Expert Systems With Applications 89 (2017) 273–285

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A two-fold rule-based model for aspect extraction

Toqir A. Rana*, Yu-N Cheah

School of Computer Sciences, Universiti Sains Malaysia, Malaysia

ARTICLE INFO

Article history:
Received 17 May 2017
Revised 26 July 2017
Accepted 27 July 2017
Available online 28 July 2017

Keywords:
Aspect-based sentiment analysis
Opinion mining
Aspect extraction
Explicit aspects
Sequential pattern-based rules
Aspect pruning

ABSTRACT

Opinion target extraction or aspect extraction is the most important subtask of the aspect-based sentiment analysis. This task focuses on the identification of the targets of user's opinions or sentiments from online reviews. In the recent years, syntactic patterns-based approaches have performed quite well and produced significant improvement in the aspect extraction task. However, these approaches are heavily dependent on the dependency parsers which produced syntactic relations following the grammatical rules and language constraints. In contemporary, users do not give much importance to these rules and constraints while expressing their opinions about particular product and neither reviewer websites restrict users to do so. This makes syntactic patterns-based approaches vulnerable. Therefore, in this paper, we are proposing a two-fold rules-based model (TF-RBM) which uses rules defined on the basis of sequential patterns mined from customer reviews. The first fold extracts aspects associated with domain independent opinions and the second fold extracts aspects associated with domain dependent opinions. We have also applied frequency- and similarity-based approaches to improve the aspect extraction accuracy of the proposed model. Our experimental evaluation has shown better results as compared with the state-of-the-art and most recent approaches.

1. Introduction

Continuous expansion of World Wide Web (WWW) and tremendous growth of social media networks have rehabilitated the life style of a common person. Users of the Internet in general, and social media networks in particular, are turning to these platforms to share their experiences, emotions and feelings about different events, places or products in the form of reviews. In other words, the Internet and social media have become the prime source for decision making and information gathering rather than conducting traditional surveys. People are increasingly relying on the experiences of other users to buy new product, to travel, or to choose an appropriate hotel. However, it is almost impossible for any user to read all such reviews and make a right decision. Therefore, sentiment analysis plays a vital role in analyzing these reviews and producing an overall summary of these reviews (Cambria, 2016).

Among the different granularity levels of sentiment analysis, aspect-based sentiment analysis has attracted a large number of researchers (Liu, 2012; Pang & Lee, 2008). Aspect-based sentiment analysis deals with the users' sentiments or opinions and the targets of these opinions which are often referred to as aspects, and generates an overall summary of these aspects along with the positive or negative polarity of users' opinions towards that specific aspect. Among different tasks of aspect-based sentiment analysis, aspect extraction is the most important task and studied by a large number of researchers (Rana & Cheah, 2016a). This task involves how to identify aspects which are related to the specific entity and what are the users' opinions which are related to those aspects.

Hu and Liu (2004) have identified two types of aspects: explicit and implicit. Explicit aspects are those aspects which are expressed explicitly. For example, in the sentence: "The phone is great", "phone" is the aspect. While in this sentence: "Phone is small", the user is talking about the "size" of the phone but did not mentioned it explicitly in this sentence, and such aspects are called implicit aspects. Most of the researchers have focused on explicit aspects and only a very small number of studies have focused on the implicit aspects (Rana & Cheah, 2016a).

In recent years, linguistic patterns-based approaches have been widely studied for the extraction of explicit aspects (Bancken, Alfarone, & Davis, 2014; Du, Chan, & Zhou, 2014; Kang & Zhou, 2016; Liu, Gao, Liu, & Zhang, 2015, Liu, Gao, Liu, & Zhang, 2016; Liu, Liu, Zhang, Kim, & Gao, 2016; Poria, Cambria, Ku, Gui, & Gelbukh, 2014; Qiu, Liu, Bu, & Chen, 2011). Most of these studies used dependency parser-based approaches to explore relations among aspects and opinions. These dependency parsers heavily depend on the grammatical rules and language constraints. On the other hand, reviewer websites do not restrict users to follow these rules and

* Corresponding author.
E-mail addresses: toqirr@gmail.com (T.A. Rana), yncheah@usm.my (Y.-N. Cheah).

http://dx.doi.org/10.1016/j.eswa.2017.07.047
0957-4174/

therefore, users tend to write in casual manner and sometimes violate such rules. For example: "same story as everybody else when trying to get service from apex - nothing". In this sentence, "service" is the aspect which is associated with the opinion word "nothing", but dependency parsers are unable to relate them. Machine learning approaches, like conditional random fields (CRF), have also been applied to solve the problem (Chen, Qi, & Wang, 2012; Choi & Cardie, 2010; Huang, Liu, Peng, & Niu, 2012; Jakob & Gurevych, 2010; Yang & Cardie, 2013). However, CRF-based approaches are supervised and required a large number of aspects to perform well.

In this paper, we have proposed a two-fold rule-based model (TF-RBM) which performs the task in three steps; (1) using sequential patterns-based rules (SPR) (Rana & Cheah, 2015, 2016c) for the extraction of explicit aspects which are associated with regular opinions; (2) improving aspect extraction accuracy with a frequency-based approach along with normalized Google distance (NGD) (Cilibrasi & Vitanyi, 2007); and (3) in the second fold, extracting aspects which are associated with domain dependent opinions, we called this phase as concept extraction. Although, WordNet (Fellbaum, 1998) and Word2Vec[1] similarity measures have been used by Kang and Zhou (2016) and Liu et al. (2016) respectively, but both of these tools require a huge corpus and trained model. On the other hand, NGD does not require any corpus or trained model. NGD uses Google as the search engine and calculates the similarity between two terms on the basis of the hits returned by Google.

The proposed approach first uses SPRs to extract all aspects and opinions. Although, a large number of aspects can be extracted, many are not related to the product. To overcome this issue, pruning is carried out by calculating the frequency of each word and eliminating all aspects which do not fulfill a minimum threshold. However, not all the non-frequent aspects are irrelevant. To find the non-frequent, but relevant, aspects, we calculate the NGD score for only those aspects which are nouns and select those which have NGDs lower than the minimum score, even though they are not frequent as proposed in Rana and Cheah (2018). In the second fold, we search for aspects which are not associated with any regular opinion, but are associated with some domain dependent opinions e.g., small, large, tiny. We have compared our approach with the state-of-the-art and most recent approaches. Experimental results of the proposed model have shown the improvement in the aspect extraction phase.

The remaining of the paper is organized as follows: Section 2 outlines the related work. In Section 3, the proposed methodology for aspect extraction is discussed. In Section 4, the experimental results are presented while Section 5 concludes the paper.

2. Related work

2.1. Unsupervised approaches

Aspect extraction in aspect-based sentiment analysis was first studied by Hu and Liu (2004). They considered nouns as potential aspects and calculated the frequency of each aspect. They used association rule miner (CBA) (Liu, Hsu, & Ma, 1998) for the generation of candidate aspects. CBA generates frequent noun phrases without considering the position of each noun in the sentence. Therefore, they used aspect pruning to eliminate aspect phrases where the noun words were not closely related. Furthermore, the frequency of each aspect was calculated and only those aspects were selected which had the support greater than minimum given threshold. For each selected aspect, the nearest adjective was considered as the opinion and these opinions were further used to identify the non-frequent aspects. Popescu and Etzioni (2007) later improved the accuracy by calculating Pointwise Mutual Information (PMI) for each aspect. If the respective calculated PMIs were too low, those aspects were eliminated. Li, Zhou, and Li (2015) proposed a web-based similarity method along with frequency pruning method to improve the aspect extraction accuracy. They adopted the same approach for frequency pruning as proposed by Hu and Liu. For web-based similarity, number of hits returned by a search engine, for entity term and aspect terms, were used to evaluate similarity among two terms. They calculated the PMI-IR score and those aspects were eliminated which have a score less than the given threshold.

Many researchers followed Hu and Liu's work. Li, Zhang, Ma, Zhou, and Sun (2009) combined NLP with statistical methods to identify aspects from Chinese online reviews. Raju, Pingali, and Varma (2009) proposed a three step clustering-based approach to identify product aspects. Their approach grouped similar nouns/noun phrases into clusters and these clusters were used to identify aspects.

Reviewing websites provide additional information related to the customer reviews e.g., rating. Such additional information was collected from the review sites to identify product aspects (Meng & Wang, 2009; Moghaddam & Ester, 2010). Eirinaki, Pisal, and Singh (2012) also extracted aspects considering the most frequent aspects and Marrese-Taylor, Velásquez, and Bravo-Marquez (2014) also adopt a similar approach for the tourism domain. Other than these approaches, bootstrapping techniques were explored by Bagheri, Saraee, and De Jong (2013), Li, Qin, et al. (2015) and Zhu, Wang, Zhu, Tsou, and Ma (2011). Gunes, Furche, Shoreditch, and Orsi (2016) proposed a frequency-based noun-clustering method to detect structured aspect terms. Quan and Ren (2014) used PMI along with document frequency to explore the association between aspects and opinions. Yang, Liu, Lin, and Lin (2016) combined local context information, i.e., within a sentence, and global context information, i.e., within multiple sentences in a document, for aspect extraction and ranked them on the basis of their score and frequency.

2.2. Semi-supervised approaches

Wang and Wang (2008) and Hai, Chang, and Cong (2012) used a bootstrapping method to learn both product aspects and opinions from Chinese customer reviews. Wei, Chen, Yang, and Yang (2010) followed a semantic-based approach which identified the most frequent aspects and further eliminated all aspects which were irrelevant to the seed opinion words. Ma, Zhang, Yan, and Kim (2013) combined Latent Dirichlet Allocation (LDA) with a lexicon of synonyms to extract aspects from Chinese reviews. Liu, Xu, et al. (2015) proposed a word alignment-based model which identified aspects from the position of the words in the sentence. Yan, Xing, Zhang, and Ma (2015) used association among aspects and opinion words and used a synonym lexicon to expand the aspect list. Samha, Li, and Zhang (2014) built a list of aspects, from the information provided by the manufactures, to identify similar aspects from customer reviews.

