Research Journal of Applied Sciences, Engineering and Technology 7(12): 2400-2410, 2014
DOI:10.19026/rjaset.7.543
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: July 27, 2012 Accepted: September 03, 2012 Published: March 29, 2014

Review Article

A Review of Unsupervised Approaches of Opinion Target Extraction from Unstructured Reviews

Khairullah Khan, Baharum Baharudin and Aurangzeb Khan
CIS Department, University Technology PETRONAS, Malaysia

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 study presents a review of the state of the art unsupervised approaches for opinion target identification due to its potential applications in opinion mining from web documents. This study compares the existing approaches that might be helpful in the future research work of opinion mining and features extraction.

Keywords: Features extraction, machine learning, opinion mining, opinion targets, sentiment analysis

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 (Aurangzeb *et al*., 2011b; Baharum and Khairullah, 2011).

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

- Shahid Afridi is a good player.
- She is not a good actress.
- The breakfast 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 (Aurangzeb *et al*., 2011a; Khan *et al*., 2009). 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 (Khan *et al*., 2010b; Pang and Lee, 2008). 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 trend of online shopping portals is increasing day by day. The vendors collect customer feedback for future trend prediction and product

Corresponding Author: Khairullah Khan, CIS department, University Technology PETRONAS, Malaysia

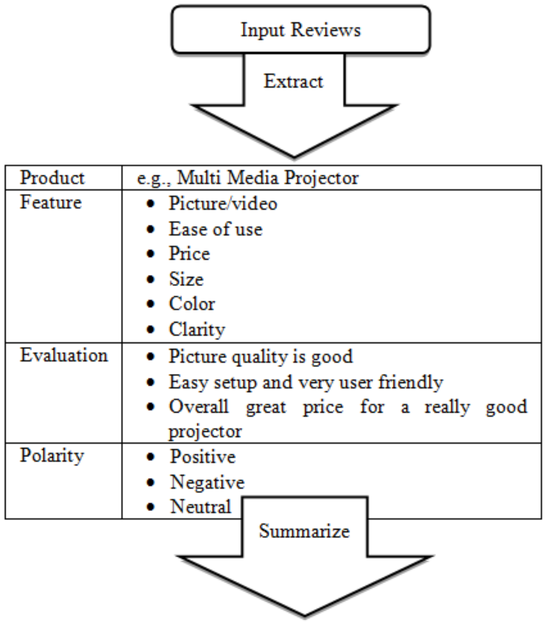

Fig. 1: Overview of opinion mining process

improvement through these portals. Opinion is the key element which has provided the inspiration for this study.

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 (Baharum and Khairullah, 2011; Baharum and Baharudin, 2010).

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” Pang and Lee 2008 (Baharum and Baharudin, 2010). 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 (Khan *et al*., 2010b). 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 Fig. 1.

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 (Balahur and Montoyo, 2008). 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 (Khan *et al*., 2010a; Zhang and Liu, 2011). Hence, feature mining of products is important for opinion mining and summarization. The task of feature mining provides a base for opinion summarization (Khan *et al*., 2012; Somprasertsri and Lalitrojwong, 2010). 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 (Aurangzeb *et al*., 2011b) 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. However this study only focuses on explicit features.

Opinion target identification is basically a classification problem which is defined as: to classify noun phrase or term as opinion target or not (Baharum and Baharudin, 2010; Goujon, 2011). 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 (Baharudin *et al*., 2010; Qiu *et al*., 2009; Zhai *et al*., 2011). This study 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 study is to identify potential techniques for opinion targets extraction that might be helpful in the future research work in opinion mining.

METHODOLOGY

Unsupervised approaches for opinion targets identification: The unsupervised techniques has been popularly used for opinion target identification (Ben-David *et al*., 2007; Blitzer *et al*., 2007; Bloom *et al*., 2007; Carenini *et al*., 2005; Ferreira *et al*., 2008; Holzinger *et al*., 2006; Hu and Liu, 2004; Popescu *et al*., 2005; Wei *et al*., 2010; Wong and Lam, 2009; Yi *et al*., 2003; Zhai *et al*., 2011).

