$ except the frequency of x.

Then the ratios of relevancy of the BNP x to topic and non-topic, which are presented by $r_1$ and $r_2$ respectively, can be calculated as below.

$$r_1 = \frac{n_1}{n_1+n_2} \qquad (4)$$

$$r_2 = \frac{n_3}{n_3+n_4} \qquad (5)$$

Thus the combined ratio is calculated as:

$$r = \frac{n1 + n_3}{n_1 + n_2 + n_3 + n_4} \qquad (6)$$

Hence to normalize the ratios with log:

$$lr = (n_1 + n_2)\log(r) + (n_3 + n_4)\log(1-r) - n_1\log(r_1) - n_3\log(1-r_1) - n_2\log(r_2) - n_4\log(1-r_2) \qquad (7)$$

Hence the likelihood ratio is calculated as below.

$$-2\log\Lambda = \begin{cases} -2 * lr & if\ r_2 < r_1 \\ 0, & if\ r_2 \geq r_1 \end{cases} \qquad (8)$$

The likelihood is directly proportional to the value of $-2\log\Lambda$.

### D. *Likelihood Approach by Ferreira et al. (2008)*

A more extensive study of the LRT approach for opinion target identification is presented by this paper. As mentioned in the previous sub section, the LRT was employed by [7]; however, due to non-availability of proper data sets for evaluation measures the author only calculated precision.

Ferreira et al. (2008) performed an evaluation on the state-of the art datasets, which are manually, annotated corpuses created by [8]. Furthermore, they have modified the algorithm using subsequent similarity measures based on the following two rules.

### *Identification of Feature Boundaries for Patterns*

The earlier work [7] used BNPs, dBNPs and bBNPs for candidate feature identification. Noun phrases in these patterns are considered as candidate features. However, there is no rule mentioned for multiple matches. For example, in the pattern "battery life", three features can be reflected: "battery life", "battery", and "life". The recent work [11] extended the earlier algorithm, which only selects the longest BNP patterns. For example, in the above expression this rule considers only "battery life" as a feature.

### *Classification of Patterns with an Adjective Noun (JJNN)*

Most of the candidate BNPs is combinations of JJNN patterns. The adjective sometimes represents features e.g. "digital images" and sometimes it represents an opinion e.g. beautiful image; hence, it is required to classify the subsequent adjectives in the candidate patterns. Subsequent similarity rule is employed by [11], which have improved the results. Another main contribution of this paper is the new annotation scheme of the features in the existing dataset that were originally employed by [8]. According to the revised annotation scheme, the number of features was increased as their focus was on all features.

## III. Comparative Analysis

This section describes the analysis of the unsupervised approaches that has been potentially employed for opinion targets extraction. As explained in section II there are most popular used techniques that have been employed for opinion targets extraction.

### A. *Analysis of Factors Affecting the Existing Approaches*

This section explains the analysis of the factors affecting the existing unsupervised techniques of opinion targets extraction. We have performed analysis on the bench mark dataset that have been employed by the existing approaches. The experimental setup is divided into two broad categories. The first category is related to candidate selection based on linguistic patterns while the second one is focusing on features selection based relevance scoring.

### B. *Datasets*

This section describes the datasets that have been used for the analysis and evaluation in this work. In this work, benchmark datasets of the customer reviews about five different products are employed. These datasets have been reported in numerous works for opinion mining and target identification. These datasets are crawled from amazon review sites and are manually annotated by [8]. The datasets are freely available from the authors' website[1]. In these datasets, each product feature with opinion scoring is properly tagged in each sentence through a manual process according to a prescribed annotation scheme as shown below.

- A sentence is considered as opinionated if it contains positive or negative comments about features of the product.
- Positive and negative comments are opinion statements containing adjectives that either have a positive or negative orientation.
- A product feature is the characteristic of the product about which opinions are expressed by the customers.

The datasets contain customer reviews about four different electronic products, i.e. Camera (Canon G3 and Nikon Coolpix 4300), DVD player (Apex AD2600 Progressive-scan), mp3 player (Creative Labs Nomad Jukebox Zen Xtra 40GB) and cell phone (Nokia 6610). The summary of each dataset is given in Tables 1: including the total number of reviews (number of documents), total number of sentences, number of sentences with opinions and targets with percentage, total distinct base noun phrases which count each distinct BNP as 1; the total target features shows the count of all target features in each dataset, the average target features shows target features out of the total distinct BNPs, the target types show the number of distinct target features in each dataset and the ratio of target features to the total target occurrence.