2.3. Supervised approaches

A dictionary-based supervised approach was proposed by Kobayashi, Inui, and Matsumoto (2007) to identify aspects and syntactic patterns were used to find associations among aspects, opinions and products. Cruz, Troyano, Enríquez, Ortega, and Vallejo (2013) generated a taxonomy for aspects, specific for every domain, and validated the opinions by the assumption that opinions appeared somewhere near to aspects in the sentence.

[1] https://code.google.com/p/word2vec/

Jin, Ho, and Srihari (2009) proposed a lexicalized HMM-based model for the extraction of aspects and opinions from customer reviews. Other frameworks were based on CRFs proposed by Chen et al. (2012), Choi and Cardie (2010), Huang et al. (2012), Jakob and Gurevych (2010), Li et al. (2010) and Yang and Cardie (2013). Li, Wang, and Zhou (2012) used several heuristic rules based on binary classifier to map the opinion targets from the parse tree of the sentence. Poria, Cambria, and Gelbukh (2016) combined neural networks with a set of linguistic patterns for aspect extraction. They used a 7-layer deep convolutional neural network and combined linguistic patterns to tag each word in a sentence as an aspect. Rana and Cheah (2016b) used manually built list of class labeled patterns to identify aspects which have an associated opinion word but are objective in nature and did not symbolize users' opinion. The supervised approaches require manually built aspects list and depends heavily on the trained datasets to perform well.

2.4. Topic modeling-based approaches

Topic modeling-based approaches have been widely studied in the recent years (Rana, Cheah, & Letchmunan, 2016). These approaches are mainly based on LDA and search the document for any potential relevant topics while very few approaches have employed Probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 2001) e.g., Moghaddam and Ester (2011), where they extended pLSA by incorporating latent rating information of reviews.

Brody and Elhadad (2010) treated each sentence in the document as a separate document and applied the traditional LDA approach. The topics generated from each sentence were considered as aspects. Zhao, Jiang, Yan, & Li (2010) proposed a hybrid approach which jointly used maximum entropy and topic modeling to identify aspects and opinions. Jo and Oh (2011) proposed two models to identify aspects and grouping them by considering that all aspects in single sentence represent the same topic. For aspect and aspect-dependent sentiment extraction, a Joint Aspect/Sentiment (JAS) by extending the LDA, was proposed by Xu, Tan, Liu, Cheng, and Lin (2012). Lin et al. (2014) also proposed to jointly extract aspects and sentiments. They assumed that, different reviews shared the same topic distribution in different languages. This assumption helped to identify different topics in the reviews. Mukherjee and Liu (2012) used seed opinion words to jointly extract and categorize aspects. Bagheri, Saraee, and De Jong (2014) proposed an aspect extraction model by considering each word in the sentence as a state of a Markov chain (Gruber, Weiss, & Rosen-Zvi, 2007). Shams and Baraani-Dastjerdi (2017) proposed Enriched LDA model which incorporates the aspect co-occurrence relations as prior domain knowledge into LDA for aspect extraction. Poria, Chaturvedi, Cambria, and Bisio (2016) presented Sentic LDA which integrates common-sense in LDA to calculate the word distributions.

In the product review domain, multiple products may share the same aspect, e.g., "price". Therefore, Chen et al. (2013) proposed a knowledge-based topic model using the must-link (aspects in the same document) and cannot-link (aspects in different documents) state. On the basis of these knowledge-based systems, a self-learning model was proposed (Chen, Mukherjee, & Liu, 2014) which learns the knowledge from multiple domains which share the same kind of aspects. This self-learning model was followed by Chen and Liu (2014b) to propose a lifelong topic model which extracts the knowledge from multiple domains and then learn from the extracted knowledge to identify aspects. This work only focused on the must-link state. To handle, both, must- and cannot-link states, a lifelong learning model, which covered both links, was proposed (Chen & Liu, 2014a). Wang, Chen, and Liu (2016) proposed two topic modeling-based models: holistic fine-grained topic model and second model is based upon lifelong learning model (Chen & Liu, 2014b). This two-model approach learns knowledge from multiple domains to assist the target domain.

LDA distribution is unable to capture correlations among topics. Therefore, topic modeling-based techniques find topical terms related to the topic but each individual word may not be an aspect and hence reduced the accuracy of the system. For example, "time" and "day" are the topical terms for "battery life" but have no relation with the "phone".

2.5. Syntactic pattern or rule-based approaches

Wu, Zhang, Huang, and Wu (2009) introduced a phrase dependency parser-based approach which generates a tree representing relations among different phrases in the sentence. Similarly, Yu, Zha, Wang, Wang, and Chua (2011) used a list of aspects generated from pros and cons reviews to identify aspects, from free text reviews, which were ranked by most of the users. Qiu et al. (2011) and Qiu, Liu, Bu, and Chen (2009) proposed a double-propagation (DP) algorithm which used dependency parser-based rules to extract aspects. This algorithm used a small set of opinion words to find associated aspects and used these aspects to find new opinion words. The newly find opinion words were added to the seed opinion list and process was repeated again until no new aspect or opinion was identified. Zhang, Liu, Lim, and O' Brien-Strain (2010) improved DP by adding new syntactic rules. DP performs well on small datasets but its performance starts decreasing as the size of dataset increased. Liu et al. (2016) used DP approach as baseline and improved the aspect extraction by similarity- and association-based recommendations. They used Word2Vec to train their model from a large number of reviews and selected aspects with maximum similarity score. The association-based recommendation selects aspect by exploring association among aspects and opinion with the help of association rule mining algorithm (Agraval & Srikant, 1994).

Many approaches have exploited the relationship between aspects and opinions using phrase dependency parser (Wu et al., 2009; Yu, Zha, Wang, & Chua, 2011) which generates a tree to explore connections among different words in a sentence. Liu, Xu, Liu, and Zhao (2013) proposed a partially supervised model which used syntactic patterns, with high frequency and low recall, along with a word alignment model. Unlike other syntactic patterns-based approaches, their model utilized the advantages of the word alignment model i.e., position of the words, frequencies of words co-occurrences, etc. Xu, Liu, Lai, Chen, and Zhao (2013) used a graph-based algorithm to find the syntactic patterns between aspects and opinions. Moreover, Liu, Xu, and Zhao (2012) used a word-based translation model, Htay and Lynn (2013) defined a number of patterns to identify relations among aspects and opinions and (Hai, Chang, Kim, & Yang, 2014) defined certain rules for Chinese language reviews. Liu, Xu, et al. (2015) and Liu et al. (2016) proposed an automated rule selection approach, based on dependency tree parsing, to select the most appropriate rules for every aspect. These rules selection algorithms required extensive computational cost to select the most appropriate rule for every aspect. Kang and Zhou (2016) used dependency parser-based rules to extract explicit aspects and combined frequency and semantic similarity-based techniques for aspect pruning to improve aspect extraction accuracy.

The dependency parsers produce dependency relations among words by analyzing the grammatical structure of the sentence. On the other hand, in online reviews, sentences with the same structure can represent different meaning. For example: "The phone has a great screen" and "The screen has a great resolution". These both sentences are different but have the same grammatical structure and hold the same aspect "screen". The dependency parsers will generate dependency relation between "great" and "screen" for the

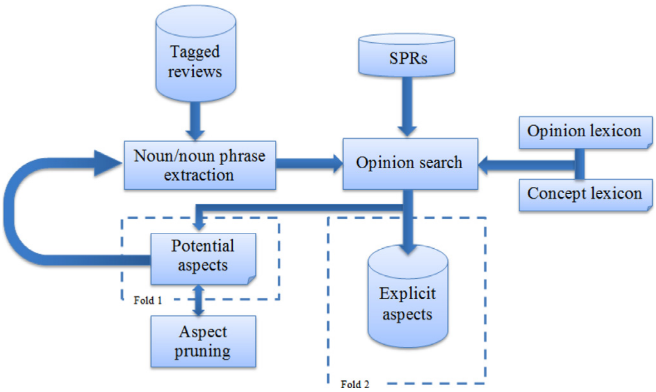

Fig. 1. Two-fold rule-based model.

first sentence and "great" and resolution" for the second sentence. While in the second sentence "great" is expressed against aspect "screen". However, the dependency parsers-based approaches are unable to distinguish between these two sentences and identified "screen" and "resolution" as product aspects respectively.

3. Proposed methodology

The proposed methodology performs aspect extraction task in three steps as shown in Fig. 1: (1) aspect extraction; (2) aspect pruning; and (3) concept extraction.

In the first fold, SPRs (Rana & Cheah, 2016c) are been used for the extraction of nouns/noun phrases as aspects and associated opinions using opinion lexicon. Only those nouns/noun phrases which are associated with an opinion are considered as potential aspects. Frequency and similarity-based approaches are applied to eliminate non-frequent and irrelevant aspects. In the second fold, the proposed model searches aspects associated with the domain dependent opinions using the concept lexicon.

3.1. Aspect extraction

One of the key tasks for aspect-based sentiment analysis is to identify appropriate opinion targets. Most of the researches have considered the nouns/noun phrases as the potential aspects in any sentence. This task heavily depends on the accuracy of the part-of-speech (POS) taggers. These POS taggers tagged each word of the sentence according to its relationship with the adjacent or related words. In customer reviews, users do not usually follow grammatical or language rules which makes the tagging process vulnerable. For example, consider the customer review: "excellent sound quality". This sentence clearly holds an opinion "excellent" which has a target "sound quality". However, with the Stanford NLP parser[2] following is the tagged output:

excellent/JJ sound/JJ quality/NN

In this sentence, "sound" is tagged as an adjective (JJ) which is the part of the noun (NN) aspect. Similarly, in the review: "the vibrate setting is loud", "loud" is the opinion for aspect "vibrate setting" associated with a copula verb (VBZ) "is" while "the" is the determiner (DT), but the parser tagged "vibrate" as an adjective as shown in the following tagged sentence:

the/DT vibrate/JJ setting/NN is/VBZ loud/JJ

Therefore, it is not sufficient to solely rely on nouns for selecting appropriate potential aspects. From the above examples, it is quite clear that the adjectives which are actually part of the aspect are not opinion words. Hence, for the selection of noun phrases, which are the potential aspects, we have adopted the assumption that if any adjective is associated with any noun with direct association and that adjective is not in the opinion lexicon[3] then that adjective will be the part of the noun phrase to which the associated noun belongs.