Popescu *et al*. (2005) used an unsupervised technique to extract product features and opinions from unstructured reviews. This study 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 *et al*. (2005) developed a model based on user defined knowledge to create taxonomy of product features. This study 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 *et al*. (2006) 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 *et al*. (2007) 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 study proposed evaluative expressions to extract opinion targets. The system effectively exploited the adjectival appraisal expressions for target identification.

Ben-David *et al*. (2007) 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 *et al*. (2007) extended the structural SCL algorithm for opinion target identification.

Lu and Zhai (2008) proposed automatic integration of opinions expressed in a well-written expert review with opinions scattered in various sources such as blogs and forums. The study 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 study is to develop a generalized model that should be effective on multiple domains for extraction of opinion targets.

Ferreira *et al*. (2008) describe an extended pattern based feature extraction using a modified Log Likelihood Ratio Test (LRT), which was initially employed by Yi *et al*. (2003) for target identification. This study also presented an extended annotated scheme for product features, which was initially presented by Hu and Liu (2004) and a comparative analysis between feature extraction through Association Mining and LRT techniques.

The association rule mining for target extraction is initially implemented by Hu and Liu (2004) for target extraction and extended by Wei *et al*. (2010) using semantic based patterns for frequent feature refinement and identification of infrequent features.

One of the latest works on feature level analysis of opinion is reported by Zhai *et al*. (2011). This study 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 (2011) presents a text mining approach based on linguistic knowledge to automatically detect opinion targets in relation to topic elements. This study 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) (Agrawal and Srikant, 1994) and Likelihood Ratio Test (LRT) approach (Dunning, 1993). The following sub sections provide a detail overview these two approaches.

Association mining approach: The Association Mining approach for product features Extraction (AME) was employed by Hu and Liu (2004) for the first time. In this study, they extract frequent features through association rule mining technique Agrawal and Srikant (1994). 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 Hu and Liu (2004) 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

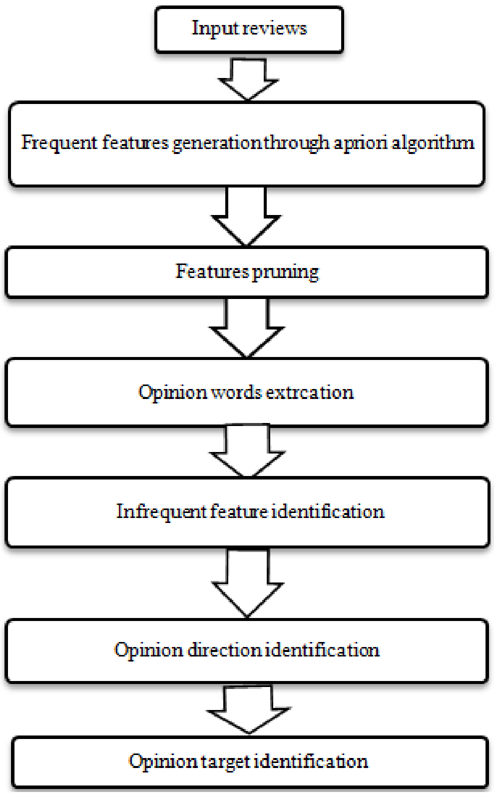

Fig. 2: Association mining approach for opinion target extraction Hu and Liu (2004)

by Wei *et al.* (2010) 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 Fig. 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.

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 study 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 study only considers nouns for the features and this rule does not consider any other lexical categories at all.

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 study 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 Hu and Liu (2004) 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:

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

where,

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

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:

$$\text{If } \exists \ \mathrm{np_i} \in \mathrm{CN} \text{ such that } \mathrm{np_i} \in \mathrm{PF}, \text{Then} \quad \forall \ \mathrm{np_j} \in \mathrm{CN} \text{ and } \mathrm{np_j} \neq \mathrm{np_i} \ \mathrm{PF} \cup \{\mathrm{np_j}\} \tag{3}$$

where, np_i and 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.

