Journal of AL-Qadisiyah for computer science and mathematics Vol.10 No.3 Year 2018
ISSN (Print): 2074 – 0204 ISSN (Online): 2521 – 3504

Proposed aspect extraction algorithm for Arabic text reviews

Ahmed bahaa aldeen abdul wahhab

Middle technical university
Technical college of management
ahmed80.ab@gmail.com

aliaa kareem abdul Hassan

University of Technology
Computer Science department
110018@uotechnology.edu.iq

Recived : 29\8\2018 **Revised : ** **Accepted : 12\9\2018**

Available online : 26 /9/2018

DOI: 10.29304/jqcm.2018.10.3.440

Abstract :

Opinion mining from reviews is a very crucial area in NLP. This area has many applications in social networks, business intelligence, and decision making. Aspect extraction is the main step to achieve opinion mining. This paper proposed an algorithm for aspect extraction from reviews in the Arabic language, to determine the aspects that the reviewers are described in their comments. The proposed algorithm begins with analyzing the comments dataset using latent Dirichlet analysis (LDA) to identify the aspects and its essential representative words, then extracting nouns and its' adjectives as a possible aspect phrase in a review. After that the categorizing process to categorize the extracted phrases according to the words specified from LDA analysis. The proposed approach has been tested by using two standard Arabic reviews datasets. The result was auspicious in spite of the difficulties if the Arabic natural language processing.

Key Words. Natural language processing, opinion mining, sentiment analysis, aspect extraction, latent Dirichlet analysis, LDA.

Journal of AL-Qadisiyah for computer science and mathematics Vol.10 No.3 Year 2018
ISSN (Print): 2074 – 0204 ISSN (Online): 2521 – 3504

1. Introduction.

Opinion mining means extracting and analyzing people's opinion about an aspect of an entity, while sentiment analysis (SA) is analyzed people's emotions towards entities such as products, services, and topics. SA can be classified into three levels: documents level assumes that each document holds opinions about one entity. Sentence level SA, which aims to classify the sentiment in a clause to positive or negative. The third class is aspect level an SA where the system finds what the writer like or dislike about the entity. It's also known as feature-based or attribute based sentiment analysis[1]. The goal of aspect level sentiment classifications is to specify aspects along with their sentiment. For example, "the food is delicious, but the service is very slow", reflects the opinion of the reviewer about two aspects: the food and the service, the sentiment toward the food is positive while service sentiment is negative. Aspect level SA includes many tasks, aspect extraction, aspect categorization, and aspect sentiment classification [2]. The aspect extraction process works to find the aspect and opinions about them before sentiment analysis step. Unfortunately, the opinion mining in the Arabic language did not receive enough attention, due to the limited number of tools and other challenges that relate to the nature of the language. This paper would manage two issues; the first is aspect extraction from the Arabic sentence, the second issue is aspect categorization.

2. Contributions

This work contributes to the field of Arabic sentiment analysis, firstly; by proposing a method to identify the aspect terms, and opinion words related to them. This target is accomplished by using latent Dirichlet analysis(LDA) to specify the most critical aspects mentioned by the reviewers and the words that used to describe these aspects. The second contribution is design a specific Arabic parsing system that works to extract the nouns and adjectives or in Arabic (الصفة و الموصوف), which is chunking system for the Arabic language that chunk bigrams and trigrams for a specific pattern like noun phrases patterns, then categorize the extracted phrase using specified words from LDA step.

3. Related works

Al-Samadi et al., worked on his first paper to classify sentiment of Arabic news by extracting aspect to classify news topics. For aspect, extraction he used N-gram feature pruning and Stanford POS tagger. His work relies on searching for nouns to classify news more than sentiment aspect extraction [3]. The second paper of Al-Samadi et al. was about aspect-based sentiment analysis for hotel reviews. His work depends on using SVM classier to classify extracted noun phrases. The system uses a training data set that has XML form extracted noun aspect by training SVM classifier. XML is a structured document while our work is more complicated as it uses unstructured text reviews about hotels and books[4]. Manahel et al. have built an aspect based sentiment analyzer for Arabic tweets depending on the parsing system to extract noun and adjectives from n-gram, then categorize this aspect using lexicon made by her [4]. Shima et al. worked on developing a root lexicon to lemmatize sentiment words in Arabic by collecting patterns that used to be sentiment words like (جميل = فعيل) and (افضل = افعل) but this work has a weakness mentioned by the author. The reason for this weakness is that the word orientation depends on the subject and the context that guide the sentiment of the word. For example, the word (big =كبير) has a positive orientation when talking about the hotel but negative orientation when talking about technology or an electronic device [6]. Abdul-Majeed et al. used the SVM approach for subjectivity and Arabic sentiment analysis. The feature used in that study are POS tagging, gender, and lemma as features, and polarity from a lexicon. The highest accuracy for sentiment classification was about 71% [7]. Al-Subaihin et al. built a system in two steps; the first is gaming an aa approach to build the lexicon through player annotation. The second step is a sentiment analyzer through word segmentation then calculate the accumulated score for the sentence. The precision reached 6,0.32 [8].

4. Challenges in the Arabic language

There are three types of Arabic language: classical Arabic, which is not used in our daily life, modern standard Arabic (MSA) and dialect Arabic [1]. MSA is a standard form that is used in official letters and schools. MSA is a simplified form of classical Arabic which is the language of the Quran and the old Arabian scriptures [9]. The dialect Arabic are local dialects used in Arabic countries, and this dialect not standard enough, and differs from each other in many idioms.

