A Combination of Frequent Pattern Mining and Graph Traversal Approaches for Aspect Elicitation in Customer Reviews

SEPIDEH JAMSHIDI NEJAD1, FATEMEH AHMADI-ABKENARI2, AND PYMAN BAYAT1

1Department of Computer Engineering, Islamic Azad University, Rasht Branch, Rasht 414764919, Iran
2Department of Computer Engineering and Information Technology, Payam-Noor University, Rasht 41635-4315, Iran

Corresponding author: Fatemeh Ahmadi-Abkenari (fateme.abkenari@gilan.pnu.ac.ir)

ABSTRACT Due to the remarkable increase in e-commerce transactions, people try to have an appropriate choice of purchase through considering other people’s reflected experience in product’s or service’s reviews. Automatic analysis of such corpus requires enhanced developed algorithms based on natural language processing and opinion mining. Moreover, the linguistic differences make extending existing algorithms from one language to another challenging and in some cases impossible. Opinion mining focuses on different subjects of review analysis such as spam detection, aspect elicitation and polarity allocation. In this article, we focus on detection of explicit aspect and propose a methodology to overcome some difficult and problematic aspect compounds in the form of multi-words format in Persian language. Our approach proposes the construction of a directed weighted graph (ADG structure) based on some yielded information from FP-Growth frequent pattern identification algorithm on our corpus of Persian sentence. Traversing some special paths within the ADG graph according to our developed rules could lead us to the extraction of problematic multi-word aspects. We utilize Neo4j NoSQL graph database environment and its Cypher query language in order to create the ADG graph and access the desired paths that reflects our developed rules on the ADG structure which lead us to extract the multi-word aspects. The evaluation of our methodology with the existing approaches on the issue of aspect derivation in Persian language including ELDA, SAM, an MMI-based and an LRT-based algorithms indicates the robustness of our approach.

INDEX TERMS Aspect extraction, explicit aspects, multi-word aspects, opinion mining, Persian aspects, sentiment mining.

I. INTRODUCTION

Opinion mining as a sub domain of data mining is tightly related to natural language processing and has a numerous applications in various domains including customer relationship management and marketing. Elicitation of online sentiments and votes of customers makes the initial corpus as a base on which the opinion mining approach works. Tourism related activities could take advantage of this trend and attract their customers via utilizing social medias and web sites in which customers and guests could freely express their experience on different hotels’ or other tourism’ services. Nowadays users face with overloaded information of this type since before decision making on using a special service or purchasing a product, they consider other users’ experience. Due to the great number of web sites that lists users’ ranking, votes and opinions on tourism services, reading all of them in a non-automated manner is an impossible and a tedious task for capturing a general picture of a tourism service.

Opinion mining tries to extract users’ opinions on a product or service. The problem is that due to the great differences in linguistic structures, an accurate and trustworthy text mining and NLP based approach on a language could not be extended to the others. One challenge in this domain is that users express positive or negative sentiments on different features of a product or service but this doesn’t mean that they have positive or negative opinions on that product or service as a whole [1], [2]. So opinion mining on aspect level tries to derive those textual concepts that users express their thoughts and votes on them. Some example of aspects on a hotel...

The associate editor coordinating the review of this manuscript and approving it for publication was Xin Luo.
entity could be hotel’s location or architecture and some aspects on cell phone entity could be battery life or camera. Moreover, aspects are categorized into two classes of explicit and implicit ones. For example in the sentence “This hotel is expensive but the service quality is fantastic”, the focus in the first place is on the hotel entity as a whole, then the word “expensive” as an opinion word mentions “price” as an implicit aspect while the rest of the sentence comments on “service quality” as an explicit aspect. The focus of this article is on the extraction of explicit aspects.

This article’s aim is to address the aspect extraction challenges in Persian language. So, the initial data set is gathers based on the online submitted users’ opinions from iran-hotelonline.com and egardesh.com2 web sites on tourism and in particular on hoteling domain. Next, the preprocessing phase includes text normalization, informal phrase identification and transformation of them to formal ones and part of speech tagging are implemented. Then our sentimental field oriented vocabulary (here on hoteling domain) is constructed for the first time on Persian language and the associated graph is built for faster traversal and information retrieval especially for opinion spam detection. The third stage includes subjectivity categorization and separation of subjective and objective sentences. The forth phase consists aspect extraction. So an initial candidate set is populated through utilizing a compound method based on high frequent word’s occurrence and the nouns that happens with a variable radius from the opinion words. Then after employing a pre-pruning process, the single-word aspect words have been extracted. The main challenge lies on multi-word aspects extraction. We utilize the dependency graph to capture those multi-word aspects that includes noun phrases. But based on the linguistic structural complexity in Persian language, employing dependency graph has its own challenges. For example while the nominal phrase of “personnel’s behavior/ᄌرف‌گر کارکنان” is an aspect, the phrases such as “good personnel’s behavior/_good_ᄌرف‌گر خوب کارکنان” happens in Persian in a way that the adjective comes in between the two words that makes the multi-word aspect extraction a challenging task. Hence in this article, the related challenges for aspect extraction from problematic aspect compounds in Persian language are addressed. To do so, the FP-Growth algorithm is employed for identification of dependent words that come together with the notion of frequent item sets in the market basket. We also develop some rules as the post pruning task to omit some candidate multi-word aspects. These stages is used to set up a weighted and directed graph for multi-word aspect elicitation in Persian language through traversing some special paths according to our developed rules. We compare the results with four approaches of ELDA, SAM, MMI-based and LRT-based approaches which were proposed in literature for multi-word aspect extraction in Persian language. The evaluation results shows the robustness of our approach in compare to the others in extracting aspect compounds.

In continue, we review the related literature on aspect extraction in English and Persian languages since the most algorithms are proposed for English language. Then the methodology of this article will be discussed in detail. Finally, the evaluation results and the conclusion will be presented.