Likelihood ratio test approach: The other potentially employed unsupervised classification technique is the Likelihood Ratio Test (LRT). The LRT was introduced by Dunning (1993) and has been reported in different NLP tasks. The LRT was employed by Yi *et al*. (2003) for product feature extraction and sentiment analysis. One of the latest approaches for product feature identification using the LRT technique is described by Ferreira *et al*. (2008). 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 Fig. 3.

Yi *et al*. (2003) 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

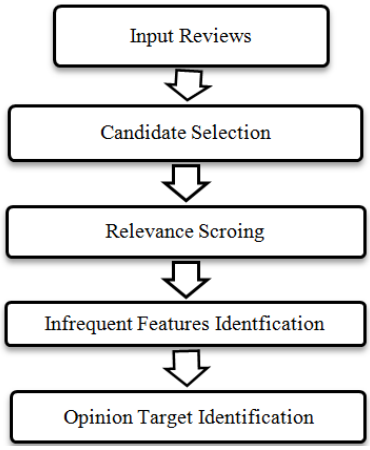

Fig. 3: Opinion targets extraction (Ferreira *et al*., 2008; Yi *et al*., 2003)

candidate features is based on noun phrase patterns. The following patterns are employed in this study.

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 study "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 study "the" and a verb are mostly observed as features.

Relevance scoring: Yi *et al*. (2003) presented unsupervised technique for relevance scoring of candidate features. This study 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:

$$r_1 = \frac{n_1}{n_1+n_2} \tag{4}$$

$$r_2 = \frac{n_3}{n_3+n_4} \tag{5}$$

Thus the combined ratio is calculated as:

$$r = \frac{n_3}{n_1+n_2+n_3+n_4} \tag{6}$$

Hence to normalize the ratios with log:

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

Hence the likelihood ratio is calculated as below:

$$-2\ log = \begin{cases} -2 * lr\ if\ r_2 < r_1 \\ 0, if\ r_2 \geq r_1 \end{cases} \tag{8}$$

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

Likelihood approach by Ferreira *et al.* (2008): A more extensive study of the LRT approach for opinion target identification is presented by this study. As mentioned in the previous sub section, the LRT was employed by Yi *et al*. (2003); 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 Hu and Liu (2004). Furthermore, they have modified the algorithm using subsequent similarity measures based on the following two rules.

Identification of feature boundaries for patterns: The earlier study (Yi *et al*., 2003) 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 study (Ferreira *et al*., 2008) 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 Ferreira *et al*. (2008) which have improved the results. Another main contribution of this study is the new annotation scheme of the features in the existing dataset that were originally employed by Hu and Liu (2004). According to the revised annotation scheme, the number of features was increased as their focus was on all features.

COMPARATIVE ANALYSIS

This section describes the analysis of the unsupervised approaches that has been potentially employed for opinion targets extraction. As explained in above section there are most popular used techniques that have been employed for opinion targets extraction. Table 1 provides the summary of the existing approaches.

Datasets: This section describes the datasets that have been used for the analysis and evaluation in this study. In this study, 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 Hu and Liu (2004). The datasets are freely available from the authors' website[1] (http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html) 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 40 GB) and cell phone (Nokia 6610). The summary of each dataset is given in Table 2: 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.

Experimental setup: Although the results are of the aforementioned techniques have been already given in the respective studies and there is no need to reproduce it. However in order to empirically prove the factors affecting the existing approach the following tools and experimental setup is employed.

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.

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 Combination Base Noun Phrase (cBNP). The first four patterns have already been employed as discussed in above section. While the cBNP pattern is a novel hybrid patterns 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 text STAT software are used.

