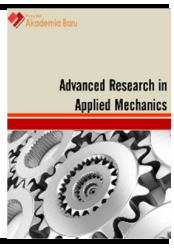
Penerbit Akademia Baru

Journal of Advanced Research in Applied Mechanics

Journal homepage: www.akademiabaru.com/aram.html
ISSN: 2289-7895



Opinion Lexicon Automatic Construction on Arabic language

Open Access

Fahd Alqasemi[1,*], Amira Abdelwahab[1], Hatem Abdelkader[1]

[1] Information Systems Department, Menoufia University, Menoufia, Egypt

ARTICLE INFO

Article history:
Received 5 June 2017
Received in revised form 10 July 2017
Accepted 4 December 2017
Available online 12 March 2018

Keywords:
Natural language processing, opinion mining, sentiment lexicon, lexicon-based SA, KNN, sentiment seeds

ABSTRACT

The most important web content is textual data, especially for decision makers in many domains. Text mining demands a significant focus in opinion mining or Sentiment analysis (SA), wherein text is analysed to understand text writer implied opinions. In this paper, we present a new approach for automatic opinion lexicon constructing on Arabic language. Since opinion lexicon is used for lexicon-based sentiment analysis approach, popular KNN search algorithm is inquired via some features to choose lexicon's terms. Constructed lexicon contents are gathered, initiating from a few number of terms seeds. And a term discrimination vector (TDV) is composed for each corpus term. The experimental results are very promising compared to other opinion lexicons.

1. Introduction

The most important web content is textual data, especially for decision makers in many domains like marketing, sport, business ... etc. Text opinion mining requires a significant focus. Opinion mining is called also sentiment analysis (SA). Two SA main approaches are machine learning Sentiment Analysis (MLSA) and lexicon-based Sentiment Analysis (LBSA). LBSA utilizes a terms list; i.e. opinion lexicon, which includes number of opinion-bear terms. This lexicon is used to count those terms appearance on target text. Then decide a text polarity, which has two cases; positive and/or negative [1].

In this paper, we had automatically built opinion lexicon which had utilized in LBSA approach. LBSA is unsupervised learning due to no need for labelled data set to discover text sentiment polarity [2]. So, our proposed approach is for automatic constructing opinion terms list that formed an opinion lexicon. A discrimination method is exploited four techniques based on the position of various terms over a text data set; i.e. corpus, without any knowledge about text polarity. Those four techniques are Semantic orientation (SO), total term frequency (TTF), sentiment keywords co-

* *Corresponding author.*
E-mail address: fhdahmd16@yahoo.com (Fahd Alqasemi)

occurrence measure SKCM, and inverse documents frequency (IDF). Every discrimination technique applied on each term features. The SKCM technique was presented in [3].

Those four techniques are preceded by Root-Based Word Find out algorithm (RBWF), which we had developed in [2]. RBWF is used for augmenting the number of sentiment seeds terms based on the roots of basic seeds to find out those seeds inflections. So, we had used number of seeds more than only basic selected seeds. Seeds Root Inflections Words (SRIW) are the seeds used here. They are the output of RBWF algorithm on basic seeds. Terms selection that composed our lexicon is done by KNN search or K Nearest Neighbour search algorithm. Where we choose an empirical value for k which is the number of seeds terms that we looked for their neighbours via corpus unique terms. Cosine distance measurement is used for distinguishing between neighbours. It is applied for KNN search inquiry technique.

The rest of paper is organized as follows; related works in section 2. Then, proposed approach is illustrated in section 3. After that, experiments and results are evaluated in section 4. Finally, conclusions are presented in section 5.

2. Related Work

Mahyoub *et al*., [4] had used Arabic WordNet semantic dictionary for giving a score to the lexicon terms, beginning with number of sentiment seeds. Then, they applied semi-supervised classification. This leads to exploit synset relations in Arabic WordNet to propagate plenty opinion bearing terms. At the end, they evaluated resulted sentiment lexicon using machine learning classification. The lexicon terms were on three polarities: positive, negative, and neutral.

Abdulla *et al*., [5] had integrated three constructing lexicons methods; manual, translation, and automatic methods for presenting an Arabic automatic opinion lexicon starting by manual collecting. Then, translating some sentiment terms from English. Automatic construction is done using terms frequency TF, which merged with former two lexicons. Then, tested after stemming comparing with the results before stemming.

In Duwairi *et al*., [6], a sentiment analysis framework was proposed. It dealt with Arabizi, Arabic dialects and emoticons. For this framework, three sentiment lexicons were collected. These lexicons were Jordanian Arabic dialect, Arabizi, and emoticons lexicon. Each lexicon had a link between its terms with modern standard Arabic (MSA). Arabizi are Arabic terms written in English Alphabet. They had begun with crowd-sourcing on twitter gathering big dataset. And then building those three lexicons manually that helped, finally, in the task of sentiment analysis.

