Journal of Computing and Security

January 2019, Volume 6, Number 1 (pp. 1--11)

http://www.jcomsec.org

Identification of Explicit Features in Persian Comments

Atefeh Mohammadi [a], Mohammad-Reza Pajoohan [a,*], Morteza Montazeri [b], MohammadAli Nematbakhsh [b]

[a] *Department of Computer Engineering, Yazd University, Yazd, Iran.*
[b] *Department of Computer Engineering, University of Isfahan, Isfahan, Iran.*

ARTICLE INFO.

Article history:
Received: 12 August 2018
Revised: 17 September 2019
Accepted: 29 September 2019
Published Online: 18 November 2019

Keywords:
Explicit Feature, Implicit Feature, Association Rules, Co-Occurrence.

ABSTRACT

Recently, the approach towards mining various opinions on weblogs, forums and websites has gained attentions and interests of numerous researchers. In this regard, feature-based opinion mining has been extensively studied in English documents in order to identify implicit and explicit product features and relevant opinions. However, in case of texts written in Persian language, this task faces serious challenges. The objective of this research is to present an unsupervised method for feature-based opinion mining in Persian; an approach which does not require a labeled training dataset. The proposed method in this paper involves extracting explicit product features. Previous studies dealing with extraction of explicit features often focus on lexical roles of words; the approach which cannot be used in distinguishing between an adjective as a part of a noun or a sentiment word. In this study, in addition to lexical roles, syntactic roles are also considered to extract more relevant explicit features. The results demonstrate that the proposed method has got higher recall and precision values compared to prior studies.

1 Introduction

Before making decisions to purchase new products or services, many individuals prefer to know the opinions of previous users [1]. Comparing the reviews and comments, consumers are able to find the product which best matches their needs, desires and requirements. In addition, prosperous retailers and service providers ravenously look after the information indicating what their consumers think about the received products or services. This privilege facilitates a thorough evaluation of their products together with a chance to identify their weaknesses; all of which assist them to improve their products quality and maintain or develop market shares. Due to the unprecedented growth of electronic commerce, many products are now sold over the internet and the number of people who prefer to shop online is increasing day by day [2]. As more and more products are sold online, users' opinions continue to accumulate. This has given rise to a new field of study known as *opinion mining or sentiment analysis*.

Generally, textual information can be placed into two major categories: facts and opinions. The former includes real expressions pertaining to entities, events and characteristics; whereas the latter describes the emotions, opinions and evaluations of individuals regarding those entities, events and characteristics [3].

Sentiment analysis or opinion mining refers to com-

* Corresponding author.
Email addresses: a.mohammadi93@gmail.com (A. Mohammadi), pajoohan@yazd.ac.ir (M. R. Pajoohan), m.montazery@eng.ui.ac.ir (M.Montazeri), nematbakhsh@eng.ui.ac.ir (M.Nematbakhsh)
ISSN: 2322-4460

putational investigation of opinions and attitudes toward entities and their attributes [4]. Opinion mining aims to extract, categorize and summarize opinions and attitudes regarding the features of an entity or service [5]. Taking the obtained opinions into account, parts or subcomponents of a product represent its features. For instance, "camera" and "picture quality" are both features of a "cellphone" entity [6].

Sentiment analysis is investigated at three levels: document, sentence and feature. For most applications, document-level and sentence-level analyses are quite useful; however, since the desirable and undesirable features are not exactly specified, these two levels fail to provide adequate details. Therefore, it is necessary to conduct sentiment analysis at the feature level [2] at which the objective is to discover the sentiments about the features of a product. The sentence "The Samsung phone has good call quality but the battery life is short" evaluates two features of the "Samsung phone" entity; namely "call quality" and "battery life". The sentiment about "call quality" is positive, while the one pertaining to "battery life" is negative.

Feature identification methods are generally placed into two categories; supervised and unsupervised [2, 7]. Supervised methods require a labeled training dataset, creation of which is often a tremendously laborious and costly task. Furthermore, since public data sources are mostly unlabeled, it is necessary to develop adequate models, like the methods based on language pattern mining used for extracting product features, to work with such data [2]. Due to their aforementioned cons, those methods won't be discussed in this work [3].

Sentences may include explicit or implicit features. If a sentence contains a feature explicitly, then that feature is known as an explicit feature. In contrast, if a feature is not explicitly mentioned in a sentence but can be inferred in the context, it is known as implicit feature [8]. For instance, consider "The good thing about this phone is that it's cheap"; the adjective "cheap" refers to the implicit feature "price".

Although syntactic roles and dependency graphs carry vital information, most studies performed on explicit feature extraction using word labelling often do not take them into account. That is why some compound nouns and adjectives cannot be produced in their contexts. For example, the compound noun "مصرف/energy-efficient" is an adjective; but only if word labelling is being used, "کم/efficient" and "مصرف/energy" are extracted as adjective and noun; respectively.

Some studies employed statistical approaches, while they lack adequate information on features extraction. For example, multiword features like "کیفیت صفحه نمایش/display screen quality" may not be extracted by statistical approaches.