Table 1: Comparative summary of opinion targets extraction techniques

Approach	Candidate selection	Machine learning technique	Requires vocabulary	Can extract multi word features	Using semantic
Yi *et al*. (2003)	Noun phrases with restricted patterns	Likelihood ratio test	Yes	Partially	No
Hu and Liu (2004)	Noun phrases with nearest adjectives	Association mining	No	Yes	No
Ferreira *et al*. (2008)	Noun phrases with restricted patterns	Likelihood ratio test	Yes	Yes	No
Wei *et al*. (2010)	Noun phrases	Association mining	Yes	Yes	Yes

Table 1: Summary of the five product datasets with manually tagged opinion targets by Hu and Liu (2004)

	Dataset				
Description	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

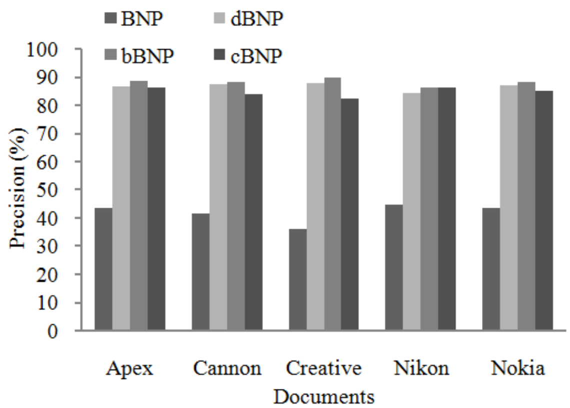

Fig. 4: Comparative percentage of precision of the five datasets based on patterns

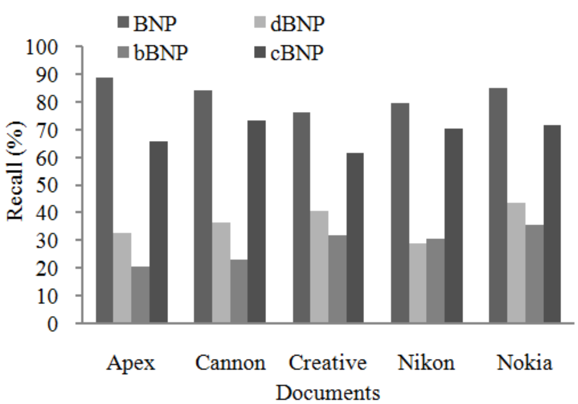

Fig. 5: Comparative percentage of recall of the five datasets based on patterns

The Stanford part of speech tagger is employed for part of speech tagging (Toutanova *et al*., 2003). This software is freeware and has been widely reported for the part of speech tagging of English language texts[1] (http://nlp.stanford.edu/software/tagger.shtml).

Text Stat 3.0 is employed for pattern extraction and test analysis. This software is open source and freely available for academic research from the author's website[1] (http:// neon.niederlandistik.fuberlin.de/en/ texttat/). This software is simple and has been used by a number of works for searching number of works for searching terms and strings in English texts