II. RELATED WORKS
Mohammadi et al. (2019) employed an unsupervised approach for detecting explicit features in Persian language. Their methodology consists text preprocessing, identification of syntactic word’s roles and drawing sentence dependency graph from which the features and sentiments were extracted together in separated windows. This research tires to distinguish between the adjectives as a part of noun or as a part of a sentiment in the sentence structure. To detect explicit multi-word features they employed LRT test with the sentence dependency graph to check the co-occurrence of words. For this aim, they considered the lexical and syntactic roles of the words together. In this research, some challenges regarding the linguistic structures of Persian sentences have been listed as represented in table 1 [3].

Bagheri (2019) developed a joint sentiment-aspect detection model (SAM) based on LDA approach. This article adds a sentiment layer in topic modeling method of LDA in which aspects were considered to be related to sentiment labels which the latter is associated to the documents. SAM generates documents by latent single and multi-word topics and models the distribution of words around each topic. For an aspect based analysis, the model in SAM learns the unigram and bigram distributions over the topic by introducing word status in order to learn samples from the type of distributions. SAM also works on document level. The author applied SAM on English and Persian languages on three domains of film, cell phone, DVD players’ and restaurant reviews. This research listed some obstacles in detecting aspects from Persian sentences as illustrated in table 1 [4].

Razavi and Asadpour (2017) proposed an unsupervised approach for aspect identification based on an embedding word method in Persian reviews. In this research, the aspect keywords are conceptually categorized by utilizing syntactic and semantic relationship in word embedding vectors in order to differentiate between implicit and explicit aspects and the elicitation of multi-word aspects [5].

Rabooki et al. (2015) introduced a methodology to extract features in Persian language on cell phone reviews and phones in university domain expressed by students. Their work include the construction of a Persian lexicon to determine the orientation of user’s reviews, preprocessing, POS tagging and syntactical dependency parsing. Their research extracted features based on the frequency and grammatical dependency notions and they found out the better

1 A Persian hotel reservation web site
2 A Persian hotel reservation web site
3 Likelihood Ratio Tests
4 Latent Dirichlet Allocation (LDA)
results from applying the latter approach but they indicated that grammatical based approach could not work well in detecting features’ polarity [6].

Bagheri et al. (2013) proposed an unsupervised domain independent model to extract implicit and explicit aspects by first detecting multi-word aspects and then measuring the impact of a sentiment word in a sentence on this process. Authors employed a bootstrap based method in order to score aspects based on their proposed PMI and frequency based metric. They used the explicit derived aspects and opinion words for extracting implicit aspects. They checked the accuracy of their algorithm on English reviews on electronic products [7].

Shams and Baraani (2017) worked on aspect extraction by proposing ELDA that works based on topic modeling algorithm of LDA and utilization of a prior knowledge of those similar aspects that occur together on the related topics in an iterative manner. They evaluated their algorithm on English and Persian languages [8].

Wu et al. (2018) suggested a hybrid supervised approach of machine learning and rules to carry out the tasks of ATEs and OTEs. First based on linguistic rules, they extracted nominal phrases as initial aspects and opinion targets. Then based on domain knowledge, they filtered out some irrelevant aspects. Then they used these data to train a GRU deep network in order to perform ATEs and OTEs [9].

Dragoni et al. (2018) suggested an opinion monitoring system for aspect extraction from real-time reviews based on unsupervised approaches encouraged users to visualize the analyzed data. The proposed architecture of this research could be used to produce a ranking system for better representation of opinions to highlight the positive and negative points of the entity. The effectiveness of this approach had been checked in SemEval challenge [10].

Marcacini et al. (2018) employed a cross-domain transfer learning approaches for aspect extraction problem through utilizing the labeled aspects of some domains in another domain in which there were no labelled aspects. To overcome the inconsistency of the mentioned approaches for feature space differences, they proposed the CD-ALPHN approach. Their network was constructed based on different information from labelled and unlabeled aspects and linguistic features [11].

Da’u et al. (2019) proposed the deep-learning approach of AODR for recommendation systems. In this method they extracted product aspects and associated weights through a deep learning method and then merged them into

---

| Challenge | Combination | Example |
|-----------|-------------|---------|
| 1         | Noun + Adjective | نوفمبر هتل طراحی شده بود. |
| 2         | noun + noun (+ noun) | نیروفرآیه برای استفاده در بخش مکانیکی |
| 3         | magnetic field strength | 1 Oe \(\rightarrow10^{5}/(4\pi)\) A/m |
| 4         | Adjective + noun (1st form) | زیبایی محوره خوب بود. |
| 5         | Adjective + noun (2nd form) | ارزش در مورد |
| 6         | Intersection | نمایش/توپن |
| 7         | Handwriting | نام/بکار/نام |
| 8         | Pictures | "د" "تکرار/فهرست " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " " 

---

5 Aspect Term Extraction
6 Opinion Target Extraction
7 Gated Recurrent Unit
8 Cross-Domain Aspect Label Propagation through Heterogeneous Networks
9 Aspect-based Opinion mining using Deep learning method for Recommender
a CF\textsuperscript{10} filtering method. Their approach included two main parts for extracting the aspects and then generating recommendations that employs TF\textsuperscript{11} method in order to fetch the overall rating prediction [12].

Wan et al. (2020) developed the AC-LDA\textsuperscript{12} approach in order to overcome the main problem of LDA method due to its incapability in deriving co-occurrence relationship such as aspect-aspect or aspect-opinion compounds. They extracted different word combinations and merged them into the LDA along with some constraints for preventing biased results caused by some general aspects for learning purpose. Their testing domain was on digital cameras from Taobao web site on Chinese language. They reported the identification of secondary aspects, low-frequency words and opinions without aspects with their approach [13].

Bagheri (2013) et al. listed some problematic linguistic related structures in Persian which made the task of aspect extraction difficult especially the existence of white spaces as inter-word or intra-word spaces as represented in table 1. They developed a Naïve Bayes based approach for polarity classification of sentiments and introduced a modified mutual information measure as MMI for feature selection. They applied their approach on Persian reviews on cell phones [14].