In this paper and in line with the previous studies, a new method for identification of explicit features and sentiment words in Persian Documents is proposed. To overcome the aforementioned challenges in this framework, this method takes advantage of considering syntactic roles and sentence dependency graphs as well as statistical approaches.

The rest of the paper is organized as follows: a literature review is provided in the next section. Section 3 shortly explains the problem definition, while the proposed methods to obtain the specified objectives are discussed in Section 4 . Section 5 summarizes the results of executing the proposed algorithms and performs a comparison between the obtained results and those of the previous studies. Lastly, Section 6 concludes and provides some suggestions for future works.

2 Related Works

Hu et al. proposed a method in [9] based on frequent item-sets. In this method, nouns or noun phrases in opinions are recognized using part-of-speech tagging and the frequent ones are extracted using the Apriori algorithm. Frequent nouns whose confidence and support are greater than the specified minimum are identified as features. In this procedure, two pruning approaches are used to remove redundant features; namely compactness pruning and redundancy pruning. For each sentence in the opinion database containing no frequent features and one or two sentiment words, the noun or noun phrase adjacent to the sentiment word is identified and extracted as an infrequent feature. Finally, the sentiment orientation identification function is applied to the extracted features, for the opinions to be classified as either positive or negative.

To extract noun features, an unsupervised technique known as double propagation was created by Qiu et al. [10]. The method works well for medium-size corpora; but precision drops in large ones due to generation of numerous irrelevant features. Contrarily, for small corpora, a large number of important features are lost and thus recall value falls.

In [11], Zhang et al. presented two methods based on "part-whole" and "no" patterns to increase the recall parameter. The former is used to indicate the cases in which one object is a part of one or more objects, while the latter is based on the fact that indi-

viduals often use short comments or opinions to describe features *e.g.* "no parasite" or "no noise". The two patterns are used to extract features missed by double propagation. Besides, the authors proposed a ranking method in order to address the low precision. Since the approach only uses an initial opinion lexicon, it is domain independent and unsupervised; thus avoids the time-consuming labeling process required in supervised learning methods.

Targeting features identification, Qiu et al. took advantage of their syntactic relations with sentiment words [12]. The proposed method uses a bootstrapping approach and is called *double propagations*; as it propagates information between sentiment words and features. The main advantage of this method is that it only requires an initial lexicon of opinions to start the bootstrapping process. Since only an initial opinion lexicon is used, the method is considered semi-supervised. The main principle of the method is based on the fact that the dependencies between sentiment words and features can be extracted using a lexicon of initial opinions and syntactic rules. The resulting sentiment words and features are then used to further extract sentiment words and features. Propagation stops when no more sentiment words or features are identified. In the next step, three rules are followed to assign polarities to the sentiment words: (1) heterogeneous rule, (2) homogeneous rule and (3) intra-review rule. Once the polarities of the sentiment words are assigned, unnecessary features are pruned using a novel technique proposed by the authors.

The co-occurrence association rule mining method for implicit feature extraction was presented by Hai et al. [13]. In this approach, nouns and noun phrases are considered explicit features while adjectives and verbs constitute sentiment words. In the first phase, two sets of words (*i.e.* sentiment words and features) are extracted from the explicit sentences in the corpus. Then a co-occurrence matrix is constructed wherein each element denotes the number of co-occurrences of explicit words and features in the sentence. After clustering the explicit features and looking for sentiment words that lack explicit features, the second phase of the method searches a matched list of rules, so that the ones in which the feature cluster with the highest frequency weight is included are discovered. Accordingly, the representative word of the cluster is identified as the implicit feature.

Hai et al. proposed a generalized approach in [7] for extraction of opinion words and features through statistical association analysis. A small seed of features is utilized in the beginning and then iteratively been extended by mining feature-opinion, feature-feature and opinion-opinion dependency relations. Two association model types known as Likelihood Ratio Tests (LRT) and Latent Semantic Analysis (LSA), used for calculation of the association between two words, form the basis on which two bootstrapping approaches namely Likelihood Ratio Tests Based Bootstrapping (LRTBOOT) and Latent Semantic Analysis Based Bootstrapping (LSABOOT) stand. Both approaches require an initial set of features to bootstrap the feature and opinion extraction process.

In [14], a method known as hybrid association rule mining was proposed by Wang et al. This approach considers numerous hybrid methods to mine a large number of association rules. Firstly, explicit sentences are collected for each feature through feature clustering. Candidate feature indicators are then extracted from the explicit sentences using word segmentation and tagging. The weight of each indicator is computed using five algorithms: frequency, Point-wise Mutual Information (PMI), frequency*PMI, t-test, and Chi-square. The feature words whose indicator weight is greater than the given threshold are added to the rule set. In fact, the method uses a combination of five rules to identify implicit features. Furthermore, a pruning algorithm is presented to remove conflicting indicators such as "good" which may refer to a variety of features. The advantage is that mere utilization of basic rules prevents indicators of lower weight or non-indicators from being extracted. To expand the initial rules, the authors use substring hypothesis, dependency grammar and semi-supervised learning, using a constrained topic model. Finally, the association rules, obtained as a combination of expanded and basic rules, are used for implicit feature extraction.