On the basis of above-mentioned explanation, Algorithm 1 elaborates the noun/noun phrase extraction process. With the provided input sets of tagged reviews R, set sequential pattern-based rules P and opinion lexicon O, the expected output of the algorithm is a set of all nouns/noun phrases M. In the Algorithm 1, we have taken an assumption for direct association as shown in Fig. 2.

If two nouns are associated with some preposition like "of", then the two nouns are directly associated. For example: "I like the quality of picture". In this sentence, "picture quality" is the aspect but extracting "quality of picture" as a single aspect makes sense. Similarly, if any opinion is associated with more than two nouns then both the nouns are extracted as single noun phrase. For example: "The screen has a great resolution". This sentence holds an

Algorithm 1 Noun/noun phrase extraction.

```
Input: Set of tagged reviews R, set of SPRs P, set of seed opinion words O
Output: M
M = {} // Initially M is an empty set
for each sentence S in review R do
  for each word w_i in S do
    if w_i is noun and w_{i-1} is adjective and w_{i-1} ∉ O then
      n = w_{i-1} + w_i //adjective is part of the noun phrase n
    else if w_i is noun then
      n = w_i //n is single noun
    end if
    for each n directly associated with w_{i+1} where w_{i+1} is noun do //direct
      association
      n = n + w_{i+1} //identify noun phrase
      w_i = w_{i+1}
    end for
    //Search for any opinion word w_i ∈ O associated with n using P
    if any w_i found where w_i ∈ O then
      if w_{i+1} is noun then
        n = n + w_{i+1} //opinion is associated with more than one noun/noun
      phrase
      end if
        M ← n //add n to M as potential aspect
    end if
    w_i = w_{i+1}
  end for
end for
Output: Set of noun/noun phrases M
```

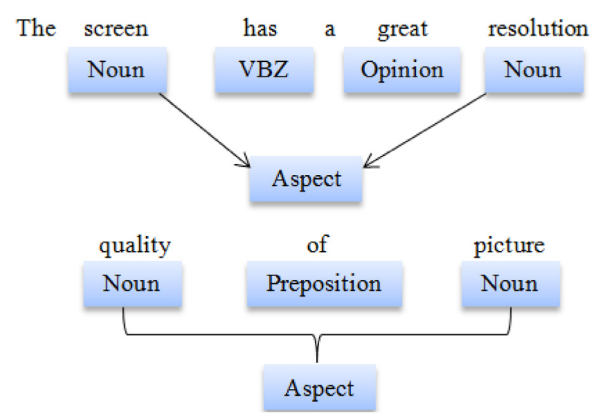

Fig. 2. Selecting opinion targets.

[2] http://nlp.stanford.edu/software/lex-parser.shtml

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

opinion "great" which is associated with "screen" and "resolution". Algorithm 1 will identify "screen resolution" as an aspect associated with the opinion "great". This makes sense as the user has expressed his/her opinion on the "resolution" which is the property of "screen" and by identifying the "screen resolution", the true target of users' opinion has identified.

The algorithm starts with reading the tagged review sentences. The algorithm searches for any word w_i which is tagged as a noun, if any noun word is identified then it checks whether any previous word w_{i-1} is an adjective or not. If any adjective is found and it is not in opinion list O, then w_{i-1} will be a part of the noun phrase n, or else the noun word w_i is considered as a common noun n. The n represents as noun/noun phrase in this algorithm. The next step is to identify any noun phrase by identifying any word w_{i+1} which is noun and directly associated with n. The process continues until no new w_{i+1} is identified as a noun. Once any noun or noun phrase is identified, then the algorithm looks for any opinion word w_i which is in opinion list O according to the rules in P. If some noun/noun phrase is associated with the opinion w_i, then the algorithm looks for the possibility of any other noun w_{i+1} associated with the opinion. If any noun w_{i+1} is identified then the noun is added to n and finally n is added to the set of nouns/noun phrases M as potential aspect. The algorithm keeps on searching for any other noun/noun phrase within the sentence. The algorithm repeats itself for each of the sentences in the tagged reviews. In the end, the algorithm returns the set of all nouns/noun phrases M as set of potential aspects.

3.2. Aspect pruning

During the noun/noun phrase extraction a large number of potential aspects are extracted but not all of these aspects are relevant. The large number of potential aspects may reduce the overall accuracy of the system. To overcome this problem, we have used a two-step pruning process. In the first step, most frequent noun/noun phrases are selected and those noun/noun phrases which did not meet the minimum threshold are pruned out. The frequency of each word in extracted aspects has been calculated and only those aspect are selected which have at least one frequent word in the term. In the second step, those nouns are reselected which are semantically similar to the entity i.e., terms having NGD lower than the threshold are semantically similar. For example, for the product "mp3 player" an explicit aspect is "instructions". However, the support of this aspect is below the given threshold while the NGD score for the entity and aspect is below minimum NGD threshold. Therefore, the aspect "instructions" will be selected as potential aspect, though it is not a frequent aspect. The semantic similarity process is applied to only nouns and not applied to noun phrases.

3.2.1. Frequency pruning

To apply frequency pruning, we calculate the frequency of each noun word in every noun/noun phrase in M. Frequency is being calculated with the occurrence of every word in a sentence. If a word appears in one sentence then the frequency is 1 and if the word appears in 10 sentences the frequency will be 10. If any word appears multiple times in any sentence, it will be considered as 1. It means that, the frequency of the words is calculated on the sentence level and not on the individual extracted noun/noun phrase level. Spell checking is applied to deal with misspelled words, fuzzy matching (Cohen, 2015) is applied to deal with the words like "auto-focus" and "autofocus" and lemmatization (Fellbaum, 1998) is applied for words like "phone" and "phones" using Natural Language Toolkit (NLTK[4]).

The proposed model searches for all the possible aspect terms. Therefore, we have set the minimum threshold level is set to 2 and 3 for nouns and noun phrases respectively to identify maximum number of aspects. The threshold level has been chosen by empirical testing of the performance of different threshold levels as shown in Section 4.5. The interval for the change of threshold is set to 1000 sentences i.e., if total number of sentences are 1000 or less, then the minimum support for any word in noun phrase is 2. For each interval the minimum support will be incremented by 1 and 2 for noun phrase and noun respectively. It means that, if any noun/noun phrase contains any word with a minimum threshold then that noun/noun phrase is an aspect and if no frequent word is there in the noun/noun phrase, then that noun/noun phrase will be pruned out. The minimum threshold varies with the number of sentences, greater the number of sentences larger the threshold will be. The experimental evaluation has shown promising results by using the above assumptions.

3.2.2. Similarity pruning

The frequency pruning is applied to both nouns and noun phrases. The frequency pruning process eliminates all those nouns/noun phrases which are not able to meet the minimum threshold. For noun phrases this seems pretty logical but in the case of single word nouns, frequency pruning will eliminate all noun aspects which were not common in the review dataset either they were relevant or not. To overcome this issue, we have applied similarity matrix, using NGD, to save such nouns.

NGD calculates the similarity score by using the number of hits returned by the Google search engine. If two terms never appear on the same page then the NGD will be infinite. If two terms are always appeared on the same page then the NGD will be 0, means that the distance between two terms is zero. Besides other similarity matrices like WordNet or Word2Vec, which required a huge corpus for training and used the same corpus to calculate the similarity, NGD does not require any corpus for training and testing. NGD only requires is the number of hits returned by the search engine for every term. The following is the formula for calculating NGD (Cilibrasi & Vitanyi, 2007).

$$\mathrm{NGD}(x, y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log N - \min\{\log f(x), \log f(y)\}} \tag{1}$$

In Eq. (1), x is the first term and y is the second term. f(x) represents the number of hits returned by Google for term x, f(y) for term y and f(x, y) for searching both the terms collectively, while N is the total number of webpages. In this study, we have used product reviews for the evaluation and hence, we are interested in only those aspects which are related to the product. However, Google searches the whole web to calculate NGD irrespective the domain of the terms. Therefore, we have limited the search to Amazon[5] website because Amazon is the largest source for online products and their reviews. The search query in Google search engine will become "term site: amazon.com". This query will search for any term within the Amazon website only and return the number of hits for any term from Amazon website. Hence, NGD calculates the distance between two terms on the basis of hits returned by Google from Amazon website.

For all noun aspects, we calculate the NGD of each noun with the entity. For example, the product entity is "camera" and one of the aspect "size" was eliminated during frequency pruning step. We calculated the NGD between "camera" and "size" and if the NGD is below the given threshold then, whether such noun has the frequency lower than the minimum support, that noun will be considered as potential aspect and all other nouns, where NGD is

[4] https://www.nltk.org

[5] https://amazon.com

Algorithm 2 Aspect pruning.

Input: Set of noun/noun phrases M, set of tagged reviews R, Set of thresholds T
Output: A, W, F
A = {} // Initially A is an empty set
for each word w_i in each noun/noun phrase $n_i \in$ M **do**
  Calculate the frequency f for each w_i with respect to R
  W ← w_i
  F ← f(w_i)
end for
for each noun/noun phrase n ∈ M **do**
  if n is a noun phrase **then**
    if frequency of any word w_i in n ≥ given threshold t_k where $t_k \in$ T **then**
      A ← n
    end if
  else if n is a single word and f(n) ≥ t_k where $t_k \in$ T **then**
    A ← n
  else if n is single noun and f(n) < t_k and NGD(E, n) < 0.1 **then**
    A ← n
  else
    remove n from M
    remove f(n) from f for all words in n
  end if
end for
Output: Set of aspects A, W as a set of all words in A, F as a set of frequencies of all words in W

Algorithm 3 Concept extraction.