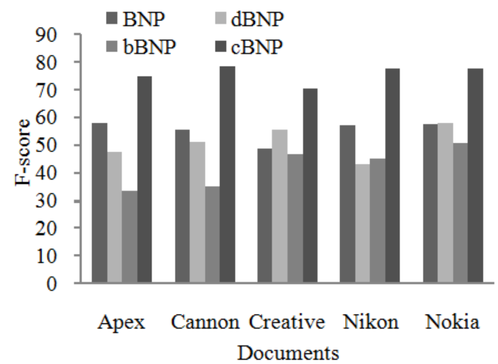

Fig. 6: Comparative percentage of F-score of the five datasets based on patterns

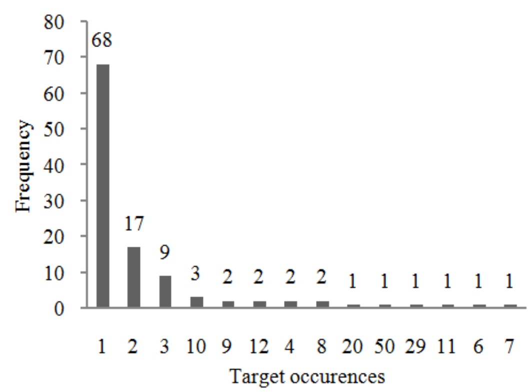

Fig. 7: Target occurrences in apex dataset

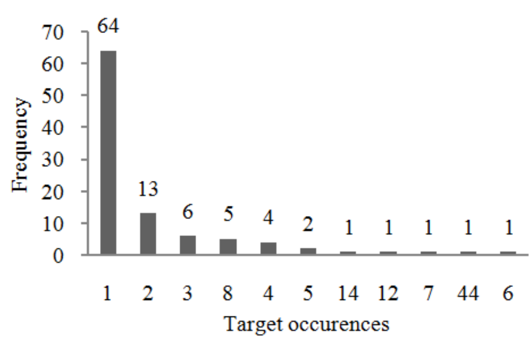

Fig. 8: Target occurrences in canon dataset

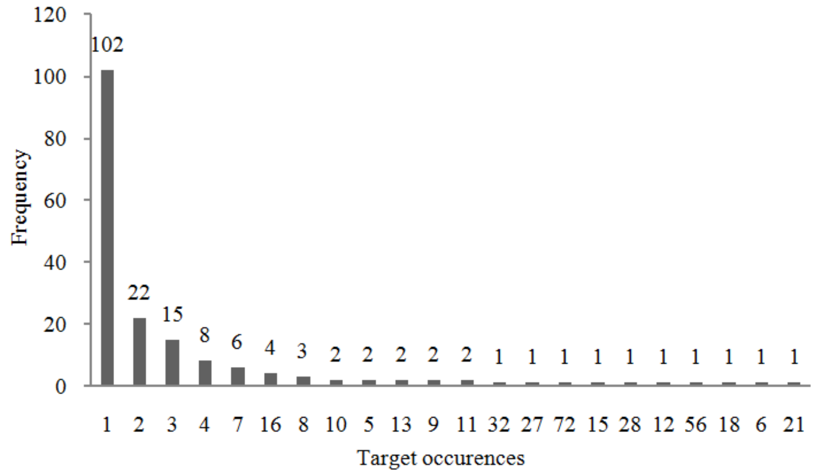

Fig. 9: Target occurrences in creative dataset

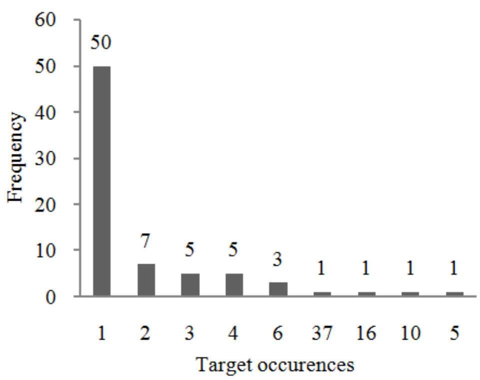

Fig. 10: Target occurrences in Nikon dataset

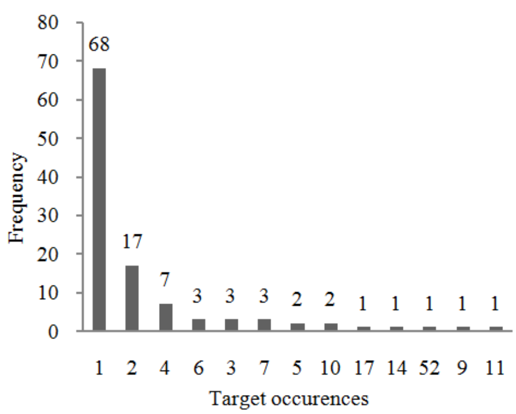

Fig. 11: Target occurrences in Nokia dataset

(Diniz, 2005). This software accepts any type of regular expression to extract sub strings from a corpus or text documents.

For comparison we use the precision, recall and f-score as measure of accuracy. The comparative results are shown in Fig. 4 to 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.