Bagheri et al. presented an unsupervised model of feature and sentiment identification, capable of extracting explicit and implicit features in English [2]. In this model, nouns are considered as features while adjectives, adverbs and verbs express sentiments. Multi-word features are identified using Frequency Modified Left Right (FMLR). Two pruning rules are then used to complete the pre-processing procedure. In the next stage, a bootstrapping algorithm which receives seed features as input and extracts further features accordingly is presented. The features are ranked according to A-score. Once the list of features is finalized, redundant ones are removed using two types of pruning strategies; namely subset-support and superset-support pruning scenarios. A graph is drawn for the set based on the set of sentiment words and implicit features. An edge connects a feature and an adjective provides the situation of their co-occurrence. The initial weight of the edge is determined based on the frequency of the co-occurrence. Ultimately, the graph is used to identify implicit features.

In order to extract features from product reviews, a rule-based approach was presented by Poria et al. in [15]. The authors employed common-sense knowledge and sentence dependency trees for extracting both implicit and explicit features and reported higher accuracies for the two datasets. The proposed method is completely unsupervised; its accuracy is dependent on the accuracy of the dependency parser and the opinion lexicon.

The model proposed by Schouten et al. in [16] uses a co-occurrence matrix for identifying implicit features. In this method, the training data are examined to construct a list of all implicit features (F), a list of all words (O) and a matrix containing the features and words that co-occur in one sentence (C). Once the lists (*i.e.* F and O) and the matrix (*i.e.* C) are constructed, the test data are analyzed yielding a score denoted by f_i , with i representing the number of every implicit feature. The value is basically obtained as the sum of co-occurrences for each word divided by the number of times it occurs in the sentence.

In [8], A semi-supervised method for identifying implicit features in users opinions was developed by Xu et al. for Chinese language. The method receives the product along with pertinent opinions as inputs and extracts the explicit sentences and the corresponding features using labeling. In the next step, the features are clustered; words and expressions with synonymous domains fall into same clusters. In this method, sentences having no features are considered implicit. Then constraint sets, such as must-link and cannot-link constraints as well as syntactic prior knowledge are extracted from explicit sentences and combined with a constrained topic model. Finally, a number of SVM classifiers are generated and used for identification of implicit features.

A novel unsupervised feature-based opinion mining method for product reviews in Persian was devised by Baba Ali et al. in [17]. This approach involves three stages: (1) explicit feature identification, (2) implicit feature identification and (3) determining semantic orientation. The output of the method is put into five levels (*i.e.* very good, good, medium, poor, very poor) all of which cover polarity as well as strength (severity) of opinions orientation. The proposed method extracts implicit features using co-occurrence association rule mining [13] with some modifications such as using a set of synonym words in pace of clustering or employing the maximum confidence rule instead of the feature cluster with the greatest frequency weight. In this approach, the co-occurrence matrix is constructed by joining sentiment words and explicit features. Then for each sentiment word occurring in a sentence including an explicit feature, a considerable set of association rules is constructed in the form of sentiment word) → explicit feature. Weak rules are adequately pruned based on minimum confidence and support values. In the second stage, sentences without explicit features containing sentiment words are examined and a list of strong rules associated with them is developed. As a result, association rules with high confidence values are extracted. Finally, representative word(s) are identified as implicit features.

3 Problem Definition

Extracting explicit features in Persian texts often meets numerous inherent challenges; listed as follows:

Challenge 1- Sometimes an adjective occurs near a noun and the combination forms a second adjective; in such cases, the noun cannot be regarded as a feature for the sentence. In a similar vein, when coming next to each other, certain nouns and adjectives form compound nouns. For instance, the sentence "**عالی است سیستم خنک کننده** / The cooling system is excellent", the adjective "**خنک کننده** / cooling" is not a sentiment for a feature since the combination of "**سیستم** / system" and "**خنک کننده** / cooling" results in the new noun "**سیستم خنک کننده** / cooling system".

Challenge 2- In some sentences, the sentiment word comes between two nouns; still, despite being separated, the nouns create another noun (feature) and the adjective becomes the sentiment word for the compound noun. For example, consider the sentences "**اولین چیزی که نظرمو بشدت جلب کرد عمر فوق‌العاده باتریش است** / the first thing that caught my eye was the battery's extraordinary life." and "**خوبش، کیفیت بالای ساختش است یکی از ویژگی‌های** . / a good thing about it is the design's high quality". The nouns "**باتری** / battery" and "**عمر** / life" in the former and the words "**ساخت** / design" and "**کیفیت** / quality" in the latter are combined to create the features "**عمر باتری** / battery life" and "**کیفیت ساخت** / design quality"; respectively. In these sentences, the adjectives "**فوق‌العاده** / extraordinary" and "**بالا** / high" are sentiment words for "**عمر باتری** /battery life" and "**کیفیت ساخت** / design quality"; respectively.

As it has been shown, ignoring the sentences' structures and syntactic roles causes some important features and sentiment words being missed. We present an explicit feature extraction approach which relies on statistical approaches, syntactic roles and sentence structures to overcome existing challenges. Addressing these issues results more accurate explicit feature extraction.