Journal of AL-Qadisiyah for computer science and mathematics Vol.10 No.3 Year 2018
ISSN (Print): 2074 – 0204 ISSN (Online): 2521 – 3504

For example, the English word (many), in Tunisian dialect (Barsha = برشة) to represent adjective many, while the same word in Iraqi dialect (hwaya=هواية) and in Egyptian dialect (keteer =كتيير) and son on [9]. The second challenge is the lack of Arabic lexicon for sentiment and also the absence of reliable tools for part of speech tagger, for example, tashfeen tagger can tag only verbs and nouns which is not much useful in the case of sentiment analysis [9]. Stanford tagger has some issues in its accuracy and sense of the right tag of some words. Also, there is a technical issue is that until now there are no reliable NLP tools for this language. There are also differences between researchers about how to deal with the Arabic corpus of text. For example, Abdulraheem, and Al-khlaifan recommend stemming to reduce the size of the lexicon corpus [10], while Rushdie Saleh et al., does not recommend stemming for the task of opinion mining, because stemming may alter the meaning of the word in context and may alter its POS tagging [11]. The other obstacle that most of the reviews in Arabic social networks written in Arabic dialect form, dialect form causes many problems in that it required different lexicon corpus for each local, national dialect [1].

5. Latent Dirichlet analysis

Latent Dirichlet analysis is a generative statistical model that let sets of observations in the text to be explained by unobserved groups, to explain the reason of behind some chunks of text is similar. LDA is a topic modeling method. LDA assumes that each review is a mixture of a small number of topics and each word participates in one of the reviews topics [12]. This method is identical to probabilistic latent semantic analysis, except in LDA topic distribution is assumed to have sparse Dirichlet prior. Dirichlet priors encode the intuition that reviews cover only a small set of topics and that topics used just a small set of words frequently. This method tries to find a statistical distribution for topics inside the document and a model for each topic in a document. Figure (1) explain the LAD topic word distribution model[12].

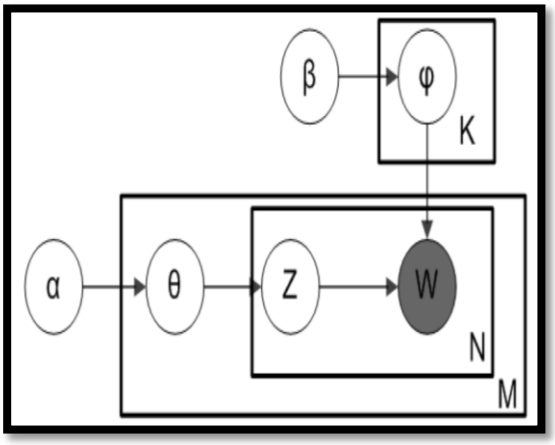

Figure (1) LDA documents analysis model

The probabilistic model in figure (1) represents the dependencies among the variables. The outer plate represents documents (reviews), while the inner plate represents the repeated word positions in a specific document. Each word position is associated with a choice of topic and word. M is the number of documents, and variables are: α is the parameter of the Dirichlet prior on the per-document topic distributions, β is the parameter of the Dirichlet prior on the per-topic word distribution, θm is the topic distribution for document, φk is the word distribution for topic k, Zmn is the topic for Nth word in document m and Wmn is specific word. Entities represented by θ and φ are matrices coming from decomposing the original document word matrix. θ Consist of rows of documents (reviews) and columns defined by topics. φ Consist of rows of topics and columns of words, so φ ……. φk refers to set of rows, each of which is distribution over topics .خطأ! لم يتم العثور على مصدر المرجع. Now the fully generative procedure for LDA: Assume that $\bar{X}$ is a document or a review in the case study of this paper, Gr topic, and t is a term

Then:

1. Start
2. S= number of topics
3. Generate the n tokens in ith document form a Poisson distribution
4. Generate relative frequencies $\Theta = (\theta 1, \theta 2. \ . \ . \theta k)$ of different topics in ith document from an Dirchilet distribution. This step is like generating$\theta r = p(Gr|\bar{x})$ for all topics r for a specific document (review). Note that$\theta r = p(Gr|\bar{x})$ probability of topic given document.
5. For each of the nth tokens in the document, first select rth latent component with probability P ($Gr|Xi$) and then generate jth term with probability P ($tj|Gr$), P ($tj|Gr$) where the probability of term given topic.

Journal of AL-Qadisiyah for computer science and mathematics Vol.10 No.3 Year 2018
ISSN (Print): 2074 – 0204 ISSN (Online): 2521 – 3504

6. Proposed system

The proposed system consists of two parts, the first part is the aspect extraction algorithm and the second part is the aspect categorization process. Flowchart (2) is a general view of these two processes:

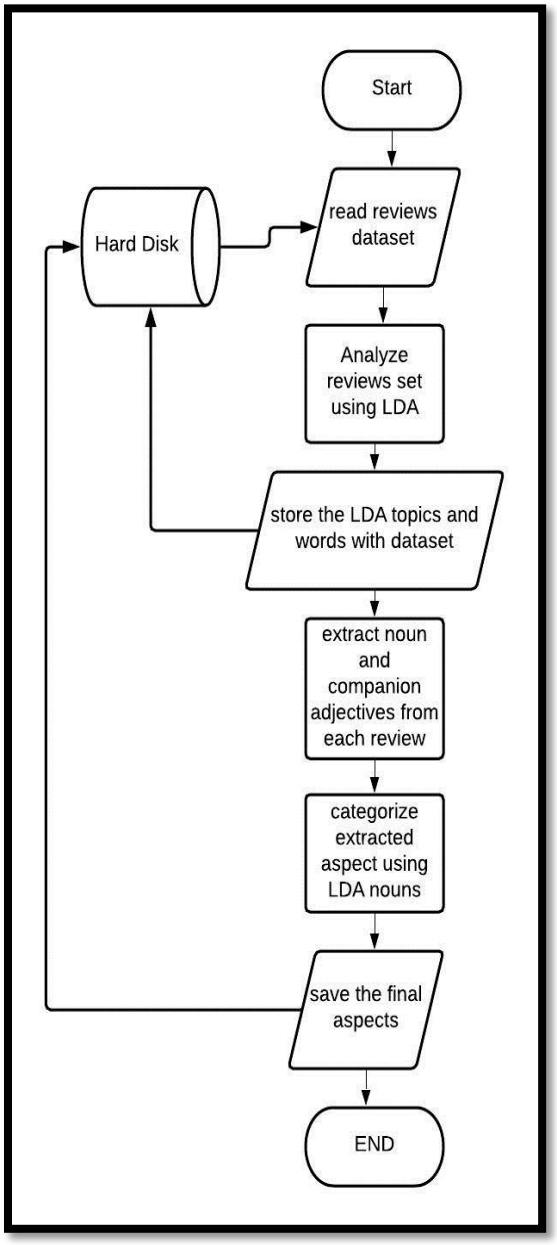

Figure (2) proposed system flow chart

The aspect extraction algorithm is a probabilistic approach that depends on the latent Dirichlet analysis model to identify the aspects, and a parser that parses the reviews to extract pattern of (described and description) or in Arabic (الصفة و الموصوف) by using the patterns of (NN, JJ) or (NN, NN, JJ), note that NN for noun and JJ for adjective. The proposed approach for aspect extraction exploit a specification in Arabic language which is that the attribute (adjective) come adjacent to the described noun like (الفندق رائع = hotel is wonderful), or adjacent to a composite noun in case of (NN, NN, JJ) in (خدمة الغرف جيدة) which is mean (room service is good). The Arabic language does not use auxiliary verbs, so the JJ (adjective) comes adjacent to a noun or composite noun. In the English language, the descriptor may be blocked from the noun by either auxiliary verb as in (Hotel is good) or by auxiliary verb and exaggeration formula (hotel was extremely good), so this would make aspect extraction process is little harder. After aspect extraction the step of aspect categorization is coming, firstly LDA must be done on the data to extract the central aspect mentioned by all reviewers, and specify the representative words for each aspect. The use of LDA analysis makes the proposed method probabilistic. There is one obstacle to extract a demanded pattern from Arabic text, is that there are no chunking tools to extract demanded pattern. NLTK has chunking tools for the English language only, which can extract noun phrases or any pattern effectively. For this reason, there is a need to build an Arabic chunking system or shallow parser that take bigrams, or trigrams and parse each word in these pieces and looking for a pattern of (NN+JJ) and trigrams with (NN+NN+JJ) as in

The following algorithm1:

Algorithm 1: Arabic chunk parser algorithm

Input: Arabic reviews corpus
Output: corpus of bigrams and trigrams of aspect phrases

1. Start
2. For all reviews in the corpus DO
3. Read a review
4. Clean text review from punctuation marks and non-Arabic text and numbers
5. Divide the text review into bigram chunks

   For i in range (0, length of review -1) DO

   Take every two adjacent words Bigram = [text [i], text [i+1]]

   Check the bigram to observe if hold aspect and opinion word

   IF bigram [0] in ['DTNN','NN','NNS','NNP'] AND bigram [1] like ['JJ']

   THEN: Save bigram to final list
6. Divide the text review into Trigram chunks

   For I in range (0, length of review-2) DO

   Take every three adjacent words Bigram = [text [i], text [i+1], text [i+2]]

   Check the Trigram to observe if hold aspect and opinion word

   IF Trigram[0] in ['DTNN','NN','NNS','NNP'] AND Trigram[1] in ['DTNN','NN'] AND Trigram[2] in ['JJ','ADJ'] THEN : Save Trigram to final list
7. Save the extracted chunks to the data frame
8. END.

Note that DTNN is DT for determiner followed by a noun, NN noun, NNS means noun singular, and NNP means noun plural. Now the second part of the process classifies the aspects according to the most aspects that the user has focused on his reviews about the hotel and from human experience that are:

- The hotel: contain general opinion about the hotel
- Rooms: user opinion about the room
- Staff: one of the most important aspects that most of the reviewers mention.
- Services: an aspect of general services like Wi-Fi internet, taxi, swimming pools, and spa.
- Price: is another crucial aspect of choosing the hotel
- Food: also is an aspect always described by visitors about the breakfast, restaurant, and bars.
- Location: location aspect consists the location of the hotel and its closeness from the city center and markets and other famous tourist places.

To specify the words that represent these aspects, we use the probabilistic approach by using latent Dirichlet analysis. LDA used to analysis the corpus reviews to determine the essential keywords through topics. To reach to the optimal number of topics, we have to run the LDA with a various number of topics on the same dataset with measuring coherency each time to choose the best amount of topics that give an excellent coherency of words. As in the following figure (3), the highest portion of coherency 0.45 when the number of topics reaches to 35 topics.