Input: Set of aspects A, W as a set of all words in A, F as a set of frequencies of all words in W, set of seed concepts C, set of tagged reviews R, set of thresholds T
Output: Extended set of aspects A
for each sentence S in review R **do**
  for each word w_i in S **do**
    if w_i is noun **and** w_{i-1} is adjective **and** $w_{i-1} \notin$ O **then**
      $n = w_{i-1} + w_i$ //adjective is part of the noun phrase n
    else if w_i is noun **then**
      $n = w_i$ //n is single noun
    end if
    for each n directly associated with w_{i+1} where w_{i+1} is noun **do** //direct association
      $n = n + w_{i+1}$ //identify noun phrase
      $w_i = w_{i+1}$
    end for
    if any $w_i \in$ W in n **and** f(w_i) > t_k where $t_k \in$ T **then**
      //Search for any concept word $w_i \in$ C and $w_i \notin$ O associated with n using P
      if any concept w_i is found **then**
        if w_{i+1} is noun **then**
          $n = n + w_{i+1}$ //opinion is associated with more than one noun/noun phrase
        end if
        A ← n
      end if
    if any $w_i \in$ W in n **and** no concept is found **and** a negation is associated with n **then**
      A ← n
    end if
    $w_i = w_{i+1}$
  end for
end for
Output: Set of aspects A

greater than the minimum threshold, will be eliminated. For comparing the two terms, we have set the minimum NGD to 0.1. The threshold level has been chosen by empirical testing of the performance of different threshold levels as shown in Section 4.5.

Algorithm 2 outlines the approach adopted for the aspect pruning. For this algorithm, M is the set of all nouns/noun phrases produced by Algorithm 1 and R is the set of tagged reviews used for calculating the frequency of each word from M. We have used tagged reviews because M contains adjectives and preposition as part of noun phrases. However, we are interested in only nouns. Therefore, Algorithm 2 works only for noun words and does not consider adjectives or prepositions from M. The set of threshold T contains the minimum support for each of the dataset according to their number of sentences. The outputs of the Algorithm 2 are a set of aspects A, a set of frequent words W and a set of frequencies F of the set of frequent words W.

The algorithm starts with the calculation of the frequencies of each word w_i in each noun/noun phrase n_i. Once the frequencies are calculated, the next step is to eliminate all such terms which are not frequent. Therefore, for each noun/noun phrase, if the term is a noun phrase n then the algorithm searches for any word w_i has the frequency f(w_i) higher than the given threshold t_k. If the term is a single noun n then the algorithm verifies the frequency of n. If the frequency is higher than the threshold t_k, n is an aspect and added to set A. Finally, it verifies those terms which are single nouns n where f(n) is lower than the given threshold t_k. The algorithm calculates the NGD (n, E), where E is the entity and NGD is less than 0.1 then n is an aspect and added to set A. If none of the conditions are satisfied then the algorithm removes n from the set M and the frequency of each word in n from F.

3.3. Concept extraction

During the noun/noun phrase extraction, we focused on only opinion words which are domain independent opinions, i.e., these opinions have positive or negative polarity irrespective of the domain, e.g., good, bad, and excellent. However in online reviews, users not only use regular opinion words to express their experiences, they also use domain specific opinion words, e.g., small, large, tiny. These domain specific opinions are often called as concepts. Therefore, these concepts also express some kind of users' sentiment, e.g., "small size" may refer to a positive opinion in the camera domain but a negative opinion in the TV domain. Also, negations can perform an important role in selecting the appropriate aspect. For example: "I can only hear the sound but no picture!" This review holds an aspect "picture" but no opinion or concept is directly associated with it, but still it holds a negative sentiment which is represented with a negation.

On the basis of above-mentioned explanation, the proposed approach searches for any concept or negation associated with the aspects generated after the aspect pruning phase. For concept validation, we have used the SenticNet4 (Cambria, Poria, Bajpai, & Schuller, 2016) lexicon which contains around 50,000 concepts. SenticNet4 is applied to identify those words which are not regular opinions. Therefore, if an opinion word is associated with an aspect and the opinion word is not in the seed opinion list O while the opinion word is in SenticNet4 lexicon, then the opinion is considered as domain dependent concept or opinion.

The aspect pruning phase produced three outputs, i.e., a set of aspects A, set of frequent words W and their frequencies F. As observed, not every concept is interesting and we are interested in only those concepts which express the users' opinion. Hence, concepts related to only those aspects which are the most frequent and have the frequency higher than the given threshold are selected. In our scenario, we have set the minimum support to 6 with the interval of 1000 sentences and after each interval the minimum support will be doubled. The threshold level has been chosen by empirical testing of the performance of different threshold levels as shown in Section 4.5. For the negations, there is no threshold but the aspect should have at least one frequent word from the set W.

With the above explanation, Algorithm 3 outlines the process of extracting concepts and aspects with negation. The main idea of noun/noun phrase extraction remains the same as explained in Algorithm 1. However, the purpose is to extract concepts and aspects with negation only for frequent terms. Therefore, we pro-

Table 1
Detailed description of the dataset.

Data	Product	# of sentences	# of opinionated sentences	# of non-opinionated sentences	# of explicit aspects
D1	Canon digital camera	597	238	359	237
D2	Nikon digital camera	346	160	186	174
D3	Nokia cell phone	546	265	282	302
D4	Creative MP3 player	1716	720	996	674
D5	Apex DVD player	740	344	396	296

vided the set of all frequent words, their frequencies, and set of aspects which were the output of Algorithm 2. Along with these sets, we have provided the concept lexicon and set of threshold to extract concepts with frequent aspects.

Noun/noun phrase extraction is the same in Algorithm 3 as in Algorithm 1. Once the noun/noun phrase n is extracted and it verifies the frequency of the n and if it does meet the minimum threshold then it searches for any associated concept. If any concept which satisfies any rule in P is identified then add n to the list of aspects. If no concept is identified then the algorithm searches for any negation and if any negation is identified then n is added to set of aspects A. The algorithm continues for all the sentences and returns the extended set of explicit aspects A.

4. Experimental evaluation

This section explains the experimental evaluation of the proposed approach to assess the performance of the TF-RBM and compare the results with related approaches.

4.1. Evaluation corpus

We used the customer review dataset which has been widely used by the researchers for aspect extraction task and considered as benchmark dataset[6] for evaluation (Rana & Cheah, 2016a). This dataset contains customer reviews from different products which include five electronic products (Hu & Liu, 2004). Table 1 shows the detailed information of these datasets.

All these datasets are already annotated, i.e., each sentence has been marked whether it contains any aspect or not. The sentences, which contain any aspect, are labeled with all the aspects within the sentence and also these aspects are labeled whether they are explicit or implicit. Table 1 shows total number of sentences in column three and total number of opinionated and non-opinionated sentences in column four and five respectively according to the annotated datasets. Colum six shows the total number of explicit aspects as reported by Liu et al. (2016).

4.2. Evaluation metrics

We have used precision, recall and F_1-score as evaluation metrics as these have been used by the majority of the researchers for evaluation purposes. For the comparison, we have followed two ways to compute the results as adopted by Liu et al. (2016): (1) based on multiple occurrences of each term, and (2) based on distinct occurrence of each term.

In customer reviews, some aspects are frequently mentioned by a large number of users, on the other hand, some aspects are being mentioned by a very small number of users. This scenario helped to identify the most important aspects. For example, in the camera domain, “picture” is the most common aspect and a large number of users have expressed their views on this aspect. However, very few have addressed the aspect “color”. Therefore, for (1), if any occurrence of “picture” is identified, it means that all the occurrences of this aspect has been extracted and if none of the occurrence of “picture” is identified then all the occurrences have been lost. For (2), if any occurrence of “picture” is identified then it will be considered as one extraction and if any occurrence is not extracted then it will be considered as one loss. Both the scenarios have their own importance. In (1), the most frequent aspects have been regarded with their frequency and extracting (or losing) the most frequent aspects means extracting (or losing) the most important aspects. While in (2), all the aspects are ranked with the same importance and all are treated equally.

Precision, recall and F_1-score is calculated using true positive (TP), false positive (FP) and false negative (FN) values for the system. To calculate these values, we have a set of extracted aspects A and let T be the set of manually annotated aspects from the dataset. Hence, TP will be $| A \cap T |$, FP will be $| A \setminus T |$ and FN will be $| T \setminus A |$.

For (1), we follow the same precision and recall metrics as reported by Liu et al. (2016). For (2), the following are the standard precision (Eq. (2)), recall (Eq. (3)) and F_1-score (Eq. (4)) metrics.

$$P = \frac{TP}{TP + FP} \tag{2}$$

$$R = \frac{TP}{TP + FN} \tag{3}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{4}$$

4.3. Related approaches

For experimental evaluation, we have compared our results with the state-of-the-art and most recent approaches, in two different ways. For the first comparison, same resources are used against all the approaches i.e., datasets, Stanford NLP parser and opinion lexicon, as used in Liu et al. (2016). The results are compared with Double Propagation (DP) (Qiu et al., 2011), Conditional Random Field (CRF) (Jakob & Gurevych, 2010), RSG, RSG^+, RSLS and $RSLS^+$ (Liu et al. (2016). The comparison is conducted using the both evaluation criterion i.e., based on multiple occurrences and distinct occurrences of each term.

DP is the state-of-the-art algorithm which used dependency parser-based rules to perform the task. We used CRF for the comparison as our approach, to some extent, performs supervised learning. RSG select rules from all the rules of DP using the greedy algorithm while RSG^+ did the same selection but with an extended set of rules. Similarly, RSLS select rules from all the rules of DP using the local search algorithm while $RSLS^+$ did the same selection with extended set of rules.

For the second comparison, we have used the original results as reported in the studies. The results are compared with Popescu and Etzioni (2007), DP, rule-based extraction (RubE) (Kang & Zhou, 2016) and convolutional neural networks with linguistic patterns (CNN+LP) (Poria, Cambria, et al., 2016). Popescu and DP are the state-of-the-art methods while RubE used linguistics patterns along with sematic similarity pruning methods for aspect

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

Table 2
Performance evaluation of the proposed approach.