### C. *Experimental Setup*

Although the results are of the aforementioned techniques have been already given in the respective papers and there is no need to reproduce it. However in order to empirically prove the factors affecting the existing approach we have performed analysis on the factors that affect the performance of the existing approaches.

As mentioned in the existing approaches there are two phases of the target extraction techniques. The first phase is related to candidate selection while the second phase is related to relevance scoring. In the candidate selection process patterns of language elements with grammatical relations are employed to identify candidate features. In relevance scoring phase the candidate features are refined using unsupervised machine learning techniques. Hence our experimental setup is divided into the following two phases to identify strength and limitations of the existing approaches in each phase.

### D. *Analysis of Patterns for Candidate Selection*

This section provides a comparative analysis of the linguist patterns that have been employed for candidate selection. As mentioned earlier both AME and LRT approaches are using noun phrase for candidate selection. However there is a difference between the selections. AME uses association between the noun phrases and top features with highest frequency is selected that qualify the minimum support as target features. While The LRT select the noun phrases based on grammatical sequence of terms. In order to investigate best patterns for candidate selection the following patterns are examined: Base noun phrase (BNP), Definite based noun phrases (dBNP), Beginning definite base noun phrases (bBNP), and Combined base noun phrase pattern (cBNP). The first four patterns have already been discussed. While the cBNP pattern is employed by [23] which is set of patterns defined as below.

- Noun Phrase-Verb Phrase-Adjective (NP VB JJ)
- Noun Phrase-Verb Phase-Adverb Adjective (NP VB RB JJ)
- Noun Phrase-Verb Phase-Adverb Adjective NN (NP VB RB JJ NN)
- Definite Base Noun Phrase (dBNP)
- Preposition Based Noun Phrase (iBNP)
- Subjective Base Noun Phrase (sBNP)

In order to extract these patterns from the datasets the Stanford part of speech tagger and textSTAT software has been used, The Stanford part of speech tagger is employed for part of speech tagging [21], while TextStat 3.0 is employed for pattern extraction and test analysis [2]. This software is simple and has been used by a number of works for searching terms and strings in English texts [22].

The comparative results are shown in figures 4, 5 and 6. The precision of bBNP is higher than the other patterns as it extracts fare number of features. While the recall of BNP pattern is higher as it extracts all BNPs, however, its recall is very low due to its false negative features. The F-score of our proposed cBNP is significantly higher than the other patterns. Thus the overall performance of cBNP is good.