El-Halees [7] had exploited an SA approach included three methods because they practically found out that using one method on Arabic SA would give weak results. In contrast with English language, they started with LBSA then supervised classification with maximum entropy. After that, they utilized KNN popular classifier on that sequenced order. For LBSA task, they initially used translated lexicon from English to Arabic, which missed some essential words. So, some of these words are manually added. Then, some unwanted are eliminated.

3. Proposed Approach

3.1. Semantic Orientation (SO)

Semantic orientation (SO) is computed via Point-wise mutual information PMI. So, we named it SOPMI. PMI is a very popular method that used in sentiment analysis [8], especially, in automatic lexicon construction process [9]. SOPMI is found by computing PMI relation for each corpus term

with every found SRIW. Then, subtracting negative PMI relations from positive SRIW ones for every term in corpus

3.2. TTF and IDF

Terms Total Frequency (TTF) means the total frequency of each unique term in whole corpus. It is found by counting each term appearance over corpus documents. Inverse document frequency (IDF) is a very known term weight used to give each unique term a single value. This value helps to impact term semantic weight on the whole corpus by the following equation [10]

$$IDF(t_k) = log\left(\frac{D}{DT_k}\right) \quad (1)$$

where D is the total number of corpus documents. DT_k is the number of documents term t_k belongs to. Also, IDF is found for every unique term in the corpus.

3.3. SKCM

SKCM stands for sentiment keywords co-occurrence measure [3]. It was used for enhancing the feature selected method for MLSA. However, in this work, we used it as a part of term discrimination vector. SKCM algorithm is supposed to take SA lexicon as an input. But in this work, we passed a small lexicon consisted of SRIW terms.

SKCM is simply looking for previously known terms polarity in corpus documents and computing values related to every unique corpus terms. Such value is a resulting weight for each term by counting term co-occurrence with each lexicon term. Then the operation finishes after subtracting negative terms number from positive number. This gives the weight of indicated term which must repeated for all unique corpus terms. All of these terms weights are named SKCM values and utilized in the domain-specific SA lexicon construction process here.

3.4. Lexicon Construction

The steps of the proposed approach are employed to discriminate between corpus terms based on how much they were close to SRIW. For achieving this objective, our approach presents those steps to reach the best possible results. So, we constituted what we named term discrimination vector TDV that consists of four values related to each term. They are SOPMI, SKCM, TTF, and IDF. We can denote TDV by the following equation

$$TDV(t_i) = [SOPMI(t_i), SKCM(t_i), IDF(t_i), TTF(t_i)] \quad (2)$$

where $TDV(t_i)$ is the term discrimination vector of term (t_i). $SOPMI(t_i)$ is the semantic orientation of term (t_i). $SKCM(t_i)$ is the sentiment keyword co-occurrence measure of term (t_i). $IDF(t_i)$ is Inverse Document Frequency of term (t_i). TTF (t_i) is a Total Term Frequency of term (t_i) from the unique corpus terms list. Figure 1 is showed the steps of proposed approach.

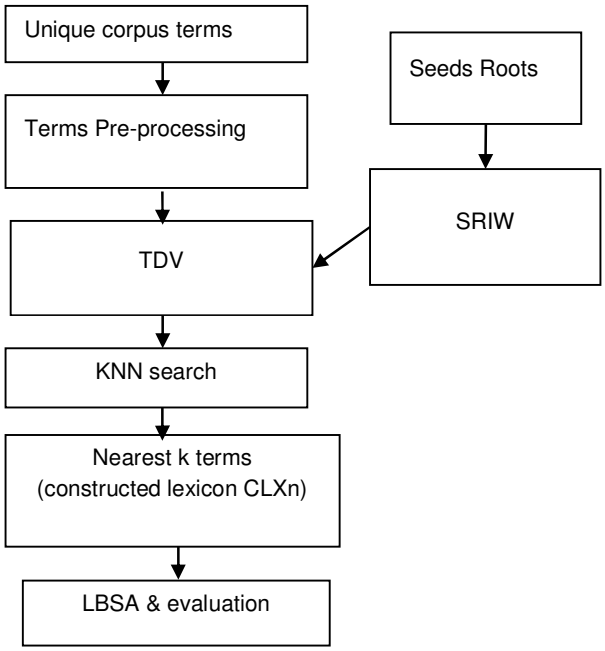


Fig. 1. The steps of the proposed approach

4. Experiments and Results

OCA corpus is a popular sentiment Arabic data set which had collected by Rushdi-Saleh *et al.*, [11]. It is a movie review corpus consisted of 500 documents, half of them are negative reviews and the other half are positive reviews.

Two sentiment lexicon are utilized here. The first one was previously generated automatically then refined manually in Nabil *et al.*, [12]. Whereas, the other is domain-adapted from the first one [2]. Hence, we compared resulted sentiment lexicons to these two premade lexicons on OCA corpus.