Data	Distinct						Multiple					
	SPR		SPR + P		SPR + P + CE		SPR		SPR + P		SPR + P + CE	
	P	R	P	R	P	R	P	R	P	R	P	R
D1	0.57	0.75	0.70	0.68	0.71	0.78	0.67	0.92	0.82	0.86	0.80	0.89
D2	0.65	0.89	0.84	0.82	0.83	0.87	0.70	0.96	0.87	0.89	0.87	0.93
D3	0.72	0.79	0.92	0.73	0.90	0.80	0.79	0.95	0.93	0.92	0.92	0.93
D4	0.60	0.76	0.73	0.74	0.70	0.80	0.75	0.91	0.88	0.88	0.86	0.93
D5	0.60	0.59	0.84	0.55	0.83	0.73	0.72	0.83	0.91	0.79	0.88	0.90
Avg	0.63	0.76	**0.81**	0.71	0.79	**0.80**	0.72	**0.92**	**0.88**	0.87	0.87	**0.92**

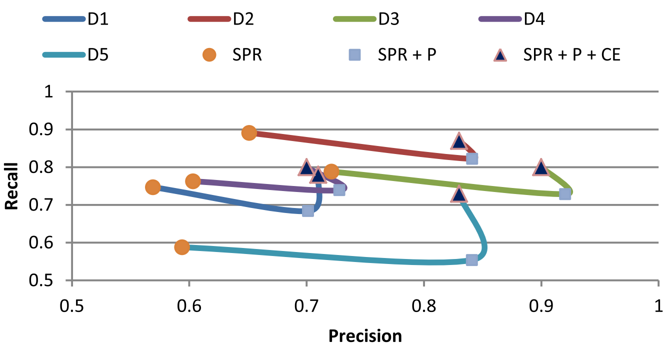

Fig. 3. Performance of the proposed approach (distinct).

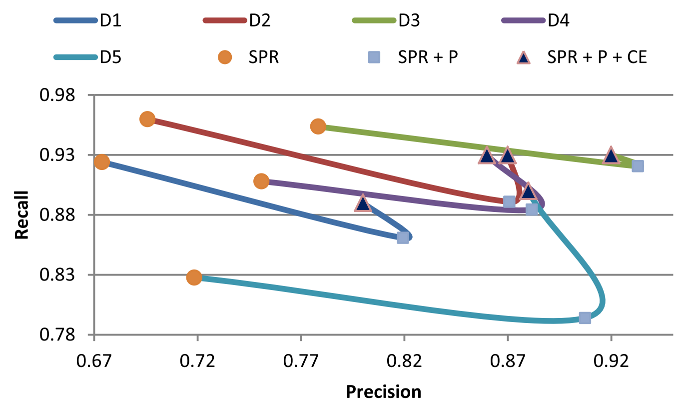

Fig. 4. Performance of the proposed approach (multiple).

extraction. CNN+LP is the supervised neural networks based approach which incorporates linguistic patterns to improve the aspect extraction. These approaches used same dataset with unique number of aspects. The aspects reported in this study are the multiple occurrences of these unique aspects as reported in the original study (Hu & Liu, 2004). As the multiple aspect occurrences represent unique aspects in the same dataset; therefore, the comparison with these techniques is achievable.

4.4. Experimental results

Table 2 shows the performance assessment of the proposed model at every step for both distinct and multiple aspect occurrences evaluation. SPR produces a higher recall for both multiple and distinct occurrences but the precision is very low. The aspect pruning phase (SPR+P) improves the precision by 18% with the loss of 5% in recall for distinct and improves 16% in precision while recall decreases 5% for multiple aspect occurrences. The recall of the aspect extraction is further improved by extracting aspects associated with domain dependent opinions and negations in concept extraction phase (SPR+P+CE). Table 2 shows the effectiveness of the proposed methodology as the overall precision of SPR is improved by 16% with the loss of 4% in recall for distinct occurrences and 15% improvement in precision with no effect on recall for multiple occurrences.

Figs. 3 and 4 elaborate the performance evaluation of the proposed model on each dataset for distinct and multiple occurrences evaluation respectively. In Fig. 3, the precision of distinct evaluation is not consistent and in the cases of D1 and D4, the proposed approach shows a lower precision. This is because of the large number of non-opinionated sentence for these datasets as shown in Table 1. For D2, D3 and D5, the distribution of opinionated and non-opinionated sentences are much even but D1 and D4 have quite a large number of non-opinionated sentences than opinionated sentences. This makes sense because as the number of non-opinionated sentences increases, the precision of the system will decrease but there is not much effect on the recall. This also proves the effectiveness of the proposed approach.

In Figs. 3 and 4, D5 shows the lowest recall for SPR and after the pruning step, the recall jumps up to 19% and 11% for distinct and multiple occurrence evaluations respectively as shown in Table 2. The reasons of lower recall are the domain dependent opinion words. In first fold, SPR uses regular opinion lexicon for the aspect extraction which are the domain independent e.g., good, bad, and excellent. While in D5, large numbers of aspects are associated with domain dependent opinion words e.g., small, big, large, and tiny, which have different polarity in different domains. Therefore, in second fold, the proposed approach searches for aspects associated with domain dependent opinion words and recall of D5 shows better improvement as compared to other products.

4.5. Impact of threshold levels on TF-RBM

Before moving towards the result comparison with similar approaches, it is essential to explain the selection of different threshold level as mentioned in Sections 3.2 and 3.3. Fig. 5 elaborates the performance of TF-RBM for frequency pruning step as explained in Section 3.2.1. TF-RBM has been applied with different threshold levels. These threshold levels represent the minimum support required for noun/noun phrase pruning. Threshold "S@x,y" represents the minimum support "x" for noun phrase while threshold "y" represent minimum support for noun. The threshold levels are tested on all the datasets and selected according to their impact on precision and recall. Threshold "S@2,3" performs more consistently as compared to other threshold levels for D1, D2, D3 and D5 datasets. For D1 dataset, both thresholds "S@3,2" and "S@2,3" perform equally while in D2 and D5 datasets, "S@3,2" shows lower recall and precision as compared with "S@2,3" for these datasets. In the case of D4, "S@2,3" failed to produce better precision and recall as compared to higher threshold levels. D4 dataset contains highest number of review sentences as shown in Table 1. There-

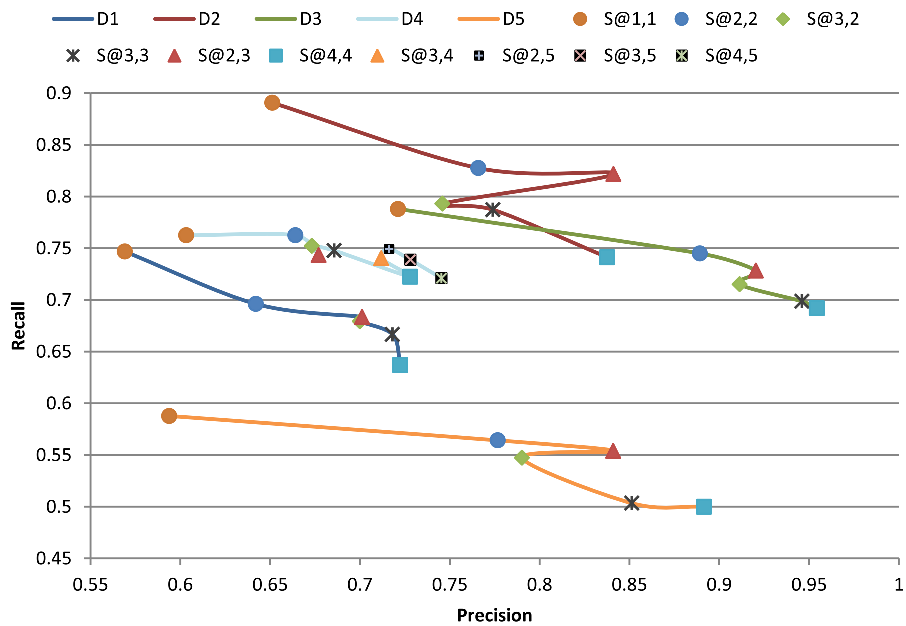

Fig. 5. Impact of threshold levels on frequency-based pruning.

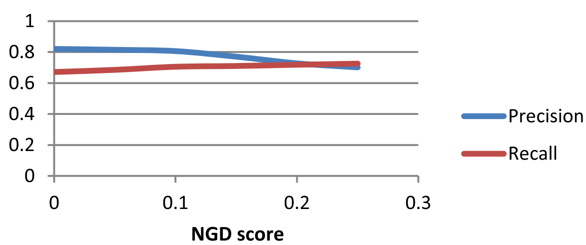

Fig. 6. Impact of threshold levels on similarity-based pruning.

fore, threshold level for datasets with lower number of reviews cannot perform better for the dataset with large number of review sentences. While threshold "S@3,5" shows better result as compared to other threshold levels for D4 dataset. Although, threshold "S@4,5" produced higher precision but shows lower recall than "S@3,5". However, TFRBM was proposed to extract maximum number of aspects therefore, threshold "S@3,5" has been selected.

Fig. 6 shows the impact of different NGD scores on average precision and recall of all the datasets. In the case of frequency pruning, the threshold level differs for different datasets and also different for nouns and noun phrases. Therefore, Fig. 5 shows the impact of different threshold levels on each dataset separately. However, the NGD score remains same for all the datasets and hence, the impact was observed on average results of precision and recall. This is clear from the Fig. 6 that with lower NGD score higher precision is achieved while the higher recall value is achieved with the higher NGD score. The change in precision is minor until 0.1 NGD score but as score changes from 0.1 to 0.15, the precision falls dramatically. Similarly, the major impact on recall is at 0.1 NGD score and as NGD score increased less impact is noticed on the recall. Therefore, the minimum NGD score has been chosen as 0.1 for all the datasets.

Fig. 7 shows the trend of different threshold levels for concept extraction on each dataset. This can be observed that the lower threshold produced higher recall while caused lower precision. The precision of datasets starts increasing with the higher support with a little impact on the recall. For D1, D3 and D5 datasets, the precision and recall start stabilizing at threshold level 6 and there is not much impact of higher threshold on these datasets as shown in Fig. 7(a), (c) and (e) respectively. Although, the precision and recall stabilized at threshold level 8 for D2 dataset however, the major impact of threshold on D2 is up to level 6 and there is not much impact between threshold level 6 and 8 as shown in Fig. 7(b). Therefore, we have selected threshold level 6 for concept extraction. In the case of D4 dataset, similar threshold levels do not perform well as for the other datasets. For D4 dataset, precision and recall stabilized at threshold level 12 as shown in Fig. 7(d) because of the large number of review sentences as compared to other datasets. Therefore, the threshold level varies with the number of review sentences (as explained in Section 3.3) and threshold level for smaller datasets does produced better results on datasets with large number of review sentences.