Shams and et al. (2012) presented a novel approach for generation of a lexical resource named PersianClues used for sentiment analysis in Persian language. Moreover, a novel unsupervised LDA-based sentiment analysis method called LDASA was proposed. In order to generate the PersianClues, at the first phase, an automatic translation approach was used to translate the existing English clues to Persian. Next, iterative refinement approach was used to correct the erroneous clues resulted from previous step. Then, topic-based polar sets were achieved from these clues and finally, each document was categorized into its related polarity using a classification algorithm [15].

Basiri and Kabiri (2017) addressed the problem of resource scarcity by introducing two new resources; a sentence-level dataset for sentiment analysis in Persian of SPerSent and a new Persian lexicon, CNRC. SPerSent contains 150000 sentences, each associated with two labels; a binary label indicating the polarity of the sentence, and a five-star rating. These labels were obtained automatically using a lexicon-based method. Specifically, three lexicons were used independently to label each sentence. Then, the majority voting and average methods were used to aggregate the results for polarity and five-star labels, respectively. Finally, Naïve Bayes, was used to evaluate the SPerSent [16].

Shams and et al. (2020) proposed three coarse-grained phases which were partitioned to manifold fine-grained operations. The first phase extracted the prior domain knowledge from dataset through selecting the preliminary polarity lexicon and aspect word sets, as representative of aspects. These two resources, as primitive knowledge, were assigned to an expectation-maximization algorithm to identify the probability of any word based on the aspect and sentiment. To determine the polarity of any aspect in the final phase, the document was firstly broken down to its constituting aspects and the probability of each aspect/polarity based on the document was calculated. They applied their approach on Persian and English reviews [17].

Liu et al. (2019) proposed a feature selection method called GI-FSw. It can be used to deal with the imbalanced classification problems, which are commonly encountered in real-world datasets. Its novelty was on the usage of imbalanced ratio dependent weight to revise the original Gini index. In other words, an embedded feature selection method using weighted Gini index (WGI) was used in their research. Its comparison with Chi2, F-statistic and Gini index feature selection methods showed that F-statistic and Chi2 reach the best performance when only a few features were selected. As the number of selected features increases, theirs proposed method had the highest probability of achieving the best performance. The ROC, AUC and F-measure were used as evaluation criteria. Experimental results with two datasets showed that ROC and AUC performance could be high, even if only a few features were selected and used, and only changed slightly as more and more features were selected. However, the performance of F-measure achieved excellent performance only if 20% or more of features were chosen. The results were helpful for practitioners to select a proper feature selection method when facing a practical problem [18].

Cai and Zhu (2018) proposed an effective feature evaluation criterion for multi-label feature selection, called neighborhood relationship preserving score. This criterion was inspired by similarity preservation, which is widely used in single-label feature selection. The method evaluated each feature subset by measuring its capability in preserving neighborhood relationship among samples. Unlike similarity preservation, authors addressed the order of sample similarities which can well be expressed the neighborhood relationship among samples, not just the pairwise sample similarity. Under this criterion, they also designed one ranking algorithm and one greedy algorithm for feature selection problem. The proposed algorithms were validated in six publicly available data sets from machine learning repository. Experimental results in six publicly available data sets from two application areas demonstrated that the proposed algorithms were superior to previous algorithms. In this implementation, they only used a simple linear kernel to construct the sample similarity matrix [19].

Zhang et al. (2019) proposed a semi-supervised STCS lexicon model based on spectral Clustering (SC). They proposed a method for constructing a topic-specific sentiment lexicon, which comprises three following models: first, they proposed a filtering text model, namely, FT model, to calculate the text influence value and obtain topic-specific hot comments as a preprocessing data set. second, in theirs proposed

\textsuperscript{10}Collaborative Filtering
\textsuperscript{11}Tensor Factorization
\textsuperscript{12}Association Constrained LDA
constructing sentiment relationship graph model, namely, CRM model, three factors, the base sentiment similarity, topic sentiment similarity, and synonym sentiment similarity between each pair of sentiment words, were proposed and calculated in theirs data set, and then they could obtain the factor of final sentiment similarity by adding the three values in proportion. Finally, they proposed a spectral clustering model, namely, SC model, to cluster the sentiment words on the basis of a sentiment relationship graph for obtaining the topic-specific sentiment lexicon, namely, STCS lexicon. The algorithm analysis and experimental results showed that the sentiment lexicon constructed by the STCS lexicon model exhibits better performance. It can accurately identify the sentiment of topic-related sentiment words [20].

Chakraborty et al. (2020) provided a multifaceted insight into the evolution of sentiment analysis into the limelight through the sudden explosion of plethora of data on the internet. This a article presented a detailed survey of social networks and its related terms. The works that had been accomplished relating to cluster, community and social networks had been described in its scope. This article mainly aims to bring out the shortfalls of the wide variety of papers making it easy for researches to apply sentiment analysis methods after accumulating data from social media. Authors also addressed the process of capturing data from social media over the years along with the similarity detection based on similar choices of the users in social networks. The techniques of communalizing user data had also been surveyed in this article. Data, in its different forms, had also been analyzed and presented as a part of survey in this article. Other than this, the methods of evaluating sentiments have been studied, categorized, and compared [21].