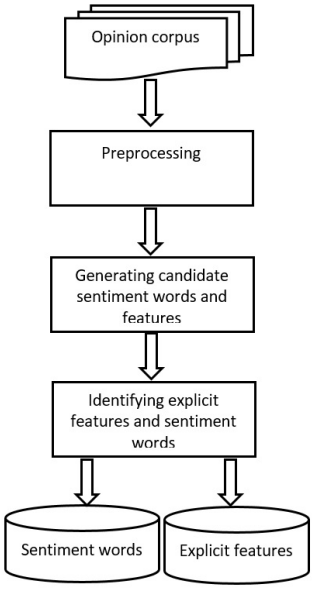

Figure 1. The Overall Process for Extracting Features and Sentiment Words.

4 The Proposed Method

The proposed method in this research involves with identifying explicit features and sentiments for each product. The main focus of this paper is the feature extraction method in which the reviews will be preprocessed so that the syntactic role of each word can be determined and a dependency graph is generated. Then, statistical approaches are used to elicit the candidate sentiment words and features. Finally, the features and sentiment words are extracted regarding the sentence structure. Figure 1 illustrates the overall process of extracting features and sentiment words. As shown, all opinion texts undergo preprocessing prior to extracting sentiment words and product features.

5 The Preprocessing Stage

As the initial step of the preprocessing stage, the set of opinions about each product is given as input to the stemmer tool, developed at Ferdowsi University of Mashhad for normalization [18]. The sentences are then parsed into one or several smaller sentences based on punctuation marks such as periods, exclamation marks, question marks and colons; all as separators of the sub-sentences. In the next step, the Hazm tool presented in [19] is used to stem and label the words and to create a dependency graph between them. It is notable that both stemmed and original words are labeled and stored. Since the tool does not suite long sentences, they will be broken into smaller units such that each unit contains only one verb. For instance, consider the sentence "**ندارد**

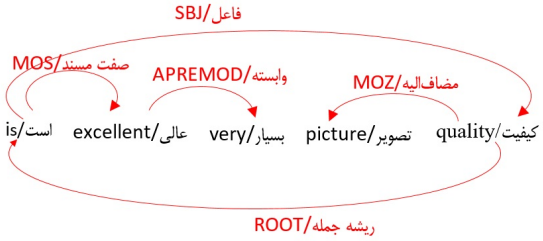

Figure 2. Sentence Dependency Graph.

کیفیت تصویر بسیار عالی است ولی دوربین مناسبی / picture quality is very excellent but the camera is not good." To determine the syntactic role of each word and generate a dependency graph, the opinion is broken into sentences: "**کیفیت تصویر بسیار عالی است** /picture quality is very excellent" and " **ولی دوربین مناسبی ندارد**/but the camera is not good". The syntactic roles in the first sentence are "**کیفیت**/quality/ **تصویر** /picture/MOZ **بسیار**/very/APREMOD **عالی**/excellent/MOS **است**/is/ROOT"; the corresponding dependency graph of which is shown in Figure 2.

As shown in Figure 2, the word "**کیفیت**/quality" is dependent on "**است**/is" while "**تصویر**/picture" depends on "**کیفیت**/quality".

6 Generating Candidate Sentiment Words and Features

Subsequent to the sentences' preprocessing stage, all adjectives and nouns are examined for the candidate sentiment words and explicit features to be extracted. Nouns and sentiments represent features and adjectives; respectively. The conditions for extracting each one is separately discussed below.

In order to extract explicit features in each sentence, word labels together with the corresponding syntactic roles are taken into account. For example, in the sentence "**کیفیت**/ its price is fair" the syntactic and lexical roles are as follows: "**قیمت** /Price/SBJ/Ne **آن** /Its/MOZ/Pro **عادلانه** /Fair/MOS/AJ **است** /Is/V/ROOT" (lexical and syntactic roles are separated with "/"). So for each sentence, all words with the following conditions are added to the list of candidate features:

(1) The words with the syntactic role of subject or genitive are added to the list of candidate features.)Note that the features are added to the list of candidate features and that of sentence identifiers).
(2) In Persian, some words yield an adjective after being stemmed. The grammatical and syntactic roles of these words are subject and adjective; respectively. For example, the

word "زیبایی/beauty" is stemmed to obtain "زیبا/beautiful". For each sentence, these words must be added to the list of candidate features rather than the candidate adjectives.

(3) Only if the noun coming before the conjunction "and" exists in the list of candidate features, the noun preceding the conjunction will be added to the list of candidate features. For example in the sentence "it has flash and focus", the word "focus" is regarded as a candidate feature of the sentence; provided that the word "flash" already exists in the candidate features list.

Once candidate features are extracted, it is time to extract the candidate adjectives. If a word in a sentence is an adjective in both stemmed and original (prior to stemming) forms, its syntactic role is also taken into account along with extracting the product features. The word is added to the list of candidate adjectives provided that one of the following conditions is met:

(1) For each sentence, words have one of these grammatical roles: predicate, post-noun adjective, object or adverb. (Note that the words are added to list of candidate adjectives together with that of sentence identifier).