4.6. Results comparison

Table 3 shows the comparison of precision of the proposed approach, TF-RBM, with the compared approaches DP, CRF, RSG, RSG$^+$, RSLS and RSLS$^+$. Similarly, Table 4 shows the comparison of recall and Table 5 shows the comparison of F_1-score. These all approaches used the same resources, as we have followed; therefore, a fine grained comparison is achievable.

DP required a set of seed opinion words and used a double-propagation algorithm to extract aspects and used same set of extracted aspects to identify new opinion words and the cycle continued until no new aspect or opinion was identified. This makes DP capable to extract domain dependent opinion words and such aspects which were not frequent. On the other hand, DP also suffered from a large number of false positive aspects which affects the precision of the system. Table 4 elaborates that DP has comparable recall with the proposed approach but in case of precision and F_1-score, the approach performed much better, as shown in Tables 3 and 5.

TF-RBM is also compared with the state-of-the-art supervised CRF. The sequential pattern rules (Rana & Cheah, 2016c) were generated by mining sequential patterns from the customer reviews. Therefore, the approach, to some extent, is supervised. As compared to CRF, the approach produce better results for all evaluation metrics as shown in Tables 3–5. Our approach not only performed better for multiple aspects but also produced good results for distinct aspects as compared to CRF.

Table 3
Precision comparison of TF-RBM with DP, CRF, RSG, RSG+, RSLS and RSLS+.

	DP		CRF		RSG		RSG+		RSLS		RSLS+		TF-RBM	
Data	mul	dis	mul	dis	mul	dis	mul	Dis	mul	dis	mul	dis	mul	dis
D1	0.71	0.60	0.79	0.64	0.86	0.83	0.85	0.81	0.87	0.83	0.85	0.80	0.80	0.71
D2	0.74	0.60	0.84	0.74	0.89	0.83	0.87	0.83	0.90	0.86	0.89	0.86	0.87	0.83
D3	0.77	0.58	0.80	0.71	0.84	0.76	0.83	0.76	0.84	0.76	0.83	0.77	0.92	0.90
D4	0.70	0.54	0.87	0.80	0.81	0.72	0.81	0.69	0.82	0.73	0.82	0.70	0.86	0.70
D5	0.63	0.53	0.74	0.66	0.84	0.75	0.86	0.78	0.84	0.77	0.86	0.78	0.88	0.83
Avg	0.71	0.57	0.80	0.71	0.84	0.78	0.84	0.77	0.85	**0.79**	0.85	0.78	**0.87**	**0.79**

Multiple aspect occurrence evaluation (labeled as "mul") and distinct aspect occurrence evaluation (labeled as "dis").

Table 4
Recall comparison of TF-RBM with DP, CRF, RSG, RSG+, RSLS and RSLS+.

	DP		CRF		RSG		RSG+		RSLS		RSLS+		TF-RBM	
Data	mul	dis	mul	dis	mul	dis	mul	Dis	mul	Dis	mul	Dis	Mul	dis
D1	0.91	0.84	0.68	0.53	0.88	0.74	0.91	0.75	0.88	0.75	0.91	0.76	0.89	0.78
D2	0.90	0.79	0.74	0.55	0.88	0.76	0.94	0.83	0.88	0.75	0.94	0.85	0.93	0.87
D3	0.90	0.81	0.51	0.45	0.83	0.67	0.90	0.73	0.84	0.69	0.90	0.79	0.93	0.80
D4	0.89	0.75	0.70	0.50	0.82	0.66	0.90	0.78	0.82	0.67	0.91	0.77	0.93	0.80
D5	0.90	0.76	0.70	0.51	0.79	0.58	0.90	0.68	0.79	0.58	0.90	0.67	0.90	0.73
Avg	0.90	0.79	0.67	0.51	0.84	0.68	0.91	0.76	0.84	0.69	0.91	0.77	**0.92**	**0.80**

Multiple aspect occurrence evaluation (labeled as "mul") and distinct aspect occurrence evaluation (labeled as "dis").

Table 5
F_1-score comparison of TF-RBM with DP, CRF, RSG, RSG+, RSLS and RSLS+.

Data	DP		CRF		RSG		RSG+		RSLS		RSLS+		TF-RBM	
	mul	dis	mul	dis	mul	dis	mul	Dis	mul	dis	mul	dis	mul	dis
D1	0.80	0.70	0.73	0.58	0.87	0.78	0.88	0.78	0.87	0.79	0.88	0.78	0.84	0.75
D2	0.81	0.68	0.79	0.63	0.88	0.79	0.90	0.83	0.89	0.80	0.91	0.85	0.90	0.85
D3	0.83	0.68	0.62	0.55	0.84	0.71	0.86	0.75	0.84	0.73	0.86	0.78	0.92	0.85
D4	0.78	0.63	0.78	0.62	0.81	0.69	0.85	0.73	0.82	0.70	0.86	0.73	0.90	0.74
D5	0.74	0.62	0.72	0.57	0.81	0.65	0.88	0.73	0.81	0.66	0.88	0.72	0.89	0.78
Avg	0.79	0.66	0.73	0.59	0.84	0.72	0.87	0.76	0.85	0.74	0.88	0.77	**0.89**	**0.79**

Multiple aspect occurrence evaluation (labeled as "mul") and distinct aspect occurrence evaluation (labeled as "dis").

Table 6
Precision (P), Recall (R) and F_1-score (F) comparison of TF-RBM with Popecsu, DP, RubE and CNN+LP.

	Popescu			DP			RubE			CNN+LP			TF-RBM		
Data	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
D1	0.89	0.80	0.84	0.87	0.81	0.84	0.87	0.86	0.86	0.93	0.85	0.88	0.80	0.89	0.84
D2	0.87	0.74	0.80	0.90	0.81	0.85	0.90	0.86	0.88	0.82	0.87	0.84	0.87	0.93	0.90
D3	0.89	0.74	0.81	0.90	0.86	0.88	0.90	0.91	0.90	0.90	0.84	0.87	0.92	0.93	0.92
D4	0.86	0.80	0.83	0.81	0.84	0.82	0.87	0.90	0.88	0.92	0.86	0.89	0.86	0.93	0.90
D5	0.90	0.78	0.84	0.92	0.86	0.89	0.90	0.85	0.87	0.93	0.88	0.90	0.88	0.90	0.89
Avg	0.88	0.78	0.82	0.88	0.83	0.86	0.87	0.88	0.87	**0.90**	0.86	0.87	0.87	**0.92**	**0.89**

RSG and RSLS followed the same set of rules as used in DP but they used rule selection algorithms to select the most appropriate rule for the identification of aspects. On the other hand, RSG+ and RSLS+ expanded the set of rules used by DP and added new linguistic rules. RSG+ used the same rule selection algorithm as used by RSG and similarly, RSLS+ used the same as RSLS. This methodology improves the aspect extraction accuracy as compared to DP but none of these approaches can perform consistently. This is evident from the results which showed that RSLS performed better in terms of precision but failed to give better recall and F_1-score; and RSG+ and RSLS+ produced better recall but RSLS+ performed better than RSG+ in terms of F_1-score. Also, these approaches required an extensive rule selection method which makes them more complex. In TF-RBM, the classification of sequential pattern rules automatically selects the most appropriate rules and does not require additional efforts.

As compared with these approaches, TF-RBM performs more consistently for all evaluation metrics for both multiple and distinct evaluations and produced better results as compared to other approaches. Only RSLS produced the same precision for distinct aspects. Fig. 8 shows the performance evaluation of TF-RBM with all the compared approaches and shows improved performance as compared to related approaches.

Table 6 shows the comparison of TF-RBM with two state of the art approaches, i.e., Popescu and DP and two recent approaches, i.e., RubE and CNN+LP. TF-RBM performs better in terms of recall and F_1-score but drops only a little in precision. The precision of TF-RBM decreases 1% as compared to Popescu and DP but improves 15% in recall and 7% in F_1-score as compared to Popescu. As compared to DP, TF-RBM shows 9% improvement in recall and 3% improvement in F_1-score. With RubE, the precision is the same but recall improves by 4% and F_1-score by 2%. As compared to CNN+LP,

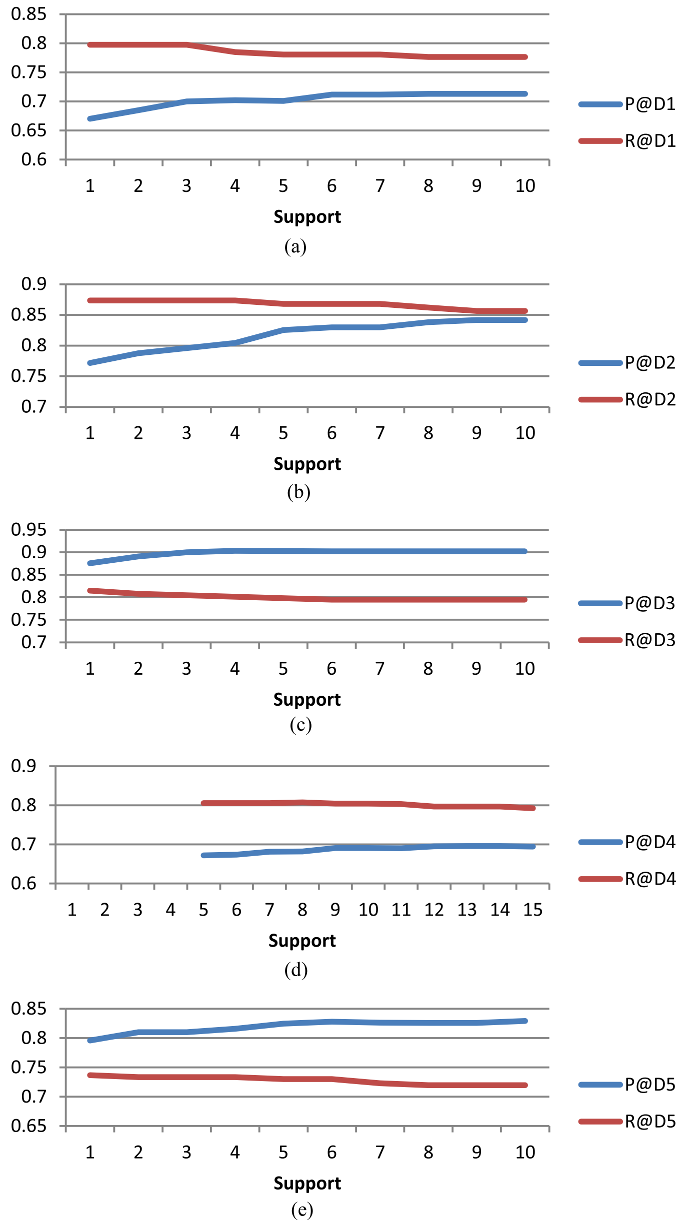

Fig. 7. Impact of threshold levels on concept extraction.