Kang et al. (2018) proposed a Bayesian inference method to explore the latent semantic dimensions as contextual information in natural language and to learn the knowledge of emotion expressions based on these semantic dimensions. They proposed two Bayesian models DWET and HDWET for exploring the latent semantic dimensioned as the context in natural language, and for learning the knowledge of emotion expressions with respect to these semantic dimensions. The basic idea was that probabilistic influence could flow between emotions and topics in Bayesian inference through a V-structure in the models, in which the emotion variable and the topic variable were located as two parents of the observed word variable. Theirs method synchronously inferred the latent semantic dimensions as topics in words and predicted the emotion labels in both word-level and document-level texts. Theirs experiment of the document-level and word-level emotion predictions, based on the Chinese emotion corpus Ren-CECps, demonstrated a promising improvement for emotion recognition compared to the state-of-the-art emotion recognition algorithms. The DWET model outperformed all base-line algorithms for word and document emotion predictions and the HDWET model, with a flexible concentration parameter injected in the hierarchy of corpus-level document emotion distribution, allowed a self-adjustment of emotion distributions through different documents, and significantly improved emotion recognition for the less common emotion categories with even better Recalls and F-scores compared to the DWET model [22]. LV et al. (2017) reviewed social media based transportation research with social network analysis and data mining methods. They summarized main research topics in this field, and reported collaboration patterns at levels of researchers, institutions, and countries, respectively. Generally this article reviewed recent advances in transportation by datamining, in which they analyzed researcher, institution, and country level collaboration networks, and research topics. To gain in depth understanding of social media based transportation research, they used social network analysis methods to analyze recent advances in this field and found out that the networks are relatively sparse which implies the future great development [23].

III. PROBLEM STATEMENT

Due to the fact that the task of opinion mining as a whole and one of its subdomain as aspect extraction could not be extended from one language to another due to the linguistic differences among languages, this article focuses on explicit multi-word aspect extraction in Persian language. For this aim, there are some problematic compounds structures that makes the task of multi-word aspect derivation a challenging task. Compounds such as noun + adjective, noun + noun and noun + adjective + noun happens as common structures in Persian sentences and a robust approach is on demand to extract them. There are some other methods but they either focused on single-word aspect elicitation or the derivation of simpler aspect compounds. We developed the ADG graph structure and some rules to traverse special paths along the graph in order to address these problematic aspect compounds’ extraction that will be discussed in continue.

IV. METHODOLOGY

In this section, our methodology for aspect extraction will be discussed. For the aim of aspect elicitation, we need to prepare our data through step one to three as depicted in figure 1. Aspect extraction will be performed in step four through four sections of A, B, C and D.

The initial dataset of this research consists the textual Persian opinions of users on domestic hoteling domain in Iran in a five years period that are collected from iranhotelon-line.com and egardesth.com web sites. Data was collected through a web crawler and saved in a database. The web crawler is developed in the C # programming language. Opinions was in the form of paragraphs of different sizes. We collected 10000 opinion paragraphs. Preprocessing of such texts includes text normalization, transformation of informal phrases to formal terms and part of speech tagging that have been carried out through NLTools13 in Persian.

13A Persian based software as a text parser, POS tagger, stemmer and text normalizer developed in Ferdosi university of Mashhad, Iran.
After POS tagging, adjectives and adverbs were extracted as sentimental words and opinions with ADJ and ADV tags to form the sentimental vocabulary in the field of hoteling. Such words were utilized to form our sentimental vocabulary.

First we separated sentimental words according to their POS as ADJ and ADV with different polarities and populated the initial vocabulary. Then through utilizing online dictionaries, we automatically searched the synonyms and antonyms of the words that we had in our initial vocabulary to extend the initial set. This approach was used in Hu and Liu (2004) [26]. This process was carried out in an iteration basis until no new words could be added to the vocabulary. After finishing the process, a manual review process was performed for checking the whole set. The result was a sentimental hoteling oriented vocabulary includes any nouns, adjectives and adverbs that users used to express their opinions for hoteling products and services. For faster and optimized utilization of the constructed vocabulary, in this phase the sentimental graph was drawn based on the sentimental vocabulary. To do so, we used graph based NoSQL database software of Neo4j. This graph includes 3062 vertices and 4763 edges. All adjectives, adverbs and any nouns was illustrated as vertices in different colors of blue, green and yellow respectively. Edges were drawn between those nodes that were in synonym or antonym relations with different colors. For example, edges of green color indicates synonym relations between any words and adverbs, edges of blue color indicates synonym relations between any words and adjectives and edges of yellow color indicates synonym relations between any words and nouns. After the manual control, we added 1508 words in hoteling domain that were not appeared in our gathered opinion paragraphs such as “Lobby”, “Architecture”, “Design”, ... in order to have a complete vocabulary in hoteling domain. As a result, our final vocabulary consists of 4570 words.
This vocabulary construction process was discussed in our previous research in detail [27]. Figure 2A and 2B depict a part of our sentimental hoteling oriented vocabulary in the form of the discussed graph structure.

For classifying the Subjectivity of sentences, they should be categorized into two groups of objective and subjective ones. An objective sentence expresses real information such as “the parking of this hotel located at the west wing”. On the other hand, a subjective sentence expresses opinions, ideas, or judgments such as “the parking of this hotel is very spacious”. We use the constructed vocabulary at the previous stage for determining the subjectivity based on Wiebe (2000) method [28]. Since most approaches for sentiment categorization are based on supervised learning, Wiebe [28] proposed an unsupervised method for sentiment categorization for which he utilized distributive similarity of Lin [29] in order to overcome the shortage in the volume of phrases including opinions in its primary corpus. We used Wiebe method for sentiment categorization and separation of objective and subjective sentences because according to its unsupervised nature, the algorithms does not need labelling all the samples. Since we constructed our sentimental dictionary at first, we does not need to employ Line method as Wiebe used. In this research, we developed a C# program based on the Wiebe method to differentiate between subjective and objective sentences based on two input files; our sentimental vocabulary and the preprocessed dataset. The result found out 21474 subjective sentences among 10000 paragraphs. The sentences such as “the personnel behavior was very polite” and “the hotel location was very good” are two examples of this phase’s output.

These phases are the needed infrastructures that first we used them in our other paper for opinion spam detection in Persian hoteling reviews [30] and now we utilize it for the task of aspect extraction that will be discussed in continue. Since aspects are the state of an entity on which users express their opinions, they usually appear within a subjective sentence as noun terms. So, upon extracting nouns and conceptual categorization of them we could reach a chaotic set of aspects in the first place.