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, the histogram of the opinion target distribution in each dataset is created through Text STAT and Excel software as shown in Fig. 7 to 11. The layouts of all the graphs are similar where the x-axis shows features and y-axis shows target occurrences in the dataset.

Figure 7 shows the frequency distribution of product features in the Apex dataset which contains a total of 347 target features out of which 194 features have a frequency of less than 2, i.e., 55.90% of the features have a low frequency, which would be classified as irrelevant features.

Figure 8 shows the frequency distribution of the target features in the Canon dataset, which contains a total of 257 target features out of which 83 features have a frequency of only one, i.e., 32.29% of the features have a low frequency, which would be classified as irrelevant features.

In Fig. 9, the graph shows the frequency distribution of the target features in the Creative dataset. This dataset contains a total of 736 target features with 287 targets having a frequency of only one, i.e., 38.99% have a low frequency, which would be classified as irrelevant features.

Similarly, Fig. 10 shows the frequency distribution of the target features in the Nikon dataset, which contains a of total 185 target features out of which 68 features have a frequency of only one, i.e., 36.76% of the features have a low frequency, which would be classified as irrelevant features.

In Fig. 11, the graph shows the frequency distribution of the target features in the Creative dataset. This dataset contains a total of 310 target features with 103 targets having a frequency of only one, i.e., 33.23% have a low frequency, which may be classified as irrelevant features.

Hence, from the above discussion it is clear that even a large dataset has features with low frequencies thus cannot be predicted by the relevance scoring technique.

CONCLUSION

This study presents a systematic review of unsupervised approaches of opinion target identification from unstructured reviews. This study shows that besides the significant improvements in the accuracy of opinion target identification, the problem of infrequent feature identification is not completely solved due to the dependency on a threshold value. The frequency measure also affects the accuracy as a review can have multiple topics under discussion. During this study it has been pointed out that the results of candidate selection cand be improved with the boundary conditions on the patterns. The analysis also shows that the target extraction is greatly affected by threshold values and the infrequent features cannot be detected by simply using the distribution similarity.

REFERENCES

Agrawal, R. and R. Srikant, 1994. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases. Morgan Kaufmann Publishers Inc., September 12-15, pp: 487-499.

Aurangzeb, K., B. Baharum and K. Khairullah, 2011a. Sentiment classification from online customer reviews using lexical contextual sentence structure. Proceeding of ICSECS 2011. Springer-Verlag, Berlin, Heidelberg, Pahang, Malaysia.

Aurangzeb, K., B. Baharum and K. Khairullah, 2011b. Sentiment Classification Using Sentence-level Lexical Based Semantic Orientation of Online Reviews. T. Appl. Sci. Res., 6(10): 1141-1157.

Baharudin, B., L.H. Lee and K. Khan 2010. A review of machine learning algorithms for text-documents classification. J. Adv. Inform. Technol., 1(1): 4-20.

Baharum, B. and K.K. Baharudin, 2010. Automatic Extraction of Features and Opinion Oriented Sentences from Customer Reviews. World Acad. Sci. Eng. Technol., 4(62): 457-461.

Baharum, B. and K. Khairullah, 2011. Mining customer data for decision makingusing new hybrid classification algorithm. J. Theor. Appl. Inform. Technol., 27(1): 54-61.

Balahur, A. and A. Montoyo, 2008. A feature dependent method for opinion mining and classification. Proceeding of International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE '08), pp: 1-7.

Ben-David, S., J. Blitzer, K. Crammer and F. Pereira, 2007. Analysis of representations for domain adaptation. Adv. Neural Inform. Process. Syst. Vol. 19.

Blitzer, J., M. Dredze and F. Pereira, 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, Czech Republic, pp: 440-447.

Bloom, K., N. Garg and S. Argamon, 2007. Extracting appraisal expressions. Proceedings of the Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics. Rochester, New York, USA., April 2007, pp: 308-315.