Firstly, OCA corpus is pre-processed by eliminating stop-words and singleton terms. Then, we selected seven seeds English terms for each class and translated them to Arabic. Constituting TDV is followed by finding selected SRIW terms. These terms are passed to KNN algorithm for finding k nearest terms to those seeds via TDV matrix; since we empirically assigned k value to 100. Then, we found the list of sentiment terms, automatically constructed lexicon. Table 1 showed the number of words of 5 lexicons gained with different values of k.

LBSA is the last step applied in the experiments. Thus, it had been done three times. The first two LBSA operations had been done via the two base lexicons used for comparing purpose, i.e. the two premade lexicons. The accuracy results are illustrated in Table 2.

Table 1
Constructed lexicons contents

Lexicon	Positive Words	Negative Words
CLX1	1824	2595
CLX2	2477	3628
CLX3	2922	4535
CLX4	3198	5357
CLX5	3515	6111

Table 2
LBSA accuracy of base lexicons

Lexicon	Accuracy result
Lex1	50.40 %
Lex2	70.40 %

The final LBSA operations had been done via these constructed lexicons. The LBSA accuracy result is illustrated in Table 3. Comparing between LBSA accuracy results it is obvious that the advantage of resulted lexicon, in all k values especially when *k* value is equal to 150.

Table 3
LBSA accuracy of different constructed lexicons

Constructed Lexicons (CLXn)	K value	TDV
CLX1	100	**73.20 %**
CLX2	150	**77.20 %**
CLX3	200	**76.60 %**
CLX4	250	**75.60 %**
CLX5	300	**75.60 %**

5. Conclusions

In this work, we presented a new approach for generating specific-domain lexicons in Arabic language automatically. Such lexicons are utilized on LBSA approach. Since lexicon generating process is done via KNN search method by passing a query including TDV matrix, selected terms from seeds terms. TDV is the term discrimination vector for corpus terms. It includes four values that had been found using four techniques; i.e. SOPMI, SKCM, TTF, and IDF. The proposed approach sentiment lexicon outperformed premade lexicons.

References

[1] Liu, Bing. "Sentiment analysis and opinion mining." *Synthesis lectures on human language technologies* 5, no. 1 (2012): 1-167.

[2] Alqasemi, Fahd, Amira Abdelwahab, and Hatem Abdelkader. "Adapting domain-specific sentiment lexicon using new NLP-based method in Arabic language." *International Journal of Computer Systems (IJCS)* 3, no. 3 (2016): 188-193.

[3] Abdelwahab A, Alqasemi F, abelkader H. "Enhancing the Performance of Sentiment Analysis Supervised Learning using Sentiments Keywords Based Technique." Seventh International Conference on Computer Science and Information Technology (CCSIT). (2017).

[4] Mahyoub, Fawaz HH, Muazzam A. Siddiqui, and Mohamed Y. Dahab. "Building an Arabic sentiment lexicon using semi-supervised learning." *Journal of King Saud University-Computer and Information Sciences* 26, no. 4 (2014): 417-424.

[5] Abdulla, Nawaf, Salwa Mohammed, Mahmoud Al-Ayyoub, and Mohammed Al-Kabi. "Automatic lexicon construction for arabic sentiment analysis." In *Future Internet of Things and Cloud (FiCloud), 2014 International Conference on*, pp. 547-552. IEEE, 2014.

[6] Duwairi, Rehab M., Raed Marji, Narmeen Sha'ban, and Sally Rushaidat. "Sentiment analysis in arabic tweets." In *Information and communication systems (icics), 2014 5th international conference on*, pp. 1-6. IEEE, 2014.

[7] El-Halees, Alaa. "Arabic opinion mining using combined classification approach." (2011).

[8] Turney, Peter D. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews." In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 417-424. Association for Computational Linguistics, 2002.

[9] Bai, Aleksander, Hugo Hammer, Anis Yazidi, and Paal Engelstad. "Constructing sentiment lexicons in Norwegian from a large text corpus." In *Computational Science and Engineering (CSE), 2014 IEEE 17th International Conference on*, pp. 231-237. IEEE, 2014.

[10] Al-Radaideh, Qasem A., and Laila M. Twaiq. "Rough set theory for Arabic sentiment classification." In *Future Internet of Things and Cloud (FiCloud), 2014 International Conference on*, pp. 559-564. IEEE, 2014.
[11] Rushdi-Saleh, Mohammed, M. Teresa Martín-Valdivia, L. Alfonso Ureña-López, and José M. Perea-Ortega. "OCA: Opinion corpus for Arabic." *Journal of the Association for Information Science and Technology* 62, no. 10 (2011): 2045-2054.
[12] Aly, Mohamed, and Amir Atiya. "Labr: A large scale arabic book reviews dataset." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, pp. 494-498. 2013