(2) If a sentence is without a verb and it ends with an adjective, Hazm incorrectly identifies the adjective as the sentence root; thus it is considered as the candidate adjective of the sentence and hence is added to the relevant list.

(3) If the conjunction "and" in a compound is preceded by an adjective belonging to the list of candidate adjectives, the adjective appearing after the conjunction is only added to the list of candidate adjectives. As an example, consider the sentence "the touch is excellent and smooth"; the word "smooth" is regarded as a candidate adjective of the sentence provided that "excellent" already exists in the list of candidate features.

If a word is an adjective in the stemmed sentence while previously was a noun in the original sentence, the correct part of the speech must be determined. The word is an adjective if one of the following conditions are satisfied, and a feature otherwise:

- The candidate word appears after a noun and before a verb; as adjectives do in Persian language.
- The candidate word precedes an adverb while no adjective appears after the candidate word.
- The noun-forming suffix does not appear after the adjective (*e.g.* "y" or "ness" in English).

In some cases, Hazm may be unable to correctly identify the syntactic roles of some words. Thus, using the lists of candidate features and adjectives, all the sentences are examined once again, looking for words which belong to the lists of candidate features or adjectives, but have been missed in extraction as a result of Hazm errors. The words are then extracted as candidate features or adjectives and added to the related list. Let's consider the following example for clarification:

- "باتری آن عالی است/its battery is excellent"
- "عمر باتری چندان هم خوب نبود/the battery life was not that good"
- blah, blah.

Suppose "باتری / battery" is chosen from the first sentence and added to the list of candidate features. Further, suppose that "عمر/life" is extracted from the second sentence as a feature; whereas "battery/باتری" is not. Hence, using the list of candidate features, the word "battery/باتری" is extracted and added to the list. In the following step, explicit features and sentiment words are extracted.

7 Identifying Explicit Features and Sentiment Words

Subsequent to extracting candidate features and adjectives, one must address the challenges in the process of extracting explicit features and sentiment words. Figure 3 depicts the main steps of this process. The current stage of the proposed method is distinct in that it achieves the following tasks:

(1) Distinguishing between adjectives as parts of nouns or sentiment words
(2) Identifying multi-word features
(3) Identifying complex forms of genitive cases

8 Distinguishing Between Adjectives as Parts of Nouns or Sentiment Words

In some Persian sentences, an adjective that comes before a noun may be a part of the noun itself rather than a sentiment of feature. For instance, the adjective "تصویری/visual" does not represent a sentiment or feature in the sentence "تصویری این گوشی حرف ندارد نمایش/the phone's visual display is awesome". In contrast, the adjective "inexpensive/ " in the sentence "قیمت و کارآمد می‌خواهید بخرید، توی خرید این گوشی شک نکنید / در کل اگر گوشی ارزان overall, if you want an inexpensive usable cellphone, do not doubt it" is a sentiment for the feature "قیمت/price" and cannot be combined with the feature. In other cases, the adjective is in fact a part of the noun, while representing a sentiment for a feature. As an example, in the sentence " مصرف است باتری این گوشی کم /the battery is low-energy", the adjec-

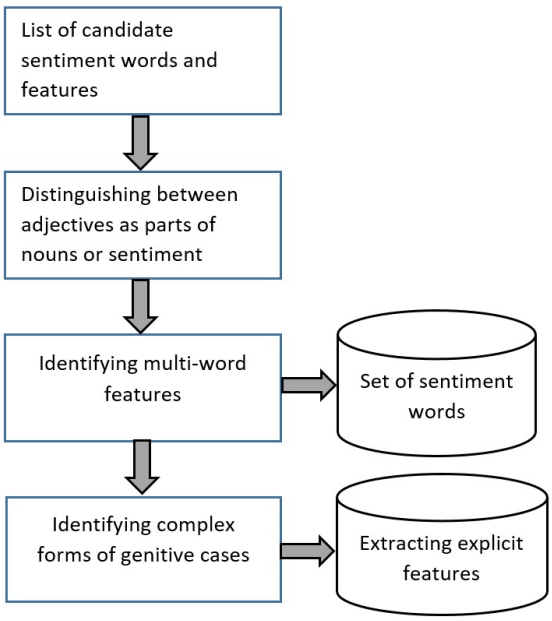

Figure 3. Identifying Explicit Features and Sentiment Words.

tive "کم/low" is attached to the noun "مصرف/energy" to form the adjective "کم مصرف /low-energy" which represents a sentiment for the feature "باتری/battery". Here, the primary objective is to separate adjectives that are combined with the nouns, so the candidate features or sentiments can be formed. For this to happen, the combinations of "adjective + noun" in every sentence are examined based on the list of candidate features and adjectives. The distinguished combination will be maintained if the two following conditions are met:

- The adjective + noun combination frequently occurs in the opinion corpus
- The Likelihood Ratio Test (LRT) [7], signifying the dependence between two words in the corpus, exceeds ∂ which denotes the dependency identifier threshold. Note that larger LRT values indicate greater dependencies (The LRT is the ratio of the probability of the two terms to that of all comments; indicating the degree of dependence between those two words. Employing LRT criterion in the calculations has led to better statistical results in texts' analyses; because of two main reasons: 1- yielding good results for short texts. 2-consideration of the dependency between rare and frequent words in the calculations).
- blah, blah.