Carenini, G., R.T. Ng and E. Zwart, 2005. Extracting knowledge from evaluative text. Proceedings of the 3rd International Conference on Knowledge Capture. ACM Banff, Alberta, Canada, October 2-5, pp: 11-18.

Diniz, L., 2005. Comparative review: TextStat 2.5, ANTCONC 3.0 and compleat lexical tutor 4.0. Lang. Learn. Technol., 9(3): 22-27.

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

Ferreira, L., N. Jakob and I. Gurevych, 2008. A comparative study of feature extraction algorithms in customer reviews. Proceedings of the IEEE International Conference on Semantic Computing. Santa Clara, CA, August 4-7, pp: 144-151.

Goujon, B., 2011. Text mining for opinion target detection. Proceedings of the European Intelligence and Security Informatics Conference (EISIC). Athens, Greece, September 12-14, pp: 322-326.

Holzinger, W., B. Krupl and M. Herzog, 2006. Using ontologies for extracting product features from web pages. Proceedings of the 5th International Semantic Web Conference. Athens, Georgia, USA, pp: 286-299.

Hu, M. and B. Liu, 2004. Mining and summarizing customer reviews. Proceeding of 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Seattle, WA, USA, pp: 168-177.

Khan, K. and B.B. Baharudin, 2012. Identifying product features from customer reviews using lexical concordance. Res. J. Appl. Sci. Eng. Technol., 4.

Khan, K., B. Baharudin and A. Khan, 2009. Mining opinion from text documents: A survey. Proceeding of 3rd IEEE International Conference on Digital Ecosystems and Technologies (DEST '09). Istanbul, pp: 217-222.

Khan, A., B. Baharudin and K. Khan, 2010a. Efficient feature selection and domain relevance term weighting method for document classification. Proceeding of 2nd International Conference on Computer Engineering and Applications (ICCEA), 2: 398-403.

Khan, A., B. Baharudin and K. Khan, 2010b. Semantic based features selection and weighting method for text classification. Proceeding of International Symposium in Information Technology (ITSim), 2: 850-855.

Lu, Y. and C. Zhai, 2008. Opinion integration through semi-supervised topic modeling. Proceedings of the 17th International World Wide Web Conference. Beijing, China, April 21-25, pp: 121-130.

Pang, B. and L. Lee, 2008. Opinion mining and sentiment analysis. Foundations Trends Inform. Retrieval, 2(1-2): 135.

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

Qiu, G., F. Zhang, J. Bu and C. Chen, 2009. Domain specific opinion retrieval information retrieval. Proceedings of the 5th Asia Information Retrieval Symposium. Sapporo, Japan, October 21-23, pp: 318-329.

Somprasertsri, G. and P. Lalitrojwong, 2010. Mining feature-opinion in online customer reviews for opinion summarization. J. Univers. Comput. Sci., 16: 938-955.

Toutanova, K., D. Klein, C.D. Manning and Y. Singer, 2003. Feature-rich part-of-speech tagging with a cyclic dependency network.. Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Edmonton, Canadada, May 27-June 01, 1: 173-180.

Wei, C.P., Y.M. Chen, C.S. Yang and C.C. Yang, 2010. Understanding what concerns consumers: A semantic approach to product features extraction from consumer reviews. Inform. Syst. E-Bus Manage., 8: 149-167.

Wong, T.L. and W. Lam, 2009. An unsupervised method for joint information extraction and feature mining across different web sites. Data Knowl. Eng., 68: 107-125.

Yi, J., T. Nasukawa, R. Bunescu and W. Niblack, 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. Proceedings of the 3rd IEEE International Conference on Data Mining. Washington, DC, USA, November 19-22, pp: 427-434.

Zhai, Z., B. Liu, H. Xu and P. Jia, 2011. Clustering product features for opinion mining. Proceedings of the fourth ACM International Conference on Web Search and Data Mining. Hong Kong, China, February 9-12, pp: 347-35.

Zhang, L. and B. Liu, 2011. Identifying noun product features that imply opinions. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. June 19-24, Portland, pp: 575-580.