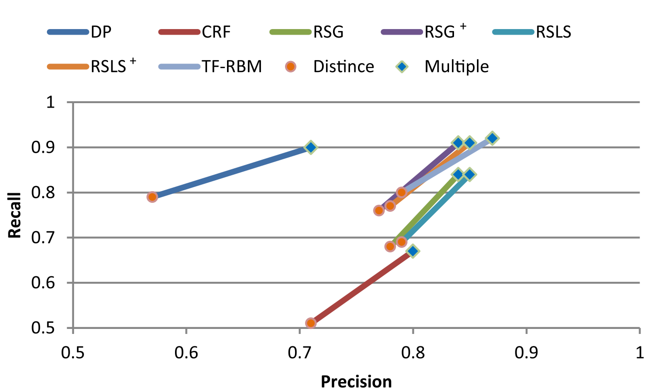

Fig. 8. Performance comparison of TF-RBM with related approaches.

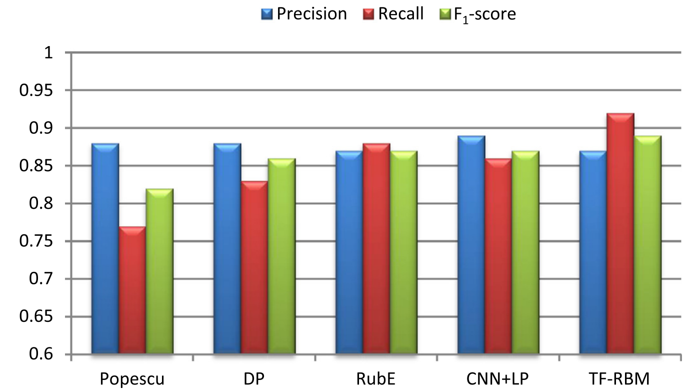

Fig. 9. Comparison of aspect extraction of TF-RBM with related approaches.

TF-RBM improves in terms of recall and F_1-score by 6% and 2% respectively with the decrease in precision by only 3%. Overall, TF-RBM dominates all other approaches both in recall and F_1-score.

Fig. 9 clearly elaborates the effectiveness of TF-RBM in terms of precision and F_1-score as compared with the state-of-the-art approaches. It must also be noticed that our proposed approach not only extracts single noun aspects but also identified aspects which are composed of noun phrases. It proves the importance of the proposed model in aspect extraction from online reviews.

5. Conclusion

In this work, we have introduced a novel two-fold approach for explicit aspect extraction from online customer reviews. The proposed model is capable of identifying single word aspects and multi-word aspects. In the first fold, the proposed approach uses sequential pattern-based rules to extract aspects using domain independent opinion lexicon. To improve aspect extraction accuracy, we applied two pruning techniques i.e., frequency-based and web-based similarity measure. The experimental evaluation has shown the significance of the aspect pruning step in the proposed model. In the second fold, we have used a list of aspects generated after the pruning process to find domain dependent opinion words and negations and extended the set of aspects. This paper presents the step by step performance evaluation of the TF-RBM along with comparison of TF-RBM with the state-of-the-art and most recent approaches. The proposed model produced better results as compared with related approaches.

In our future work, we plan to extend the work by refining and identifying more complex relationships among aspects and opinions. We also plan to integrate sequential pattern-rules with more refined semantic and syntactic relations to increase the performance of the aspect extraction phase and to extend the proposed model on other domains.

Acknowledgments

The authors wish to thank Universiti Sains Malaysia for funding this research through the Research University Grant (1001/PKOMP/8014002). Toqir A. Rana would like to gratefully acknowledge the Ministry of Higher Education (MOHE), Malaysia, for supporting his studies under the Malaysian International Scholarship (MIS) program.

References

Agraval, R., & Srikant, R. (1994). Fast algorithms for mining association rules in large data bases. In *Proceedings of the 20th international conference on very large databases, VLDB: Vol. 1215* (pp. 487–499).