Journal of AL-Qadisiyah for computer science and mathematics Vol.10 No.3 Year 2018
ISSN (Print): 2074 – 0204 ISSN (Online): 2521 – 3504

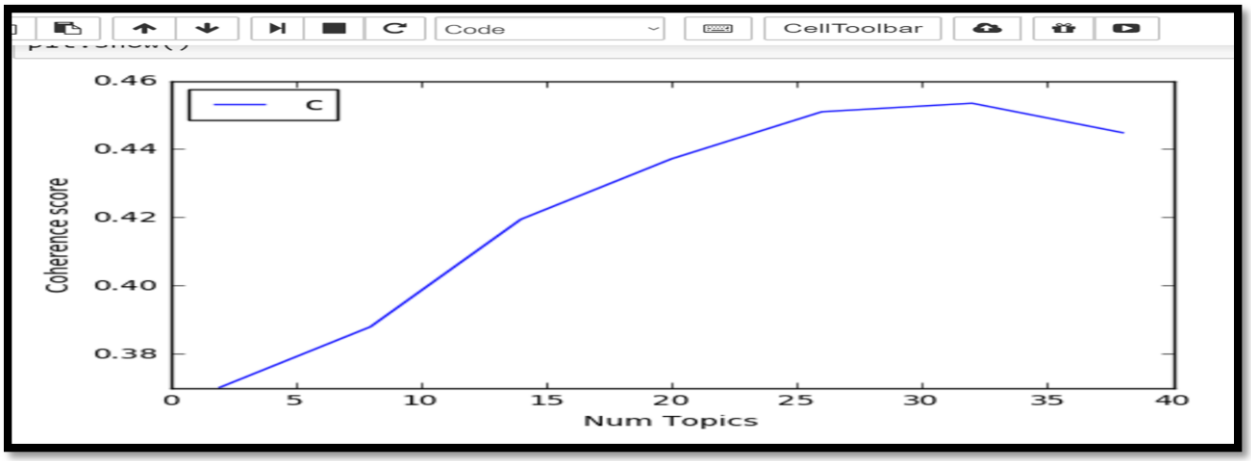

Figure (3) coherency chart for the different number of topics in LDA process

And the most important words to represent each aspect are:
Staff: 'وموظفي ' ,"عاملين"،"مدير"،'الطاقم'،'مضيافين'.... etc.
Rooms: أثاثا'،'وديكوره'،'وديكورة'،'فيلا'،'الشقه'،'الشقة'،'سرير ... etc.
Price:
رخيص'،'مجاني'،'ومجاني'،'مجانيه'،'أسعار'،'واسعار'،'وأسعار ...etc.
Location: 'وموقع'،'سيرا'،'قدميك'،'ألاقدام' بجانب ...etc.
Hotel:
'فندق'،'الفندق'،'تجربه'،'المكان'،'مكان'،'تجربة'،'منزل'،'نزل ... etc.

Food: الافطار'،'ألافطار'،'وجبه الافطار'،'وجبة الافطار'،'وجبه ...etc.
Services:
',نقل'،'مكيف_الهوا'،'مكيف_الهواء'،'دش'،'الانترنيت اللاسلكي' ...etc.
These extracted words from 35 topics would be used to categorize the extracted aspects.

The general chart bar of the most important words in the hotel reviews dataset is clear in the figure below (4):-