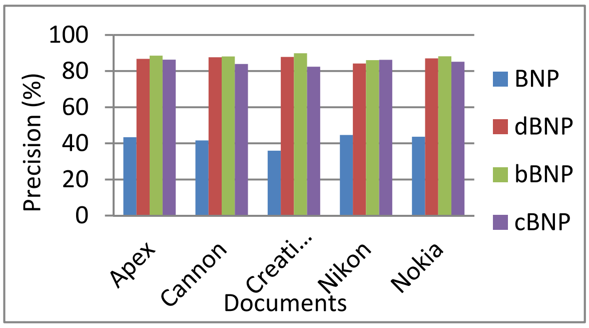

Figure 4: Precision of candidate selection based on dependency patterns

[1] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

[2] http://neon.niederlandistik.fu-berlin.de/en/textstat/

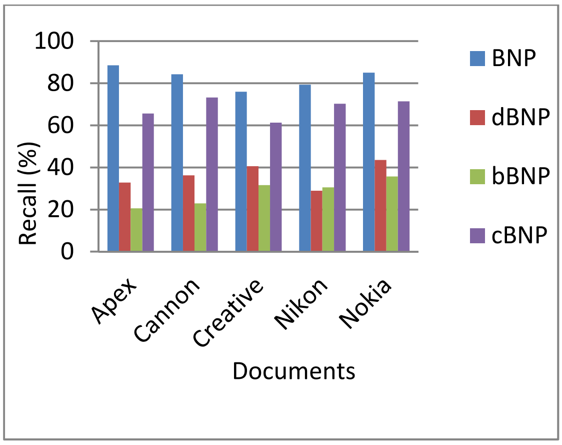

Figure 5: Recall of candidate selection based on dependency patterns

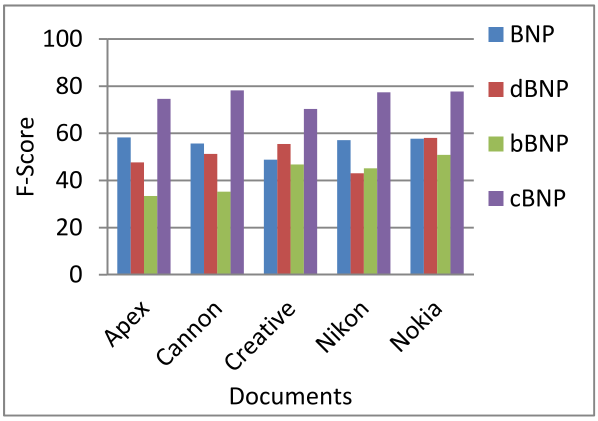

Figure 6: F-Score of candidate selection based on dependency patterns

### E. *Analysis of Frequency Based Relevance Scoring*

This section demonstrates how the target extraction techniques are affected by the threshold values. In order to analyze this problem, we conducted experiment for finding infrequent features on each data set. Table 2 shows the sample of target features which have zero LRT values due to their rear occurrence in the review dataset. Hence based on the low frequency distribution a number of target features cannot be predicted by the unsupervised learning. Table 3 shows the ratio of infrequent features classified by LRT technique.

Refer to the results in Sections B and C there are two main issues related to the extraction of opinion targets. The first issue is related to linguistic patterns that have been employed for candidate selection. The results can be greatly improved with the use of proper patterns. As shown in the graphs in figures 4, 5 and 6, the F-score is significantly improved with the use of combination of patterns.

The other main issue is related to frequency based relevance scoring for features selection. It has been observed that even within large documents there exist a lot of features which have very low frequency hence cannot be detected even by adjusting a small value of threshold. Hence the recall of the unsupervised techniques is greatly affected due high ratio of false negative value.

As discussed in the previous sections the existing unsupervised approaches exploit linguistic patterns and frequency based relevance scoring techniques to identify opinion targets. However there are certain issues related to both patterns selection and relevance scoring that might affect the performance of the techniques. Since most of the work consider base noun phrases as opinion targets. However all base noun phrase in text cannot be opinion targets. Hence the existing research work has been primarily focused on the problem of selecting dependency patterns for targets identification. For example some sentences in a review document may not have opinion while other sentences may have more than one base noun phrases with few opinion targets. For example the sentences “The/DT camera/NN comes/VBZ with/IN a/DT second/JJ battery/NN. I/PRP purchased/VBD it/PRP in/IN a/DT departmental/NN store/NN.” do not have any opinion targets although it have base noun phrase. While in the sentence “The battery/NNP is/VBZ very/RB good/JJ even/RB when/WRB using/VBG flash/NN and/CC lcd/NN” there is only one opinion target “battery” although it has three different BNPs. Hence simply selecting BNPs provides a large false positive ratio. To overcome this issue the existing worked has proposed various solutions. For example the association mining approach assumes that opinion targets are frequently discussed in reviews. However this approach suffers from two major issues i.e. frequent but not opinion target and infrequent but opinion target. As mentioned earlier, to overcome these problems the existing works have proposed pruning. Although the performance have been improved with pruning rules. However, the results show that there is still gap for further improvement.

The Likelihood Ratio Test approach assumes that the Base Noun Phrases with dependency patterns containing subjective adjective are best candidate for opinion targets instead of simply selecting base noun phrases. Hence this technique depends on opinionated expression. However, the question about how to identify opinionated expressions! is itself a challenging problem. There can be more than one noun phrases with adjective in sentences. For example the sentence “The/DT picture/NN quality/NN is/VBZ not/RB rich/JJ in/IN color/NN” have two candidate base noun phrases “Picture quality” and “Color”, and one adjective “rich”. Although the “color” is a feature that can be occurred many items in different sentences; however, in this case the “picture quality” is basically opinion target. According dBNP pattern mentioned earlier, “The picture quality” can to be correctly selected as opinion target from the above sentence. However these patterns are not effective in many cases. For example if we look into the review sentence “this dvd play is basically junk”; it has opinion targets “player” but do not satisfy the dBNP pattern rules. Since LRT based approach also depends on frequency distribution therefore it also suffer from the same two main issues i.e. frequent but none opinion targets and rarely occurred but opinion targets.