If the aforementioned combination satisfies the following condition, a new feature will be made, the old value should be replaced by the new one and the adjective is removed from the list of candidate adjectives.

- Frequently in the corpus, an adjective comes after the combination and before the verb. It is worth mentioning that in Persian language, adjectives tend to appear after nouns.

For example, if the feature "display" and the adjective "visual" met the abovementioned condition, a new feature "visual display" will be produced and replaces the old feature and the adjective for that sentence. Otherwise, if the combination satisfies the following conditions, an adjective is formed and the corresponding feature (*e.g.* "energy") is eliminated from the list of candidate features. Again, the new adjective (*e.g.* "low-energy") replaces the old one (*e.g.* "low").

- In the corpus, no adjective or adverb appears after the combination
- An adverb of degree precedes the combination; as they often come before the adjectives.

Sometimes, the "noun + adjective" combination may produce a noun. For such combinations to be properly detected, a sentence-by-sentence examination based on the candidate features and adjectives seems necessary. If the following conditions are satisfied, the old adjective is eliminated from the list of candidate adjectives and the new adjective substitutes the old noun.

- The "noun + adjective" combination is frequent
- The LRT value exceeds ∂
- In the corpus, an adjective or adverb of degree comes after the combination and before the verb

As a drawback of this scenario, multi-word features such as "عمر باتری /battery life" and "صفحه نمایش /display screen" cannot be extracted in this step; however, this challenge is addressed in the next step.

9 Identifying Multi-Word Features

Some opinions in the corpus may contain features composed of several words; for example, "صفحه نمایش کیفیت /display screen quality" or " عمر باتری/battery life". Multi-word features can be distinguished as candidates if one of the following conditions are satisfied:

(1) Several words depend on each other in the dependency graph.
(2) In cases in which the dependency graph is insufficient to identify multi-word features, LRT is used: the words are combined if LRT exceeds the threshold ∂.

Still, the algorithm presented in this step may be unsuccessful in identification of all multi-word features. To overcome this problem, all sentences are examined once more using the multi-word detection algorithm to identify any multi-word features which may have been missed.

Finally, single-word features occurring fewer than three times in the dataset are removed from the list; which in turn causes higher precision and greater recall performance [20]. This is due to the fact that the number of repetitions of a word is important in its being distinguished as a feature.

10 Identifying Complex Forms of Genitive Cases

The final step in extracting explicit features and sentiment words involves a re-examination procedure of the sentences. The opinion corpus may contain sentences wherein a sentiment word is located between two nouns which together, regardless of their distance, can form a compound noun (feature). The adjective between those two nouns is then regarded as the adjective for the compound noun (feature). For instance, let's consider the sentence "**کرد عمر فوق‌العاده باتریش است اولین چیزی که نظرمو بشدت جلب** / the first thing that caught my eye was the battery's extraordinary life". The nouns "**باتری** / battery" and "**عمر** / life" in the sentence are combined to create the feature "**عمر باتری** / battery life". In this sentence, the adjective "**فوق‌العاده** / extraordinary" is a sentiment word for " **عمر باتری** /battery life". Sentences containing such features are only identified if the following conditions are met:

- The compound noun already exists in the list of candidate features extracted in the previous step. That is, the compound noun already exists as a candidate feature for another sentence.
- An adjective or a combination of adverb + adjective appears between the two features.

11 The Evaluation Procedure

This section provides an evaluation of the method proposed in this study. The details of datasets, evaluation criteria, experimental results and the performed comparisons are presented in the following subsections.

12 The Data sets

In this paper, a set of user opinions employed in [17] (obtained from DIGIKALA website [1] ; an Iranian commercial website selling various types of goods online) is used for evaluation purposes. As shown in Table 1, the dataset contains user opinions regarding two distinct categories of products; laptops and cellphones [21]. It is worth noting that the dataset is in Persian language.

[1] http://www.digikala.com

First, all sentences associated with the understudy products are normalized by Hazm and the conversational verbs are converted into formal ones. Next, Hazm is applied once more to extract the syntactic roles and dependency graphs of all sentences. A dataset is formed to be used as the input for the proposed method.

The proposed method acts on explicit features. To evaluate the performance of the method, two expert users were asked to read all comments and extract the explicit features of each sentence (if any). The obtained features were then compared with the explicit features extracted by the method. Therefore, there are four possible states:

- TP parameter: The number of features that are properly extracted by the model.
- FP parameter: The number of features that are falsely extracted by the model.
- FN parameter: The number of explicit features that their explicitness was not identified by the model.
- TN parameter: The number of inexplicit features that are not properly extracted by the model.

13 Evaluation Criteria

To evaluate the performance of the proposed method, precision, recall and F-measure criteria have been investigated. To this end, the extracted explicit features were compared to those identified manually and the necessary calculations were carried out. In this study, ∂ threshold which has been used in the extraction process of explicit features is experimentally chosen 2.5.