Bagheri, A., Saraee, M., & De Jong, F. (2013). An unsupervised aspect detection model for sentiment analysis of reviews. In *Natural language processing and information systems* (pp. 140–151). Springer.
Bagheri, A., Saraee, M., & De Jong, F. (2014). ADM-LDA: An aspect detection model based on topic modeling using the structure of review sentences. *Journal of Information Science, 40*(5), 621–636.
Bancken, W., Alfarone, D., & Davis, J. (2014). Automatically detecting and rating product aspects from textual customer reviews. In *Proceedings of the 1st international workshop on interactions between data mining and natural language processing at ECML/PKDD* (pp. 1–16).
Brody, S., & Elhadad, N. (2010). An unsupervised aspect-sentiment model for online reviews. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics* (pp. 804–812). Association for Computational Linguistics.
Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems, 31*(2), 102–107.
Cambria, E., Poria, S., Bajpai, R., & Schuller, B. (2016). SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *COLING* (pp. 2666–2677).
Chen, L., Qi, L., & Wang, F. (2012). Comparison of feature-level learning methods for mining online consumer reviews. *Expert Systems with Applications, 39*(10), 9588–9601.
Chen, Z., & Liu, B. (2014a). Mining topics in documents: Standing on the shoulders of big data. In *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1116–1125). ACM.
Chen, Z., & Liu, B. (2014b). Topic modeling using topics from many domains, lifelong learning and big data. In *Proceedings of the 31st international conference on machine learning (ICML-14)* (pp. 703–711).
Chen, Z., Mukherjee, A., & Liu, B. (2014). Aspect extraction with automated prior knowledge learning. In *Proceedings of ACL* (pp. 347–358).
Chen, Z., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R. (2013). Exploiting domain knowledge in aspect extraction. In *EMNLP* (pp. 1655–1667).
Choi, Y., & Cardie, C. (2010). Hierarchical sequential learning for extracting opinions and their attributes. In *Proceedings of the ACL 2010 conference short papers* (pp. 269–274). Association for Computational Linguistics.
Cilibrasi, R. L., & Vitanyi, P. M. (2007). The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering, 19*(3), 370–383.
Cohen, A. *FuzzyWuzzy: Fuzzy string matching in python*. Jan. Retrieved from http://chairnerd.seatgeek.com.
Cruz, F. L., Troyano, J. A., Enríquez, F., Ortega, F. J., & Vallejo, C. G. (2013). Long autonomy or long delay? 'The importance of domain in opinion mining. *Expert Systems with Applications, 40*(8), 3174–3184.
Du, J., Chan, W., & Zhou, X. (2014). A product aspects identification method by using translation-based language model. In *Proceedings of the 22nd international conference on pattern recognition (ICPR)* (pp. 2790–2795). IEEE.
Eirinaki, M., Pisal, S., & Singh, J. (2012). Feature-based opinion mining and ranking. *Journal of Computer and System Sciences, 78*(4), 1175–1184.
Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge: MIT Press.
Gruber, A., Weiss, Y., & Rosen-Zvi, M. (2007). Hidden topic Markov models. In *International conference on artificial intelligence and statistics* (pp. 163–170).
Gunes, O., Furche, T., Shoreditch, L., & Orsi, G. (2016). Structured aspect extraction. In *COLING* (pp. 2321–2332).
Hai, Z., Chang, K., & Cong, G. (2012). One seed to find them all: Mining opinion features via association. In *Proceedings of the 21st ACM international conference on information and knowledge management* (pp. 255–264). ACM.
Hai, Z., Chang, K., Kim, J.-J., & Yang, C. C. (2014). Identifying features in opinion mining via intrinsic and extrinsic domain relevance. *IEEE Transactions on Knowledge and Data Engineering, 26*(3), 623–634.
Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning, 42*(1-2), 177–196.
Htay, S. S., & Lynn, K. T. (2013). Extracting product features and opinion words using pattern knowledge in customer reviews. *The Scientific World Journal, 2013*.
Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 168–177). ACM.
Huang, S., Liu, X., Peng, X., & Niu, Z. (2012). Fine-grained product features extraction and categorization in reviews opinion mining. In *12th international conference on data mining workshops (ICDMW)* (pp. 680–686). IEEE.
Jakob, N., & Gurevych, I. (2010). Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 1035–1045). Association for Computational Linguistics.
Jin, W., Ho, H. H., & Srihari, R. K. (2009). OpinionMiner: A novel machine learning system for web opinion mining and extraction. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1195–1204). ACM.
Jo, Y., & Oh, A. H. (2011). Aspect and sentiment unification model for online review analysis. In *Proceedings of the Fourth ACM international conference on web search and data mining* (pp. 815–824). ACM.
Kang, Y., & Zhou, L. (2016). RubE: Rule-based methods for extracting product features from online consumer reviews. *Information & Management, 54*(2), 166–176.
Kobayashi, N., Inui, K., & Matsumoto, Y. (2007). Extracting aspect-evaluation and aspect-of relations in opinion mining. In *EMNLP-CoNLL* (pp. 1065–1074). Citeseer.
Li, F., Han, C., Huang, M., Zhu, X., Xia, Y.-J., & Zhang, S. (2010). Structure-aware review mining and summarization. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 653–661). Association for Computational Linguistics.
Li, S., Wang, R., & Zhou, G. (2012). Opinion target extraction using a shallow semantic parsing framework. In *Proceedings of the twenty-sixth AAAI conference on artificial intelligence* (pp. 1671–1677). AAAI Press.
Li, S., Zhou, L., & Li, Y. (2015). Improving aspect extraction by augmenting a frequency-based method with web-based similarity measures. *Information Processing & Management, 51*(1), 58–67.
Li, Y., Qin, Z., Xu, W., & Guo, J. (2015). A holistic model of mining product aspects and associated sentiments from online reviews. *Multimedia Tools and Applications, 74*(23), 10177–10194.
Li, Z., Zhang, M., Ma, S., Zhou, B., & Sun, Y. (2009). Automatic extraction for product feature words from comments on the web. In *Information retrieval technology* (pp. 112–123). Springer.
Lin, Z., Jin, X., Xu, X., Wang, W., Cheng, X., & Wang, Y. (2014). A cross-lingual joint aspect/sentiment model for sentiment analysis. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management* (pp. 1089–1098). ACM.
Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures Human Language Technologies, 5*(1), 1–167.
Liu, B., Hsu, W., & Ma, Y. (1998). Integrating classification and association rule mining. In *Proceedings of the 4th international conference on knowledge discovery and data mining (KDD)*.
Liu, K., Xu, L., Liu, Y., & Zhao, J. (2013). Opinion target extraction using partially-supervised word alignment model. In *Proceedings of the twenty-third international joint conference on artificial intelligence* (pp. 2134–2140). AAAI Press.
Liu, K., Xu, L., & Zhao, J. (2012). Opinion target extraction using word-based translation model. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 1346–1356). Association for Computational Linguistics.
Liu, K., Xu, L., & Zhao, J. (2015). Co-extracting opinion targets and opinion words from online reviews based on the word alignment model. *IEEE Transactions on Knowledge and Data Engineering, 27*(3), 636–650.
Liu, Q., Gao, Z., Liu, B., & Zhang, Y. (2015). Automated rule selection for aspect extraction in opinion mining. In *IJCAI: Vol. 15* (pp. 1291–1297).
Liu, Q., Gao, Z., Liu, B., & Zhang, Y. (2016). Automated rule selection for opinion target extraction. *Knowledge-Based Systems, 104*, 74–88.
Liu, Q., Liu, B., Zhang, Y., Kim, D. S., & Gao, Z. (2016). Improving opinion aspect extraction using semantic similarity and aspect associations. In Thirtieth AAAI conference on artificial intelligence (AAAI)*, Arizona, USA* (pp. 2986–2992). Phoenix.
Ma, B., Zhang, D., Yan, Z., & Kim, T. (2013). An LDA and synonym lexicon based approach to product feature extraction from online consumer product reviews. *Journal of Electronic Commerce Research, 14*(4), 304–314.
Marrese-Taylor, E., Velásquez, J. D., & Bravo-Marquez, F. (2014). A novel deterministic approach for aspect-based opinion mining in tourism products reviews. *Expert Systems with Applications, 41*(17), 7764–7775.
Meng, X., & Wang, H. (2009). Mining user reviews: From specification to summarization. In *Proceedings of the ACL-IJCNLP 2009 conference short papers* (pp. 177–180). Association for Computational Linguistics.
Moghaddam, S., & Ester, M. (2010). Opinion digger: An unsupervised opinion miner from unstructured product reviews. In *Proceedings of the 19th ACM international conference on information and knowledge management* (pp. 1825–1828). ACM.
Moghaddam, S., & Ester, M. (2011). ILDA: Interdependent LDA model for learning latent aspects and their ratings from online product reviews. In *Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval* (pp. 665–674). ACM.
Mukherjee, A., & Liu, B. (2012). Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Long papers-volume 1* (pp. 339–348). Association for Computational Linguistics.
Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations Trends Information Retrieval, 2*(1-2), 1–135.
Popescu, A.-M., & Etzioni, O. (2007). Extracting product features and opinions from reviews. In *Natural language processing and text mining* (pp. 9–28). Springer.
Poria, S., Cambria, E., & Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems, 108*, 42–49.
Poria, S., Cambria, E., Ku, L.-W., Gui, C., & Gelbukh, A. (2014). A rule-based approach to aspect extraction from product reviews. In *Proceedings of the second workshop on natural language processing for social media (SocialNLP)* (pp. 28–37).
Poria, S., Chaturvedi, I., Cambria, E., & Bisio, F. (2016). Sentic LDA: Improving on LDA with semantic similarity for aspect-based sentiment analysis. In *2016 international joint conference on neural networks (IJCNN)* (pp. 4465–4473). IEEE.
Qiu, G., Liu, B., Bu, J., & Chen, C. (2009). Expanding domain sentiment lexicon through double propagation. In *IJCAI: Vol. 9* (pp. 1199–1204).
Qiu, G., Liu, B., Bu, J., & Chen, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational Linguistics, 37*(1), 9–27.
Quan, C., & Ren, F. (2014). Unsupervised product feature extraction for feature-oriented opinion determination. *Information Sciences, 272*, 16–28.
Raju, S., Pingali, P., & Varma, V. (2009). An unsupervised approach to product attribute extraction. In *Advances in information retrieval* (pp. 796–800). Springer.

Rana, T. A., & Cheah, Y.-N. (2015). Hybrid rule-based approach for aspect extraction and categorization from customer reviews. In *Proceedings of the 9th international conference on IT in Asia (CITA)* (pp. 1–5). IEEE.

Rana, T. A., & Cheah, Y.-N. (2016a). Aspect extraction in sentiment analysis: Comparative analysis and survey. *Artificial Intelligence Review, 46*(4), 459–483 Springer.

Rana, T. A., & Cheah, Y.-N. (2016b). Exploiting sequential patterns to detect objective aspects from online reviews. In *Proceedings of the 3rd international conference on advanced informatics: Concepts, theory and application (ICAICTA)* (pp. 1–5). IEEE.

Rana, T. A., & Cheah, Y.-N. (2016c). Sequential patterns-based rules for aspect-based sentiment analysis. In *Proceedings of the 3rd international conference on computational science and technology (ICCST)* (pp. 1–5).

Rana, T. A., & Cheah, Y.-N. (2018). Improving aspect extraction using aspect frequency and semantic similarity-based approach for aspect-based sentiment analysis. In *Advances in intelligent systems and computing* (pp. 317–326). Springer.

Rana, T. A., Cheah, Y.-N., & Letchmunan, S. (2016). Topic modeling in sentiment analysis: A systematic review. *Journal ICT Research and Applications, 10*(1), 76–93.

Samha, A. K., Li, Y., & Zhang, J. (2014). Aspect-based opinion extraction from customer reviews. *arXivpreprintarXiv:1404.1982*.

Shams, M., & Baraani-Dastjerdi, A. (2017). Enriched LDA (ELDA): Combination of latent Dirichlet allocation with word co-occurrence analysis for aspect extraction. *Expert Systems with Applications, 80*, 136–146.

Wang, B., & Wang, H. (2008). Bootstrapping both product features and opinion words from Chinese customer reviews with cross-inducing. In *IJCNLP: Vol. 8* (pp. 289–295).

Wang, S., Chen, Z., & Liu, B. (2016). Mining aspect-specific opinion using a holistic lifelong topic model. In *Proceedings of the 25th international conference on world wide web* (pp. 167–176). International World Wide Web Conferences Steering Committee.

Wei, C.-P., Chen, Y.-M., Yang, C.-S., & Yang, C. C. (2010). Understanding what concerns consumers: A semantic approach to product feature extraction from consumer reviews. *Information Systems and E-Business Management, 8*(2), 149–167.

Wu, Y., Zhang, Q., Huang, X., & Wu, L. (2009). Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 3* (pp. 1533–1541). Association for Computational Linguistics.

Xu, L., Liu, K., Lai, S., Chen, Y., & Zhao, J. (2013). Walk and learn: A two-stage approach for opinion words and opinion targets co-extraction. In *Proceedings of the 22nd international conference on world wide web companion* (pp. 95–96). International World Wide Web Conferences Steering Committee.

Xu, X., Tan, S., Liu, Y., Cheng, X., & Lin, Z. (2012). Towards jointly extracting aspects and aspect-specific sentiment knowledge. In *Proceedings of the 21st ACM international conference on information and knowledge management* (pp. 1895–1899). ACM.

Yan, Z., Xing, M., Zhang, D., & Ma, B. (2015). EXPRS: An extended pagerank method for product feature extraction from online consumer reviews. *Information & Management, 52*(7), 850–858.

Yang, B., & Cardie, C. (2013). Joint inference for fine-grained opinion extraction. In *Proceedings of the 51st annual meeting of the association for computational linguistics* (pp. 1640–1649).

Yang, L., Liu, B., Lin, H., & Lin, Y. (2016). Combining local and global information for product feature extraction in opinion documents. *Information Processing Letters, 116*(10), 623–627.

Yu, J., Zha, Z.-J., Wang, M., & Chua, T.-S. (2011). Aspect ranking: Identifying important product aspects from online consumer reviews. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-Volume 1* (pp. 1496–1505). Association for Computational Linguistics.

Yu, J., Zha, Z.-J., Wang, M., Wang, K., & Chua, T.-S. (2011). Domain-assisted product aspect hierarchy generation: Towards hierarchical organization of unstructured consumer reviews. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 140–150). Association for Computational Linguistics.

Zhang, L., Liu, B., Lim, S. H., & O'Brien-Strain, E. (2010). Extracting and ranking product features in opinion documents. In *Proceedings of the 23rd international conference on computational linguistics: Posters* (pp. 1462–1470). Association for Computational Linguistics.

Zhao, W. X., Jiang, J., Yan, H., & Li, X. (2010). Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 56–65). Association for Computational Linguistics.

Zhu, J., Wang, H., Zhu, M., Tsou, B. K., & Ma, M. (2011). Aspect-based opinion polling from customer reviews. *IEEE Transactions on Affective Computing, 2*(1), 37–49.