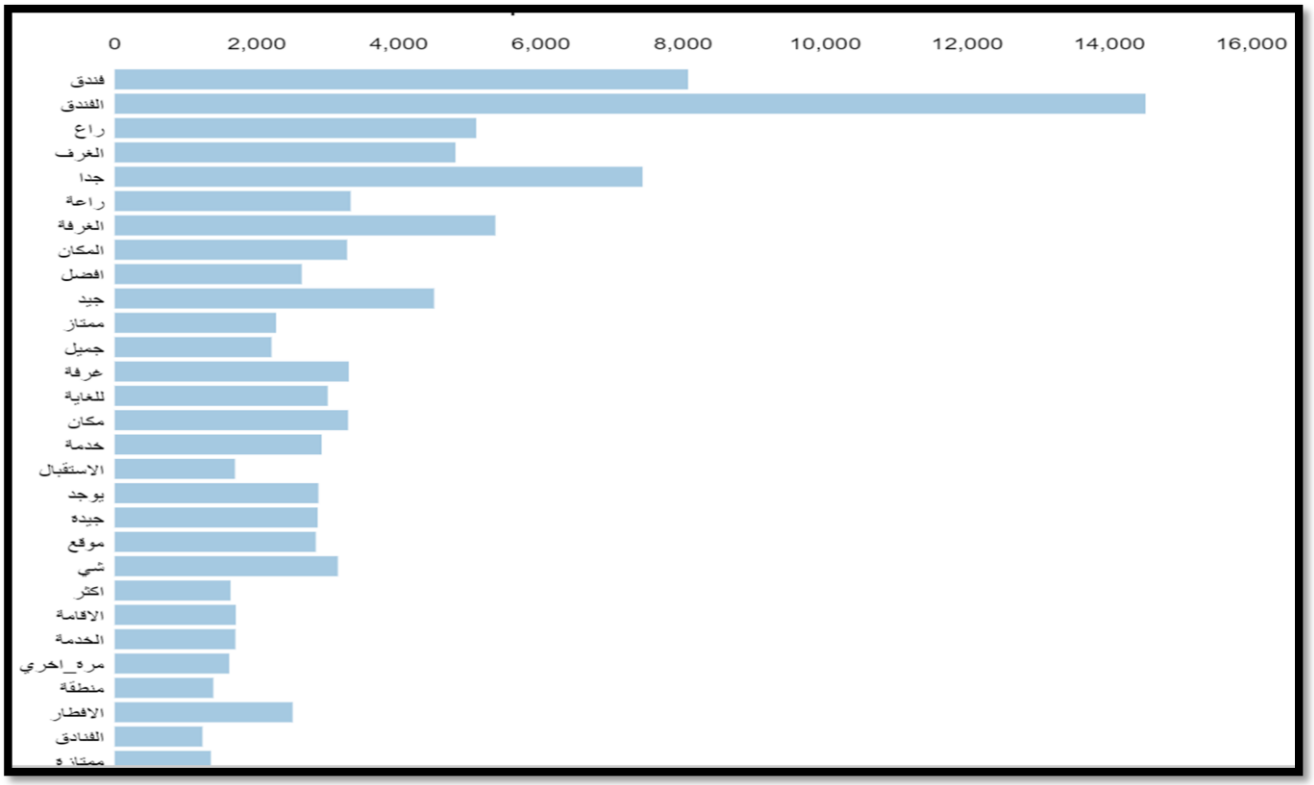

Figure (4) chart of the most essential words in the hotel's dataset using LDA

Journal of AL-Qadisiyah for computer science and mathematics Vol.10 No.3 Year 2018
ISSN (Print): 2074 – 0204 ISSN (Online): 2521 – 3504

The results of the LDA refers that there is overlapping between aspects inside these topics. This overlapping is useful, and it's powerful for our approach, it's not a weakness. The process of aspect

Categorization of aspects phrases by using the words specified from LDA analysis process. This

categorizing process is done by matching the noun in the extracted aspect phrase with the set of words that represent each aspect as in algorithm 2 as follows:-

```
Algorithm 2: aspect categorizing algorithm
Input: set of aspect phrases from reviews dataset
Output: corpus of categorized dataset
    1. Start
    2. For all extracted aspect phrases (i) :
                a. Clean phrase (i) from punctuation marks
                b. Split the words in phrase (i) to list x
                c. IF first word in list X[0] in hotel aspect word list THEN
                     Save phrase X (i) in hotel column

                   IF first word in list X [0] in rooms aspect word list THEN
                     Save phrase X (i) in rooms column

                   IF first word in list X [0] in service aspect word list THEN
                     Save phrase X (i) in service column

                   IF first word in list X [0] in staff aspect word list THEN
                     Save phrase X (i) in staff column

                   IF first word in list X [0] in prices aspect word list THEN
                     Save phrase X (i) in prices column

                   IF first word in list X [0] in food aspect word list THEN
                     Save phrase X (i) in foods column

                   IF first word in list X [0] in location aspect word list THEN
                     Save phrase X (i) in location column
    3. END
```

Now we use the same LDA analysis for the Hady al-Sahar hotel reviews dataset and calculate the coherency

to see the optimal number of topics that are yield the highest coherency which is 40 topics as in figure (5) below:

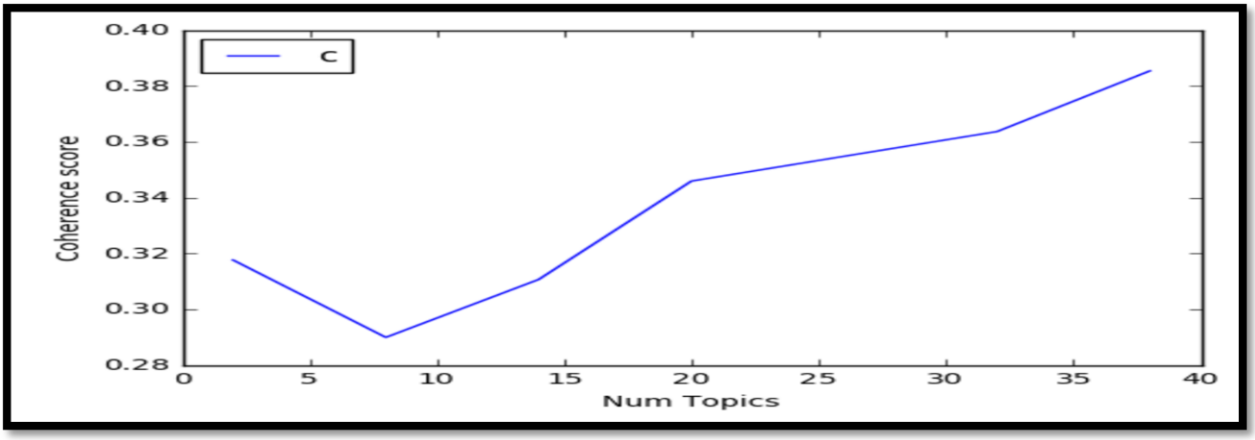

Figure (5) number of topics suitable for hotel dataset

After that, the most important words are extracted using the LDA, but in books reviews dataset we don't need to categorize the aspect, the need only to find the