Generally, the data preparation phase includes four automatic tools as: a web crawler (written in C#) for gathering the users’ opinions that could be employed frequently in an interval to refresh the corpus, an NLPTools for preprocessing step that could be done automatically, an automatic search on online dictionary for the construction of our sentimental dictionary, that could be run anytime, a written program in C# that works based on the sentimental dictionary and the initial dataset for discrimination of objective and subjective sentences in order to begin the process of extracting the
single-word or multi-word aspects. So in the interval of a week or a month, or any other time frequency, the corpus could be refreshed automatically.

This section consists of four steps of creating aspect candidate set, pre-pruning the candidate set, single-word aspect extraction, and multi-word aspect extraction as we will discuss in continue. Figure 3 depicts these steps along with the employed algorithms in each phase.

In the first part, the aspect candidate set is built based on the 10000 subjective preprocessed sentences from the previous part. Then the sentences were labelled for part of speech tagging through NLPTools [24]. This aim is carried out through utilizing a radius based notion which indicates the distance of the aspect term from the opinion words. While this radius is a constant number in many languages such as English, it is varied in different type of sentences in Persian language due to the linguistic structural complexity. Table 2 illustrates the output of NLPTools for POS tagging of some sample Persian sentences expressed by users and also indicate this various length radius in Persian language.

We develop a C# based program to extract all nouns with various radius distance from the opinion words [31]. Since aspects are those nouns that is frequently mentioned by users, they are among the words with high frequency occurrences of an entity. So upon employing TF-IDF approach via RapidMiner software as illustrated in figure 4 we reached a list of words according to their frequency based weights. In this implementation, we consider using Stemming and Tokenizer operators in order to avoid having words with the same stems or no standalone and meaningful tokens in the output. At this step, we populate the aspect candidate set based on the intersection of the output lists from applying varied radios notion and TF-IDF algorithm. The created aspect candidate set includes special nouns such as brand names and many synonym words that examples of the latter are "personnel, employee, staff" or "equipment, stuff" that belongs to the same conceptual group. So, in this section as illustrated in figure 3 with label B, the pre-pruning of the aspect candidate set was performed. The process includes the omission of brand names and conceptual grouping of synonym words which the latter was carried out based on our constructed discussed sentimental vocabulary. The remained set members will form our single-word aspects. Some examples of the...
FIGURE 4. Operation pipeline for word frequency detection with TF-IDF approach in RapidMiner software.

TABLE 2. Sample Persian sentences with various radius length distance of aspects to the opinion words which is pos tagged by NLPTools [24].

| Sentences                                                                 | POS tags labelling                                                                 | Opinion words            | various length radius |
|---------------------------------------------------------------------------|-----------------------------------------------------------------------------------|--------------------------|-----------------------|
| برخورد کارکنان در این هتل نسبت به همه مسافران هنوز عالی و محترم هست.       | پرداخت 4، کارکنان 4، در 4، این 4، هتل 4، N، نسبت 4، به 4، همه 4، مسافران 4، هنوز 4، عالی 4، محترم 4،ymph. | محترم هست.               | 5                     |
| The personnel behavior at this hotel was excellent and respectful for all the passengers of the hotel. | در 4، صحیح 4، به 4، جای 4، سوسیس 4، و 4، کلاسیک 4، از 4، گروههای سیزیجات 4 | Good                       | 9                     |
| و گریز اسفاده می‌شود که بسیار عالی و خوشمزه بود. | مناسب 4، در 4، این 4، هتل 4، فضایی 4، کمی 4، برا 4، پارک 4، در نظر گرفته شده است. | Excellent                   | 4                     |
| At this hotel's breakfast, instead of sausages, vegetable and meat groups were served, which was excellent and delicious. | در 4، صحیح 4، به 4، جای 4، سوسیس 4 | احتمال کمی این این هتل نسبت به همه مسافران هنوز عالی و محترم هست. |                        |
| و گریز اسفاده می‌شود که بسیار عالی و خوشمزه بود. | مناسب 4، در 4، این 4، هتل 4، فضایی 4، کمی 4، برا 4، پارک 4، در نظر گرفته شده است. | Good                       | 5                     |
| The facilities of this hotel were not good at all compared to its price.   | احتمال کمی این این هتل نسبت به همه مسافران هنوز عالی و محترم هست. | Good                       | 5                     |

We have already applied a pre-pruning step that not only impacts on the extraction of single-word aspects but also has...
a great impact in reducing the unwanted and noisy word in the initial candidate set before setting up our proposed ADG graph structure.

In the next stage, our aim is to find out multi-word aspects such as "sound insulation," "room cleanliness," and "restaurant's quality." This task is very challenging according to linguistic complexity in Persian language.

This section describe our approach to achieve this goal and solve the problem through three steps of: 1- utilizing dependency graph, 2- overcoming the problematic structures and 3- post pruning, that will be discussed in continue. We used the dependency graph in order to detect those multi-word aspects that are nominal phrases and exists in the initial pruned candidate set separately and they are not detected as related and dependent terms. We employed HAZM tool that is dedicated to parsing Persian sentences [32]. This tool determines words’ dependency and syntactic structure of a sentence. According to the dependency graph, some possessive relationship among words results a combined phrase of noun+ noun + noun structure such as “restaurant food quality." But the main challenge here is that in Persian linguistic structure, there are many phrases which could be categorized as multi-word aspects that they are not in the form of a pure nominal multi-word structure. For example in the sentence “The location is very good," the word "location" is two words in Persian as "loca-tional positioning" and word "loca-tional" is an adjective in Persian in aforesaid two words format. The dependency graph for a sample Persian sentence is depicted in figure 5. In this graph, "restaurant food quality" is detected as dependent words in the sentence of “The restaurant food quality was not bad.”

The dependency graph of this sample sentence is illustrated in figure 6 in which we couldn’t fetch the compound of “loca-tional positioning” as a multi-word aspect since it determine “loca-tional” as an adjective. As a result we could not rely on dependency graph as a means to overcome the problematic structures.