TABLE 1:
SUMMARY OF THE FIVE PRODUCT DATASETS WITH MANUALLY TAGGED OPINION TARGETS BY [8]

| Description | Dataset | | | | |
|---|---|---|---|---|---|
| | Apex | Cannon | Creative | Nikon | Nokia |
| Reviews | 99 | 45 | 95 | 34 | 41 |
| Total sentences | 739 | 597 | 1716 | 346 | 546 |
| Target types | 110 | 100 | 180 | 74 | 109 |

TABLE 2:
SAMPLE SET OF INFREQUENT FEATURES

| **Dataset** | **Infrequent Features** |
|---|---|
| Apex | read,look,sound,price,door,size,design,quality,support,weight,case,forward,output,product,run,unit,video,work,code,direction,disk,display,finish,machine,motor,noise, panel, recognize, service, speed, use, apex |
| Cannon | body,control,depth,design,display,feel,finish,focus,function,image,learning,look,made,noise,option,print,quality,remote,service,shape,shot,speed,use,weight,zoom |
| Creative | alarm,appearance,balance,break,build,capacity,case,change,clock,control,cover,creative,deal,design,display,equipment,feature,feel,finding,game,look,looking,manage,memory,music,name,option,panel,pause,play,product,program,quality,recognition,recording,remote,remove,style,support,switch,thing,top,unit,use,value,volume,weight,wheel,work,sorting,navigation |
| Nikon | construction,control,delay,design,function,image,learn,menu,price,quality,size,software,transfer,use,weight |
| Nokia | application,background,call,command,construction,design,game,keys,look,memory,message,network,picture,plan,quality,resolution,ring,service,software, sound, speaker, tone, use, voice, work |

TABLE 3
DISTRIBUTION OF INFREQUENT FEATURES IN EACH DATASET

| Dataset | Total | Frequent | Infrequent | %Infrequent |
|---|---|---|---|---|
| Apex | 110 | 78 | 32 | 29.09090909 |
| Cannon | 98 | 73 | 25 | 25.51020408 |
| Creative | 179 | 129 | 50 | 27.93296089 |
| Nikon | 73 | 58 | 15 | 20.54794521 |
| Nokia | 110 | 84 | 26 | 23.63636364 |

## IV. Conclusion

This paper presents a systematic review of unsupervised approaches of opinion target identification from unstructured reviews. This study shows that there two main issued in unsupervised learning of opinion targets from unstruted reviews i.e. Frequent base noun pharse but not target features and Infrequent but target features. Besides a significant improvements in the opinion target identification techniques these two prolmes are still challenging. Our analysis shows that results can be greatly improved with the imrpovement in candidate selection and relevance scroing. We have proposed hybrid patterns based candidate selection that have shown considerable improvement in the true positive. We have also the affect of threshold value on relevance scroing using Likelihood ratio test. It was found that 20 to 30 % infrequent features cannot be detected by the LRT technique due low frequency of the target feature. Hence the recall of the this method low due to high number of flase negative features. This shows that recall can be improved with the selection of infrequent features. Hence future should focus on dependecy patterns and infrequent features for the better improvement in the results.

## References

[1] Pang, B. and L. Lee, Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2008. 2(1-2): p. 135.

[2] Zhang, L. and B. Liu, Identifying noun product features that imply opinions, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2. 2011, Association for Computational Linguistics: Portland, Oregon. p. 575-580.

[3] Somprasertsri, G., Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization. . Journal of Universal Computer Science, 2010. vol. 16(6): p. 938-955.

[4] Goujon, B. Text Mining for Opinion Target Detection. in European Intelligence and Security Informatics Conference (EISIC). 2011.