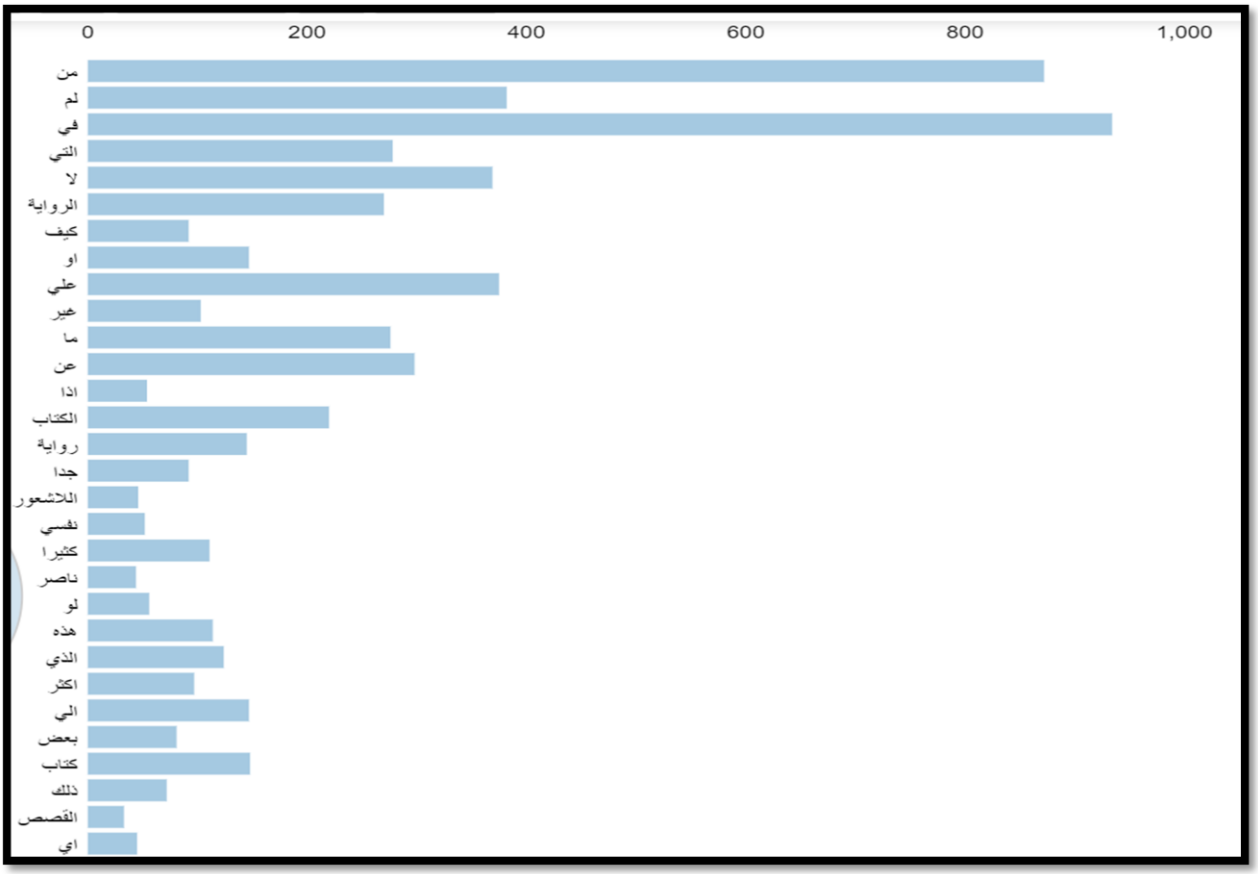

Figure (6) the most significant words in the 40 topics of hotels dataset

The result showed that the important words are:
['
الرواية','الكتاب','رواية','كتاب','القصص','قصص','الكاتب','كاتب','والحبكة','الحبكة','حبكة','الحزن','حزن','روايات','الحياه','الحياة','حياة','والحياة','الكاتبة','سرد','السرد','التاريخ','شخصية','حواسه','الافكار','الكتب','الادب','مقالات','اسلوب','اسقاطات','السرد','الادب','المقالات','الروايه','كتب','روايات','الروايات','الشعور','شعور','النص','مشاعر','تفاصيل','التفاصيل','اللغة','لغة','اللغه','الوصف','الاسلوب"]

The aspect categorization process for books dataset is different from hotels dataset because the aspect in hotels case is more detailed than books aspects. Hotels reviewers described detailed part of service or room, while the aspect of books is more general, for example, its extract aspect as (poem, novel, etc.), so it seems to be aspect name more than aspect description for an aspect of the product or service.

For obvious reason the process first is extracted aspect using LDA process, then find the important aspect by seeking for the noun with part of speech (DTN = determiner). For example, a noun like "AL-ketab" AL in Arabic is equivalent to "THE" determiner in English and "ketab" noun means "book") and if the percentage of this noun in the TFIDF table is less than 0.09 this noun would be accepted as an aspect name. The number 0.09 was specified from LDA as a threshold to specify the common nouns mentioned by reviewers as an indicator of its generality among them as an aspect.

Now all these explained in details in algorithm 3:-

Journal of AL-Qadisiyah for computer science and mathematics Vol.10 No.3 Year 2018
ISSN (Print): 2074 – 0204 ISSN (Online): 2521 – 3504


```
ALGORITHM 3: Books reviews aspect extraction
Input: Extracted aspect phrases
Output: nouns as aspect about books
1. Start
2. For i in range( 0 to length (aspect phrases)) :
      a. For noun in phrase(i):
            1- IF noun in TFIDF table THEN
                  A. IF TFIDF (NOUN) <=0.09 or noun in LDA books aspect
                        List THEN
                              Save NOUN
                  b. ELSE: ignore Noun
3. For I (0 to length (reviews)):
      For each word in reviews:
            IF part of _speech [word] ==DTNN AND TFIDF (word)≤0.09 THEN
                  Save Noun
4. END
```

But in many cases, the Arabian reviewer is not using the traditional formal approach to, the reviewer may use the narrative way to declare his opinion using verb phrase as in (" كلكم بحبكم وبشعر أني بيتي التحية خيي وسيم صالحة "). Or in English ("I love you all, and I feel like at my home, my brother "), so, this makes a misleading for any parsing system, causing an open problem.