To overcome this challenge, we first consider all compounds in Persian that could be categorized as multi-word aspects. These compounds were listed in table 1 that collected and summarized by Mohammadi et al. [3]. As illustrated in the first three rows of table 1 (row No. four is a noun + noun format and row No. 5 represent a compound adjective), three forms of noun + adjective, noun + noun (+noun) and noun + adjective + noun could yield multi-word aspects in Persian language. In this research, we focus on elicitation of compounds in these three formats. For this aim we propose the construction of a weighted directed graph that we call it ADG as aspect detection graph from now forward. To do so, we employ FP_Growth algorithm and utilize its output information in some parts of the ADG structure including the existence of edges and the weights of edges. FP-Growth is a frequent pattern detection algorithm that mostly is employed in market basket analysis tasks such as another well-known algorithm of Apriori. The speed of FP-Growth in constructing the growing tree of frequent patterns are considerable and it is among the best ones to extract association patterns of the frequent item sets. Since the co-occurrences of dependent words could be seen as a frequent item sets in a market basket,
we employ FP-Growth algorithm to detect these dependencies to show the multi-word terms. Here we set parameters as support = 1, min item set = 2 and max item set = 3 for FP-Growth algorithm since the longest multi-word aspect we faced includes three words.

Figure 7, depicts the pipeline for FP-growth implementation in RapidMiner software. Through utilizing the last operator of “Create Association rules mining”, we are able to see the output association rules that are depicted in figure 8 (A). Figure 8 (B) shows the translation of rules. Figure 8 is limited to the rules that include any compound which include the two nouns of “personnel/behavior”.

So in the ADG (V, E), V as vertexes are the terms with any POS tags that appeared so far in our dataset that we illustrate them with different labels represented different classes of nodes according to their categorized POS tags. To do so, that all tags that refer to nouns such as NN, NNP, NNS, NNPS have the same class, all tags refer to adjectives such as JJ, JJR, JJS have the same class and .... We used different colors to show these various object groups in Neo4J. Also, E as edges refers to the sequence of terms in a sentence. We allocated the edges’ weights according to the number of occurrences of the two nodes together in our corpus extracted from FP-Growth algorithm. Since verbs in a Persian sentence appear at the end of sentences, so visiting a node with verb class indicates reaching the end of a sentence in ADG graph. For simplicity, we consider a path of maximum 10 edges length as the distance between the end of our aspect compounds and the end of each sentence in our Cypher codes. We implemented the ADG structure in the graph-based NoSQL database of Neo4J. A portion of ADG graph is illustrated in three compounds in a way that figure 9 depicts the ADG graph for compounds in the form of noun+adjective such as “locational positioning (good location)/”. figure 10 illustrates the compound of noun+noun such as “personnel behavior/”. and figure 11 represents the ADG graph for compounds in the form of noun + adjective + noun such as “personnel excellent behavior/”. Table 3 illustrates some sample Cypher codes as the Neo4J query language we implemented to retrieve the three types of A, B and C compound aspects from our ADG structure.

As described above, all dependent words resulted from employing FP_Growth algorithm to the pre pruned candidate...
aspects were appended to the ADG graph to complete the graph structure. Each compounds then were compared to the rules we developed as illustrated in table 4 to be remained or omitted from the final aspect list. So, these rules are developed as a post pruning means for compound based aspects.

As could be seen from table 4, if described conditions are satisfied, the identified aspects will be remained in the aspect list (e.g. rows No. 1, 2, 3) and otherwise they will be removed (e.g. row No. 4). As another post pruning and controlling attempt, we employed the PMI formula for checking the
mutual occurrence of our derived compounds and removed those with low score. We set the threshold of PMI to zero in order to omit negative scores. We run Cypher as the query language of Neo4j in order to reflect our developed rules to achieve those paths that included the multi-word aspects. Some of the derived aspects either single word or multi-words in hoteling domain from our proposed methodology are listed in Table 5.

V. EVALUATION

For evaluation of our methodology in the first place we employ fundamental algorithms of frequency-based, POS based and LDA on our dataset to extract single-word aspects. LDA topic modeling algorithm yields clusters of words in a way that each cluster is equal to a topic [33]–[36]. The pipeline to implement LDA approach in RapidMiner is illustrated in figure 12.

Another approach for aspect elicitation is a part of speech based method that is used by Hu and Liu (2004) and Blair and Goldensohn (2008) for extraction of infrequent aspects in a way that the nearest noun(s) to the sentimental or opinion oriented words labelled as aspects [26], [37].

We also implement this method on our dataset. Another algorithm we employed was checking our results with extracted aspects based on words with high frequency notion in a way that used by Ku, et.al (2006) in paragraph and document levels according to TF_IDF scheme [38].
approach also used by Hu and Liu (2004) on nouns and nominal phrases according to a heuristically set threshold for words’ frequency rate [26]. We implemented this approach by setting the threshold to 50 and as a result, 5 aspects populated the output set. In this trend we also utilized the word count diagram to illustrate the high frequency words according to the research by Scaffidi, et al. (2007) [39]. Figure 13 (A) depicts the derived word count diagram and figure 13 (B) shows its correspondent aspects and their importance weights.
Lastly the aspects are derived based on human judgement. The result of all of these implementations are presented and compared in figure 14.

As illustrated in figure 14, our methodology extracted the most number of single word aspects in compare to the derived aspects from human judgement task as the base
and frequency based, POS based and LDA approaches. Figure 15 shows the comparison of our methodology with other three approaches of frequency based, POS based and LDA in terms of F-measure for extraction of single-word aspects. It could be seen that our methodology along with the frequency based approach has the best results according to the number of extracted aspects and their accurate concepts.

The causes of these differences are POS based approaches consider the nearest noun term to the sentimental word while as discussed earlier, Persian sentence structure demands various radius in this matter while LDA topic modeling is incapable of distinguishing between sentiment and aspects and categorized both of them under a topic.

We evaluate our approach in extraction of multi-word aspects separately. Figure 16 depicts the comparison of our approach to four methods of LRT-based [3], SAM [4], ELDA [8] and MMI-based [14] in extracting three forms of multi-word aspects since the aim of these approaches are extraction multi-word aspects in Persian language from different perspectives. As could be seen, the results of our method outstrip the results of other four approaches in detecting the type C multi-word aspects. Our methodology has an equal predominance to MMI-based approach in
detecting type A and B multi-word aspects while outperforms LRT-based, SAM and ELDA methods.