[5] Qiu, G., et al., Domain Specific Opinion Retrieval Information Retrieval, in Fifth Asia Information Retrieval Symposium. 2009, Springer Berlin / Heidelberg: Japan. p. 318-329.

[6] Zhai, Z., et al., Clustering product features for opinion mining, in The fourth ACM international conference on Web search and data mining. 2011, ACM: Hong Kong, China. p. 347-354.

[7] Yi, J., et al. Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. in Third IEEE International Conference on Data Mining (ICDM) 2003.

[8] Hu, M. and B. Liu, Mining and summarizing customer reviews, in 10th ACM SIGKDD international conference on Knowledge discovery and data mining. 2004, ACM: Seattle, WA, USA. p. 168-177.

[9] Popescu, A.-M. and O. Etzioni, Extracting product features and opinions from reviews, in Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. 2005, Association for Computational Linguistics: Vancouver, British Columbia, Canada. p. 339-346.

[10] Holzinger, W., B. Krüpl, and M. Herzog. Using ontologies for extracting product features from web pages. in 5th International Semantic Web Conference, ISWC 2006. 2006. Athens, Georgia, USA.

[11] Ferreira, L., N. Jakob, and I. Gurevych. A Comparative Study of Feature Extraction Algorithms in Customer Reviews. in Semantic Computing, 2008 IEEE International Conference on. 2008.

[12] Wei, C.-P., et al., Understanding what concerns consumers: a semantic approach to product features extraction from consumer reviews. Info Syst E-Bus Management, 2010(8): p. 149-167

[13] Wong, T.-L. and W. Lam, An unsupervised method for joint information extraction and feature mining across different Web sites. Data & Knowledge Engineering, 2009. 68(1): p. 107-125.

[14] Bloom, K., N. Garg, and S. Argamon. Extracting appraisal expressions. in In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics. 2007. Rochester, New York, USA.

[15] Carenini, G., R.T. Ng, and E. Zwart, Extracting knowledge from evaluative text, in Proceedings of the 3rd international conference on Knowledge capture. 2005, ACM: Banff, Alberta, Canada. p. 11-18.

[16] Ben-David, S., et al., Analysis of representations for domain adaptation. In Advances in Neural Information Processing Systems 2007. vol. 19.

[17] Blitzer, J., M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. in 45th Annual Meeting of the Association of Computational Linguistics. 2007. Prague, Czech Republic.

[18] Lu, Y. and C. Zhai. Opinion integration through semi-supervised topic modeling. in 17th International World Wide Web Conference (WWW '08). 2008. Beijing, China.

[19] Agrawal, R. and R. Srikant, Fast Algorithms for Mining Association Rules in Large Databases, in 20th International Conference on Very Large Data Bases. 1994, Morgan Kaufmann Publishers Inc. p. 487-499.

[20] Dunning, T., Accurate methods for the statistics of surprise and coincidence. Comput. Linguist., 1993. 19(1): p. 61-74.

[21] Toutanova, K., et al., Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network, in North American Association for Computational Linguistics (NAACL). 2003. p. 173-180.

[22] Diniz, L., Comparative Review: TextStat 2.5, ANTCONC 3.0, and Compleat Lexical Tutor 4.0. Language Learning & Technology, 2005(Vol 9 Issue 3): p. 22-27.

**Khairullah khan** received MSc Computer Science Degree from University of Peshawar Pakistan and has PhD degree from Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Malaysia. He is working as Assistant Professor at University of Science and Technology Bannu Pakistan. His current research interests include NLP, Data Mining, Opinion Mining and Information Retrieval.

**Baharum Bin. Baharudin** received his Master Degree from Central Michigan University, USA and his PhD degree from University of Bradford, UK. He is currently Associate Professor at the Department of Computer and Information Sciences, Universiti Teknologi PETRONAS Malaysia. His research interests lies in Image Processing, Data Mining and Knowledge Management.

**Aurangzeb khan** received BS-Degree in Computer Science from Gomal University D.I.Khan, Pakistan, and Master Degree in Information Technology From University of Peshawar, Pakistan, and PhD degree from Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Malaysia. He is an Assistant Professor at University of Science & Technology Bannu Pakistan. His current research interests include Data Mining, Opinion Mining and Information Retrieval.