7. Results

To test the proposed algorithm for extracting aspects from Arabic reviews text, two datasets have been used, the first al-smadi* Dataset for books reviews in the Arabic language, and the second is hady-al-sahaar**Arabic reviews dataset about hotels. This target is accomplished by calculating true positive, true negative, false positive, and then precision and recall. The true positive (TP) means the number of intersections between aspects tagged by the proposed algorithm and identified in the dataset. The false positive (FP) represent the number of aspects term occurrences specified by the proposed algorithm but not mentioned in the dataset. False negative (FN) represents the number of aspect terms occurs in the dataset but not have been identified by the proposed algorithm. Precision and recall then calculated

$$precision = \frac{TP}{TP+FP} \text{ ---------- (1)}$$

$$Recall = \frac{TP}{TP+FN} \text{ ---------- (2)}$$

The testing dataset consists of 400 reviews from the al-samadi dataset. We calculate the precision and recall for each row, then plotting the result in figure (7) and figure (8) as below:

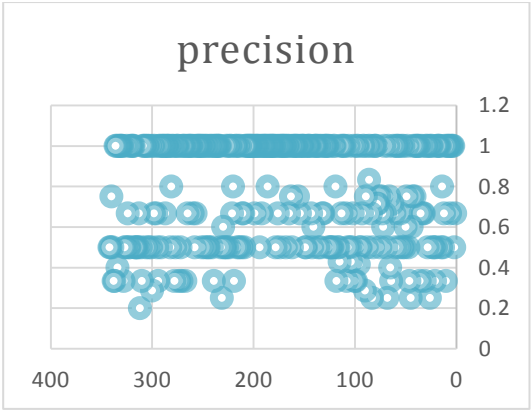

Figure 7: the precision plotting for each review in books datasets

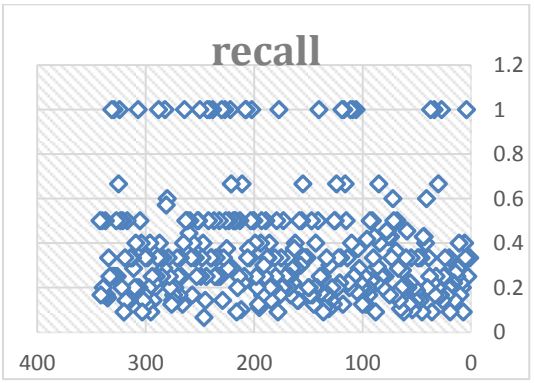

Figure 8: recall plotting for each review in books reviews dataset

From figure 7 and figure 8, the reader can notice that precision is between 0.2 and 1 while the recall taking values between 0.1 and 1. The system catch aspects from 28 reviews (8% from reviews) with accuracy reach to 100%. We can see that 106 reviews recall between 0.1 and 0.5, which make about 30% from reviews. So the precision and recall result is in the table (1) below:

Dataset	Precision	Recall	f-score
Books reviews	0.78	0.35	0.48

Table (1) accuracy of the al-samadi dataset

Hotel reviews dataset consist of 367 reviews have been analyzed, and a more detailed graph for precision and recall are in the figures (9) and (10):

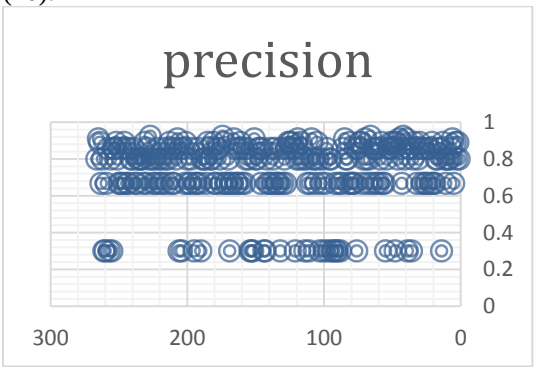

Figure (9) precision for each review in hotels dataset

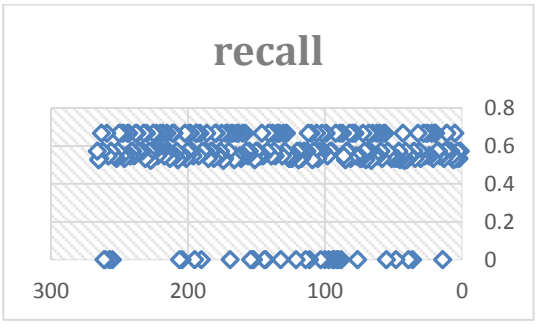

Figure (10) recall for each review in the hotel's dataset

The precision and recall for hady al_sahar hotels review dataset is on the following table (2):

Dataset	Precision	Recall	f-score
hotel reviews dataset	0.73	0.52	0.61

Table (2) accuracy of hotels dataset

From the table, the average of precision is 0.73, and most of the aspects have been identified with precision between 0.6 and 0.9 as in figure (9). In hotels dataset, the f-score is higher than f-score in books dataset because the precision and recall are more consistent.

8. Conclusions

Aspect extraction for the Arabic language is a new scope in Arabic NLP. The work on this paper revealed many issues; first, the need for a standard lexicon to specify sentiment orientation of words for different Arabic dialects, second Arabic NLP tools need more accurate taggers and consider the context of the text. During the work, we also specify the need for a shallow parser mechanism for Arabic language tools or (chunking), which can extract bigrams and trigrams with a specific pattern. The chunking software has been built in this work. Now let discuss our results, from results it's clear that precision is very good and it's varying according to the topic of the dataset. The researchers in aspect extraction should take into their consideration that there are two kinds of aspects, a general aspect that determines what is the text is talking about which is found in the books reviews dataset (AL-samadi dataset). The second type of aspect a more detailed aspect that appeared in hotel or restaurant dataset which is dealing with more deep points about the object like price, food, decoration,…etc.