It is necessary to note that the MMI-based approach only detect the two-words aspects in the form of noun + adjective and noun + noun compounds and has no strategy to detect the third form of noun + adjective + noun. The SAM approach as described previously uses LDA method to detect single-word aspects then utilizing the Markov chain, considers word orders, co-occurrence of the words and the word’s frequency. Hence it has a good results in detecting type A and C aspects while its results is not satisfactory in detecting type B aspects since it considers the co-occurrence of sentiment and aspect together in the form of noun + adjective more that noun + noun compounds. The ELDA approach has an appropriate results in detecting all three types since this method first applies LDA algorithm on the data set and then extracts all hubs and bridges based on the three notion of occurring with aspect terms, not belonging to the aspect itself and exclusiveness of a hub to one aspect. Then after removing all bridges, it improves LDA through the acquired knowledge. The ELDA and SAM approach according to the authors are not domain specific and could be extended to both English and Persian language. Hence they are not concentrated on handling aspect derivation problems in Persian.

VI. CONCLUSION
Since, in recent years, the number of transactions through utilizing e-commerce has had remarkable growth, people try to have an appropriate choice to purchase a product or service. One trend to achieve this goal is through considering other people’s reviews that express their experiences on it. In order to automatically extract the meaningful entities from short or long reviews, some linguistic-based tools are on-demand in each language, which work based on computer algorithms along with embedded linguistic knowledge. Opinion mining with the aid of natural language processing works in this domain. The problematic issue is that due to the linguistic differences, most of the applicable techniques in a source language cannot be easily extended to the others.

This article proposes a methodology for explicit aspect extraction from Persian reviews for both single-word aspects and multi-word aspects which working on the latter group is a more challenging task. To do so, we develop our methodology based on proposing the construction of a weighted directed graph as ADG structure. Through traversing some paths of ADG graph according to our developed rules, we could be able to extract the problematic aspect compounds. The graph elements including the edges and weights is derived from the output of the frequent pattern identification algorithm of FP-Growth.

Our methodology is employed to the reviews in domestic hoteling domain of Iran. We utilized the Neo4J graph database environment and its Cypher query language to create and access the desired paths of our methodology. The needed constructed infrastructure is briefly described which is used for applying our methodology for aspect derivation including the preprocessing task along with the domain-oriented sentimental vocabulary especially for detection of spam opinions and subjective sentence elicitation. The results of the evaluation of our methodology to other four approaches of ELDA, SAM, LRT-based and MMI-based notions which works on Persian multi-word aspect extraction task proves the robustness of our proposed method.

For the future work, our aim is focusing on polarity detection in Persian language and try to assign numeric polarities to sentiments instead of the three valued approach on −1, 0 and 1. To do so, we plan to utilize deep learning and the Auto Encoder (AE) neural network structure in order to receive the adjective describing features in our corpus as the input. The encoder part of the deep AE structure will be used for polarity classification purpose and the weights before applying the normalization function will be used as the real polarity of the adjective.

REFERENCES
[1] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum, “Integrating topics and syntax,” in Proc. Adv. Neural Inf. Process. Syst., vol. 17, 2005, pp. 537–544.
[2] E. Cambria, “Affective computing and sentiment analysis,” IEEE Intell. Syst., vol. 31, no. 2, pp. 102–107, Mar. 2016.
[3] A. Mohammad, M. R. Pajoohan, M. Montazeri and M. Nematabakhsh, “Identification of explicit features in persian comments,” J. Comput. Secur., vol. 6, pp. 1–11, Jan. 2019.
[4] A. Bagheri, “Integrating word status for joint detection of sentiment and aspect in reviews,” J. Inf. Sci., vol. 45, no. 6, pp. 736–755, Dec. 2019.
[5] S. A. Razavi and M. Asadpour, “Word embedding-based approach to aspect detection for aspect-based summarization of persian customer reviews,” in Proc. 1st Int. Conf. Internet Things Mach. Learn., 2017, pp. 1–10, doi: 10.1145/3109761.3158403.
[6] E. Golpar-Rabooki, S. Zaraghmifar, and J. Rezaeenour, “Feature extraction in opinion mining through Persian reviews,” J. AI Data Mining, vol. 3, no. 2, pp. 169–179, 2015.
K. Chakraborty, S. Bhattacharyya, and R. Bag, “A survey of sentiment analysis of customer reviews,” Knowl.-Based Syst., 2013, doi: 10.1016/j.knosys.2013.08.011.

M. Shams and A. Baraani-Dastjerdi, “Enriched LDA (ELDA): Combination of latent Dirichlet allocation with word co-occurrence analysis for aspect extraction,” Expert Syst. Appl., vol. 80, pp. 136–146, Sep. 2017, doi: 10.1016/j.eswa.2017.02.038.

C. Wu, F. Wu, S. Wu, Z. Yuan, and Y. Huang, “A hybrid unsupervised method for aspect term and opinion target extraction,” Knowl.-Based Syst., vol. 148, pp. 66–73, May 2018.

M. Dragoni, M. Federici, and A. Rexha, “An unsupervised aspect extraction strategy for monitoring real-time reviews stream,” Inf. Process. Manage., vol. 56, no. 3, pp. 1103–1118, May 2019, doi: 10.1016/j.ipm.2018.04.010.

R. M. Marcacini, R. G. Rossi, I. P. Matsuno, and S. O. Rezende, “Cross-domain aspect extraction for sentiment analysis: A transductive learning approach,” Decis. Support Syst., vol. 114, pp. 70–80, Oct. 2018, doi:10.1016/j.dss.2018.08.009.

A. Da’u, N. Salim, I. Rabiu, and A. Osman, “Weighted aspect-based opinion mining using deep learning for recommendation system,” Expert Syst. Appl., vol. 140, Feb. 2020, Art. no. 112871, doi:10.1016/j.eswa.2019.112871.