14 Performance Evaluation and Analysis

Table 2 compares the results of the explicit feature detection method presented in this research with those derived by Baba Ali in [17]. As shown, the proposed method enjoys a sizable lead in precision and recall values. The method proposed by Baba Ali et al. [17] is flawed in that it generates a large number of false candidates; the problems which stems from two main reasons: (1) lexically labeled nouns and adjectives are selected as explicit features and sentiment words, respectively; and (2) word orders are ignored in the process of extracting multi-word features. Furthermore, method 2 fails to propose a framework in which "adjective + noun" and "noun + adjective" pairs are combined; thus, features such as " **نمایش تصویری**/visual display", " **سیستم خنک کننده**/cooling system" or adjectives like "**کم مصرف**/low-energy" will be missed. This

Table 1. Statistical Breakdown of the Opinions.

Category	Number of Reviews	Number of Sentence
Cellphones	394	1874
Laptops	168	665
Total	562	2539

Table 2. Explicit Feature Detection Criteria: The Proposed Method (Method 1) vs. Baba Ali Method [17] (Method 2).

Category	Precision		Recall		F-measure	
	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
Cellphones	0.88	0.79	0.92	0.74	0.90	0.76
Laptops	0.85	0.78	0.91	0.66	0.88	0.72
Total	0.87	0.79	0.92	0.70	0.89	0.74

drawback can be compensated using the method presented in this study.

The precision, recall and F-measure values calculated for the proposed explicit feature extraction method are compared to those derived from the methods developed by Bagheri [2] and Hu [9] in Tables (3) and (4), respectively. The proposed method by Bagheri et al. in [2], mainly developed for English documents, cannot identify multi-word features such as "سیستم خنک کننده /cooling system". Moreover, it does not address complex genitive cases. Thus, for instance, in the sentence "کرد عمر فوق‌العاده باتریش است / اولین چیزی که نظرمو بشدت جلب the first thing that caught my eyes was the battery's extraordinary life", the adjective "فوق‌العاده / extraordinary" is incorrectly extracted as a sentiment word for "عمر / life". Hu et al. [9] extracted nouns / noun phrases and adjectives as attributes and sentiment words; respectively. They believed that other components of a sentence are unlikely to be product features. Therefore, their method failed to extract features and sentiment words such as "سیستم خنک کننده/cooling system" and "کم مصرف /low-energy".

On the negative side, since the proposed method in the present study considers more conditions and rules (*e.g.* extracting syntactic rules and dependency graphs), it is not as fast as other algorithms mentioned in the literature. Besides, although it shows an improvement compared to its previous counterparts, cases of wrong identified explicit features can still be found. The reasons for this can best be summarized as follows: First, the number of syntactic rules in Persian language is not high enough. Second, Hazm software is not accurate in recognition of the dependency between the words of a sentence. Finally, unlike English-language sites such as Amazon, the number of comments on the counterpart Persian sites is quite low which in turn reduces the accuracy of explicit features detection process.

15 Conclusions and Future Work

This paper proposed a novel method for extracting explicit features from user opinions written in Persian language; consisting of three main parts: 1) distinguishing between adjectives as parts of nouns or sentiment words, 2) identifying multi-word features and 3) identifying complex forms of genitive cases. The first part employed dependency graphs, syntactic roles in Persian and LRT value to identify explicit features and sentiment words. In the second part, dependency graphs and LRT are used to identify multi-word features. Finally the last part uses Persian syntactic roles for identification of complex forms of genitive cases. All three parts lead the accuracy of the method to be higher than that of previous ones presented in the literature.

Avenues for future researches include improving labelling and stemming which directly impact opinion mining results. Moreover, larger datasets consisting of Persian language comments can be taken into account to improve the accuracy of dependency graphs in feature detection procedures. Also, exploring how opinion texts vary over time is another area of research. Finally, since opinions tend to change as time goes by, identification of potential trends is also of a great interest for most organizations and companies; and thus is considered as a great potential for further researches and investigations.

Table 3. Explicit Feature Detection Criteria: The Proposed Method (Method 1) vs. Bagheri [2] (Method 2).

Category	Precision		Recall		F-measure	
	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
Cellphones	0.88	0.75	0.92	0.83	0.90	0.85
Laptops	0.85	0.72	0.91	0.81	0.88	0.80
Total	0.87	0.74	0.92	0.82	0.89	0.83

Table 4. Explicit Feature Detection Criteria: The Proposed Method (Method 1) vs. Hu [9] (Method 2).

Category	Precision		Recall		F-measure	
	Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
Cellphones	0.88	0.70	0.92	0.72	0.90	0.73
Laptops	0.85	0.69	0.91	0.70	0.88	0.70
Total	0.87	0.70	0.92	0.71	0.89	0.72

References

[1] E. Lloret, A. Balahur, J. M. Gómez, A. Montoyo, and M. Palomar. Towards a unified framework for opinion retrieval, mining and summarization. *Journal of Intelligent Information Systems*, 39 (3):711–747, 2012. ISSN 1573-7675. URL 10.1007/s10844-012-0209-4.