The precision for books reviews dataset is reach to 0.78 and its good accuracy for the Arabic language in the absence of lemmatizes. The recall is 0.35 in books reviews aspect extractor because the false negative is small, as the proposed method can catch aspects more than aspects (and deeper) than aspects gotten by human annotator in the dataset, and this is not a weakness, but it is a positive indicator for the proposed aspect extractor efficiency. The precision for hotel reviews dataset is 0.73 and recall 0.52 for the same reason as in books dataset, but the categorization process was much easier in the hotel because the aspect was apparent in LDA analysis through the topics.

References

[1] AL-Awami.2016. Aspect extraction for sentiment analysis in Arabic dialect, University of Pittsburgh.

[2] B. Liu. 2012. Sentiment analysis and opinion mining. Morgan and Claypool Publishers.

[3] M. AL-Smadi, M.AL-ayyoub, H, H. AL-Sarhan, Y.assuj. **2015**. Using aspect sentiment analysis to evaluate Arabic news effects on readers. In proceedings of an 8th international conference on utility and cloud computing IEEE/AC. Limassol, Cyprus.

[4] M.AL-Smadi, O. Qwasmeh, B.Talafha, M.AL-Ayyoub, Y. Jararweh, E. bin Khalifa. **2016**. An enhanced framework for aspect-based sentiment analysis of hotels' reviews: an Arabic reviews case study. In Proceedings of the 11th International Conference on Internet technology and secured transmission ICITST. The UK.

[5] M. Ibrahim, N. Salim. **2016**. Aspect-oriented sentiment analysis model of Arabic tweets. International journal of computer sciences trends and technology (IJVST), Vol 4, issue 4.

[6] S. Ismail, A. AL-sammak, T. EL-shishtawy **2016**. A generic approach for extracting aspects and opinions of Arabic reviews. In the Proceedings of the 10th international conference on informatics and systems, pp: 173-179. Egypt.

[7] Abdul-Majeed, M Diab. **2012**. AWATIF: A multi-genre corpus for modern standard Arabic subjecting and sentiment analysis. In Proceedings of LREC. Turkey.

[8] A. Alsubaihin, H. AL-Khalifa, A. AL-Salman. **2011**. A proposed sentiment analysis tool for modern Arabic using human-based computing. In Proceedings of the 13th international conference on information integration and web-based applications, Hochiminh city, Vietnam.

[9] "N. Habash**.** **2010**. Introduction to Arabic, natural language processing. Morgan and Claypool. Doi: https://doi.org/10.2200/S00277ED1V01Y201008HLT010.

[10] L. Ibraheem, H. AL-Khalifa. **2012**. Exploring problems of sentiment analysis in informal Arabic. In Proceedings of the 14th international conference on information integration and web-based app. Bali. Indonesia.

[11] M. Rushdi-Saleh, M. Martin, L. Urena, J. Perea. **2011**. OCA: opinion corpus for Arabic. Journal of the American society for information science and technology. VOL 62 issue: 10, Pp: 2045-2054.

[12] D.M. Beli, A. Ng, M. Jordan. **2003**. Latent Dirichlet allocation, Journal of machine learning research, VOL 3, pp: 993-1022.

C. Aggarwal. **2018**. Machine learning for text, e-book, Doi: https://doi.org/10.1007/978-3-319-73531-3.

Journal of AL-Qadisiyah for computer science and mathematics Vol.10 No.3 Year 2018
ISSN (Print): 2074 – 0204 ISSN (Online): 2521 – 3504

خوارزمية استخلاص جوانب الحديث المطلوبة لتحليل الاراء العربية

احمد بهاء الدين عبد الوهاب
الجامعة التقنية الوسطى
الكلية التقنية الادارية
قسم تقنيات المعلوماتية

علياء كريم عبد الحسن
الجامعةالتكنلوجية
قسم علوم الحاسوب

المستخلص :

برز في السنوات الاخيرة مجال تنقيب الاراء من تعليقات الاشخاص كمجال مهم للدراسة. وقد تعددت تطبيقات هذا المجال لتشمل تحليل الاراء في الشبكات الاجتماعية، انظمة الاعمال الذكية ، وأنظمة اتخاذ القرار.للقيام بتنقيب الاراء فأن عملية استخلاص جوانب الحديث و تحديد ما تطرق اليه المعلق عن خصائص منتج او خدمةٍ ما احد المراحل الأساسية لتنقيب الرأي وتحديد شعور المعلق من الخدمة أو المنتج. يقترح هذا البحث خوارزمية لأستخلاص جوانب الحديث للغة العربية وذلك بدءاً بعملية تحليل الموضوعات الاحتمالي latent dirichlet analysis لجميع التعليقات ،وذلك لتحديد أهم الجوانب المشتركة التي ركز عليها الجمهور المعلق حول خدمة او منتوجٍ ما . يلي ذلك عملية استخراج جمل الاراء بنظام استخراج الانماط الاعرابية مستفيداً من تتابع الصفة و الموصوف فيِ اللغة العربية . وأخيرا تتم تبويب كل عبارة مستخرجة حسب أهم الجوانب المحددة من عملية LDA و الكلمات الممثِلة لهذهِ الجوانب . تم قياس دقة الخوارزمية المقترحة بواسطة مجموعتي بيانات قياسيتين ، الاولى تعليقات عن الكتب و الثانية عن الفنادق باللغة العربية وكانت النتائج جيدة رغم عقبات معالجة اللغة العربية .