C. Wan, Y. Peng, K. Xiao, X. Liu, T. Jiang, and D. Liu, “An association-constrained LDA model for joint extraction of product aspects and opinions,” Inf. Sci., vol. 519, pp. 243–259, May 2020, doi:10.1016/j.ins.2020.01.036.

A. Bagheri, M. Saraeae, and F. de Jong, “Sentiment classification in persian: Introducing a mutual information-based method for feature selection,” in Proc. 21st Iranian Conf. Electr. Eng. (ICEE), May 2013, pp. 1–6.

M. Shams, A. Shakerly, and H. Falli, “A non-parametric LDA-based induction method for sentiment analysis,” in Proc. 16th CSI Int. Symp. Artif. Intell. Signal Process., Shiraz, Iran, 2012, pp. 216–221, doi: 10.1109/ASIP.2012.6313747.

M. M. Rasir and A. Kabi, “Sentiment level sentiment analysis in persian,” in Proc. 3rd Int. Conf. Pattern Recognit. Image Anal. (IPRIA), Apr. 2017, pp. 84–89, doi: 10.1109/IPRIA.2017.7983023.

M. Shams, N. Khoshavi, and A. Baraani-Dastjerdi, “LISA: Language-independent method for aspect-based sentiment analysis,” IEEE Access, vol. 8, pp. 31034–31044, 2020, doi: 10.1109/ACCESS.2020.2973587.

H. Liu, M. Zhou, and Q. Liu, “An embedded feature selection method for imbalanced data classification,” IEEE/CAA J. Automatica Sinica, vol. 6, no. 1, pp. 793–715, Mar. 2019, doi: 10.1109/ACCESS.2019.1911147.

Z. Cai and W. Zhu, “Feature selection for multi-label classification using neighborhood preservation,” IEEE/CAA J. Automatica Sinica, vol. 5, no. 1, pp. 320–330, Jan. 2018.

B. Zhang, D. Xu, H. Zhang, and M. Li, “STCS lexicon: Spectral-Clustering-Based topic-specific chinese sentiment lexicon construction for social networks,” IEEE Trans. Comput. Social Syst., vol. 6, no. 3, pp. 403–1189, Dec. 2019.

K. Chakraborty, S. Bhattacharya, and R. Bag, “A survey of sentiment analysis from social media data,” IEEE Trans. Comput. Social Syst., vol. 7, no. 2, pp. 450–464, Apr. 2020.

X. Kang, F. Ren, and Y. Wu, “Exploring latent semantic information for textual emotion recognition in blog articles,” IEEE/CAA J. Automatica Sinica, vol. 5, no. 1, pp. 204–216, Jan. 2018.

Y. Li, Y. Chen, X. Zhang, Y. Duan, and N. Li, “Social media based transportation research: The state of the work and the networking,” IEEE/CAA J. Automatica Sinica, vol. 4, no. 1, pp. 19–26, Jan. 2017.

E. Asgarian, R. Saeedi, A. Stiri, B. Bahmadi, and H. Ghaemi. NLPTools. Accessed: Jul. 1, 2016. [Online]. Available: https://wrlab.utm.ac.ir.

S. Jamshidi-Nejad, F. Ahmadi Abkenari, and P. Bayat, “Pre-processing challenges for persian opinion of E-tourism users using opinion mining and natural language processing solutions,” in Proc. 5th Int. Conf. Knowl. Eng. Innov. (KBEI), 2019, pp. 1–7.

M. Hu and B. Liu, “Mining and summarizing customer reviews,” in Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2004, pp. 168–177.

S. J. Nejad, F. A. Abkenari, and P. Bayat, “Employment of dictionary-based approach in opinion word elicitation: A case study in the hoteling domain,” in Proc. 6th Int. Conf. Web, 2020, pp. 1–3.

J. Wiebe, “Learning subjective adjectives from corpora,” in Proc. Nat. Conf. Artif. Intell., 2000, p. 30.

D. Lin, “Automatic retrieval and clustering of similar words,” in Proc. 17th Int. Conf. Comput. linguistics, 1998, pp. 1–7.

S. Jamshidi-Nejad, F. A. Abkenari, and P. Bayat, “Detection of users’ opinion spam in Persian language using supervised learning algorithms: A case study in the hoteling domain,” in Proc. 8th Joint Conf. Fuzzy Intell. Syst., Mashhad, Iran, Mar. 2020.

S. Jamshidi-Nejad and F. A. Abkenari, “Aspect extraction based on the neighborhood radius of opinion in persian: A case study in the domain hotel area,” in Proc. Eighth Joint Conf. Fuzzy Intell. Syst., 2020, pp. 1–8.

M. Khalash and M. Imany. (2013). Persian Language Processing Tool. [Online]. Available: http://www.sohbe.ir/hazm

T. Hofmann, “Probabilistic latent semantic indexing,” in Proc. 22nd Annul. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., 1999, pp. 50–57.

D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” J. Mach. Learn. Res., vol. 5, pp. 993–1022, Oct. 2003.

T. L. Griffiths and M. Steyvers, “Prediction and semantic association,” in Proc. Neural Inf. Process. Syst., 2003, pp. 11–18.

D. Blei, L. Carin, and D. Dunson, “Probabilistic topic models,” IEEE Signal Process. Mag., vol. 15, no. 4, pp. 424–440, Nov. 2010.

S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. Reis, and J. Reynar, “Building a sentiment summarizer for local service reviews,” in Proc. WWW, 2008, pp. 1–10.

L. W. Ku, Y. T. Liang, and H. H. Chen, “Opinion extraction, summarization and tracking in news and blog corpora,” in Proc. AAAI Spring Symp., 2006, pp. 1–7.

C. Saffaki, K. Bierhoff, E. Chang, M. Felker, H. Ng, and C. Jin, “Red opal: Product-feature scoring from reviews,” in Proc. 8th ACM Conf. Electron. Commerce, 2007, pp. 182–191.