[2] A. Bagheria, M. Saraeeb, and F. de Jong. Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. *Knowledge-Based System*, 52:201–213, 2013. URL 10.1016/j.knosys.2013.08.011.

[3] Bing Liu. Sentiment analysis and subjectivity. In *Handbook of natural language processing*, pages 627–666, 2010. URL 10.1201/9781420085938-c26.

[4] Bing Liu. Sentiment analysis and opinion mining. In *Synthesis Lectures on Human Language Technologies*, pages 1–167. Morgan & Claypool Publishers, 2012. URL 10.2200/S00416ED1V01Y201204HLT016.

[5] G. Vinodhini and RM. Chandrasekaran. Sentiment analysis and opinion mining: a survey. *International Journal*, 2(6):282–292, 2012.

[6] S. Moghaddam and M. Ester. Opinion digger: an unsupervised opinion miner from unstructured product reviews. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1825–1828. ACM, 2010. ISBN 978-1-4503-0099-5. URL 10.1145/1871437.1871739.

[7] Zhen Hai, Kuiyu Chang, and G. Cong. One seed to find them all: mining opinion features via association. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 255–264. ACM, 2012. ISBN 978-1-4503-1156-4. URL 10.1145/2396761.2396797.

[8] H. Xu, F. Zhang, and W. Wang. Implicit feature identification in Chinese reviews using explicit topic mining model. *Knowledge-Based Systems*, 76:166–175, 2015. URL 10.1016/j.knosys.2014.12.012.

[9] M. Hu and B. Liu. Mining opinion features in customer reviews. In *Proceedings of the 19th national conference on Artifical intelligence*, pages 755–760, 2004. ISBN 0-262-51183-5.

[10] G. Qiu, B. Liu, J. Bu, and C. Chen. Expanding Domain Sentiment Lexicon through Double Propagation. In *Proceedings of the 21st international jont conference on Artifical intelligence*, pages 1199–1204. URL 2009.

[11] L. Zhang, B. Liu, S. Hwan Lim, and E. O'Brien-Strain. Extracting and ranking product features in opinion documents. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1462–1470, 2010.

[12] G. Qiu, B. Liu, J. Bu, and C. Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011. ISSN 1573-7675. URL 10.1162/coli_a_00034.

[13] Z. Hai, K. Chang, and J. Kim. Implicit feature identification via co-occurrence association rule mining. In *International Conference on Intelligent Text Processing and Computational*

Linguistics, pages 393–404. Springer, Heidelberg, 2011. ISBN 978-3-642-19399-6. URL 10.1145/2396761.2396797.

[14] W. Wang, H. Xu, and W. Wan. Implicit feature identification via hybrid association rule mining. *Expert Systems with Applications*, 40(9): 3518–3531, 2013. URL 10.1016/j.eswa.2012.12.060.

[15] S. Poria, E. Cambria, L. Ku, C. Gui, and A. Gelbukh. A rule-based approach to aspect extraction from product reviews. In *Proceedings of the second workshop on natural language processing for social media (SocialNLP)*, pages 28–37, 2014. URL 10.3115/v1/W14-5905.

[16] K. Schouten and F. Frasincar. Implicit Feature Extraction for Sentiment Analysis in Consumer Reviews. In *International Conference on Applications of Natural Language to Data Bases/Information Systems*, pages 228–231. Springer, 2014. ISBN 978-3-319-07982-0. URL 10.1007/978-3-319-07983-7_31.

[17] K. Schouten and F. Frasincar. Extracting Product Features in Persian. In *Proceedings of the 3rd Computational Linguistics Conference, Sharif University*, 2014.

[18] Ferdowsi University of Mashhad. Natural Language Processing Tools. http://wtlab.um.ac.ir, Web Technology Lab, 2012.

[19] Khalash M Imani M. Persian Language Processing Tool. http://www.sobhe.ir/hazm, 2013.

[20] C. Wei, Y. Chen, C. Yang, and C. C. Yang. Understanding what concerns consumers: a semantic approach to product feature extraction from consumer reviews. *Information Systems and e-Business Management*, 8(2):149–167, 2010. ISSN 1617-9846. URL 10.1007/s10257-009-0113-9.

[21] uplooder. Digikala Dataset (2018, December 13). https://www.uplooder.net/files/a3cbdeb80f2b59d51cd858ab0f4d4558/DataSet.rar.html, Retrieved February 2, 2019.

Atefeh Mohammadi received her M.Sc degree in computer engineering from University of Isfahan in 2016. Currently, she is a PhD student at the Department of Computer Engineering, Yazd University, Iran.

Mohammad-Reza Pajoohan received his PhD degree in computer science from Universiti Sains Malaysia in 2010. Currently, he is an Assistant Professor at the Department of Computer Engineering, Yazd University, Iran.

Morteza Montazeri received his M.Sc degree in computer engineering from Tehran University in 2013.

MohammadAli Nematbakhsh received his PhD degree in computer engineering from University of Arizona in 1987. Currently, he is an associate professor at the Department of Computer Engineering, University of Isfahan in Iran.